Angara Venkata Sai Singu Rishik
rishikangara@gmail.com
+91 7207167881

Data Science Intern Assignment by Zeotap

Customer Segmentation Clustering Report

- **Number of Clusters Formed:**
  - The clustering process resulted in **4 clusters**.
  - The distribution of customers in each cluster is:
    - i. **Cluster 3:** 70 customers (largest cluster)
    - ii. **Cluster 1:** 51 customers
    - iii. **Cluster 2:** 41 customers
    - iv. **Cluster 0:** 37 customers (smallest cluster)

- **Davies-Bouldin Index (DB Index): 0.9476**

  A lower DB Index indicates better clustering with well-separated clusters. Since **0.9476 is relatively low**, the clusters have **moderate separation and compactness**.

- **Other Relevant Clustering Metrics:**

  A. Silhouette Score:
  - a. **Value: 0.4319**
  - b. The silhouette score ranges from **-1 to 1** (higher is better).
  - c. A score of **0.4319 indicates moderate clustering quality**, meaning some overlap between clusters but still useful segmentation.

  B. Inertia (WCSS - Within-Cluster Sum of Squares):
  - a. **Value: 373.3513**
  - b. Inertia measures how tightly data points are grouped within clusters.
  - c. A lower inertia generally indicates better-defined clusters.

  C. Cluster Distribution:
  - a. The largest cluster (**Cluster 3**) contains **70 customers**, while the smallest (**Cluster 0**) has **37 customers**.
  - b. This suggests that certain types of customers (likely similar in behavior) dominate the dataset.

  D. Cluster Centroids:
  - a. The centroids represent the average feature values for each cluster.
  - b. [1.42 1.45 -0.14 -0.22 -0.17] - Cluster 0
  - c. [-0.11 -0.13 -0.57 -0.54 1.54] - Cluster 1
  - d. [-0.31 -0.23 -0.57 1.82 -0.64] - Cluster 2
  - e. [-0.48 -0.52 0.83 -0.54 -0.64] - Cluster 3
  - f. The differences in centroid values indicate **different spending behaviors** among customer groups.

- **Cluster Visualization Insights:**

1. **Scatter Plot of Clusters**

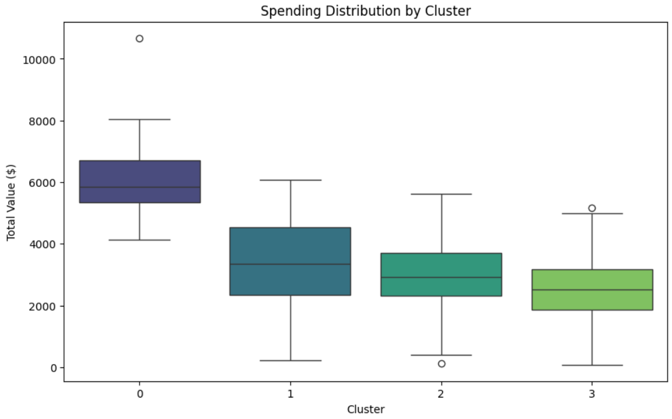
   i. Shows how customers are distributed across the first two principal components.
   ii. Red centroid markers indicate the center of each cluster.
   iii. Clusters are well-separated but show some overlap, supporting the moderate silhouette score.

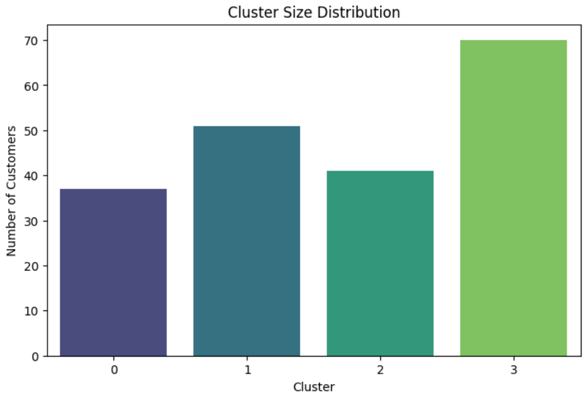
2. **Spending Distribution by Cluster (Box Plot)**

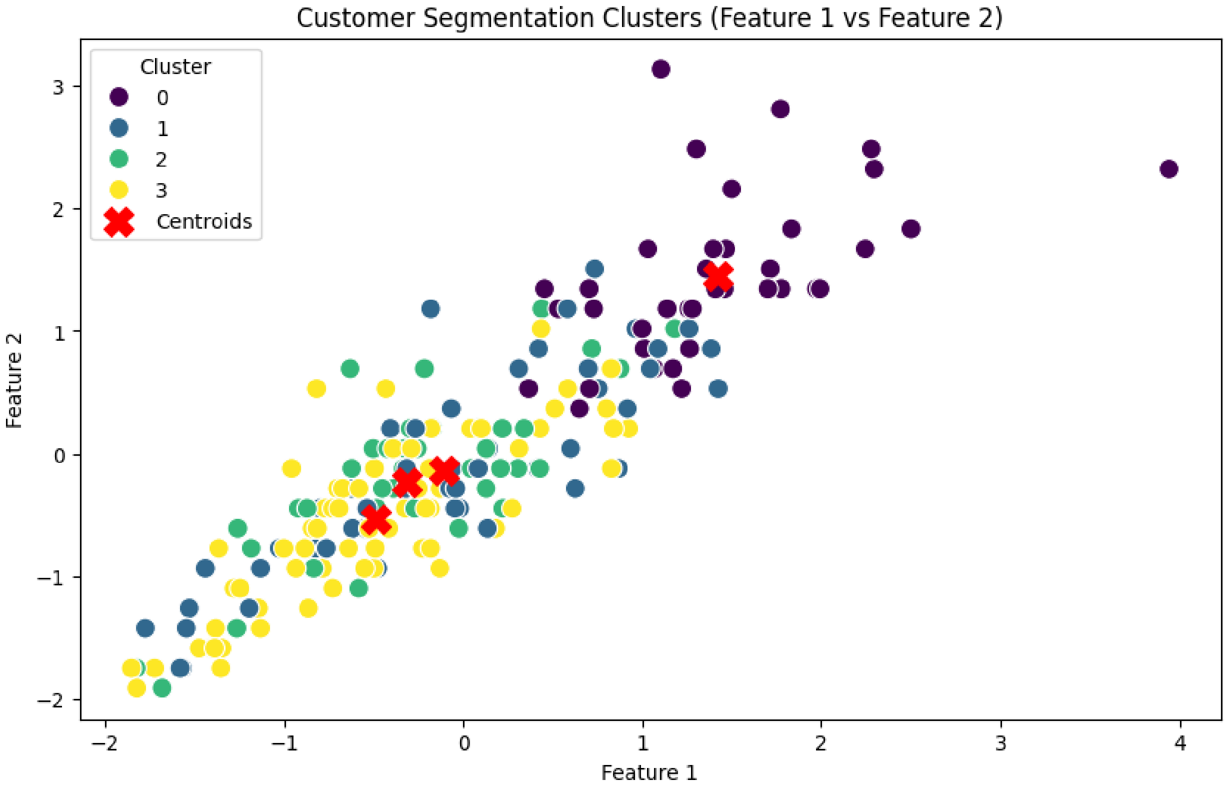
   i. Customers in Cluster 0 spend the most, as their median Total Value is the highest.
   ii. Clusters 2 and 3 have lower spending patterns, suggesting a segment of budget-conscious customers.
   iii. Some high-spending outliers exist in each cluster.

3. **Cluster Size Distribution (Bar Chart)**

   i. Cluster 3 is the largest (70 customers), meaning many customers share similar spending patterns.
   ii. The other clusters are more balanced, with sizes ranging from 37 to 51 customers.







- **Key Findings & Business Recommendations:**

    i. High-spending clusters (e.g., Cluster 0) should be targeted for premium offers, loyalty programs, and exclusive discounts.
    ii. Smaller clusters (e.g., Cluster 2) could benefit from promotional campaigns to boost engagement.
    iii. Cluster 3, being the largest, represents the most common customer type, requiring a balanced strategy of retention and growth.
    iv. Further refinement using additional features (e.g., product preferences) could improve segmentation accuracy.