Name: Likhith K Raj
LinkedIn: LinkedIn Profile Link
GitHub: GitHub Profile Link
Portfolio: Portfolio Link

Zeotap Data Science Intern Assignment

Task 3: Customer Segmentation / Clustering

Google Colab Link:
https://colab.research.google.com/drive/1PaX0MM4w8EstBIxatRJen_5LuBkzkmsV?usp=drive_link

Steps for solving the problem:

- **Step 0: Data processing**

Point: All the 3 data files was cleaned and processed in EDA part, now combining all the 3 data files by common factor (product ID, customer ID)

- **Step 1: Feature Selection**

Point: Choose relevant numerical features like Quantity, TotalValue, and Price to represent customer behavior effectively for clustering.

- **Step 2: Data Scaling**

Point: Normalize the data using StandardScaler to ensure all features contribute equally to clustering.

- **Step 3: Dimensionality Reduction**

Point: Apply PCA to reduce dimensions to two, simplifying visualization while retaining most variance.

- **Step 4: Determine Optimal Clusters**

Point: Use metrics like the Davies-Bouldin Index or the elbow method to decide the number of clusters.

- **Step 5: Apply KMeans Clustering**

Point: Perform KMeans clustering to group customers into meaningful segments based on selected features.

- **Step 6: Cluster Assignment**

Point: Add the cluster labels to the dataset for analysis and visualization

- **Step 7: Visualize Clusters (PCA)**

Point: Plot the reduced data with cluster assignments using scatter plots to interpret customer groups visually.

- **Step 8: Calculate Cluster Quality (DBI)**

Point: Compute the Davies-Bouldin Index to evaluate the compactness and separation of the clusters.

- **Step 9: Interpret Results**

Point: Analyze cluster characteristics to understand customer behavior and group distinctions.

Clustering Evaluation Results for Customer Segmentation

I have implemented the clustering in 3 parts, one with 3 clusters (0.89), next with 4 clusters (1.03) and the last one is 9 clusters (0.721)

Cluster Configurations and Davies-Bouldin Index (DBI):

1. **3 Clusters:**
    - **Davies-Bouldin Index (DBI):** 0.89

      Analysis:
      This configuration shows moderate clustering quality. While clusters are fairly compact and separated, the overall segmentation might lack granularity.
2. **4 Clusters:**
    - **Davies-Bouldin Index (DBI):** 1.03

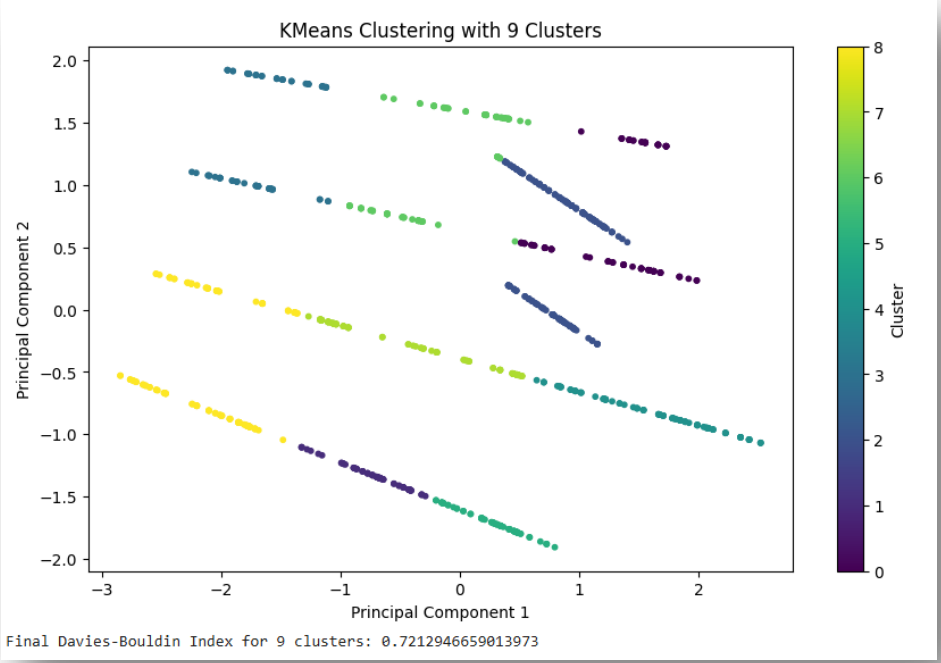
      Analysis:
      The clustering quality in this configuration is lower compared to others. Although the segmentation is simpler, cluster compactness and separation are less optimal.
3. **9 Clusters:**
    - **Davies-Bouldin Index (DBI):** 0.721

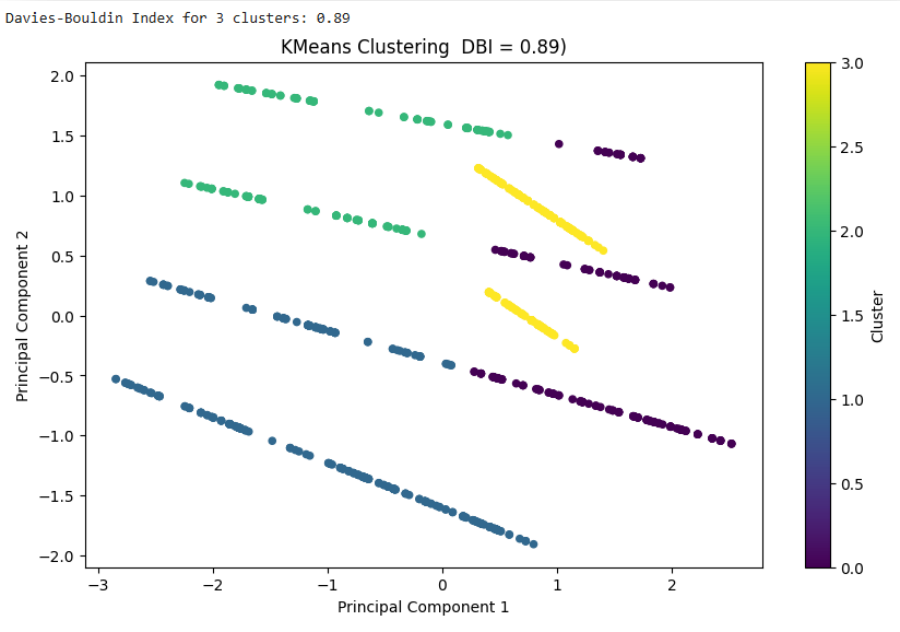
      Analysis:
      This configuration demonstrates the best clustering quality among the tested scenarios. The clusters are well-defined, compact, and adequately separated, providing granular segmentation.

Conclusion:

The **9-cluster configuration** is the most suitable option for customer segmentation, achieving the lowest DBI score of **0.721**. This indicates better-defined clusters, making it ideal for understanding and targeting distinct customer groups effectively.

Graphical Visualizations:





Please consider my profile:

Strong Data Science Background:
Extensive experience in data analysis, machine learning, and NLP, demonstrated through internships, projects, and certifications.

Proven Impact on Business Outcomes:
Delivered a 15% revenue improvement at Leucine and developed predictive models with high accuracy for stock trading and sentiment analysis.

Research and Innovation Focus:
Published papers on machine learning applications and fine-tuned models like Gemma 2 for Kannada, blending technical skills with research.

Technical Proficiency and Continuous Learning:
Skilled in Python, SQL, Tableau, and TensorFlow, with 100+ LeetCode problems solved and active participation in Kaggle competitions.