Data Science Intern Assignment | Zeotap

Task 3: Customer Segmentation / Clustering

Github- https://github.com/PriyabrataBehera-24/eCommerce_Transactions

Github code- https://github.com/PriyabrataBehera-24/eCommerce_Transactions/blob/main/Priyabrata_Behera_Clustering.ipynb

1. Introduction

This report outlines the results of a customer segmentation analysis performed using clustering techniques. The analysis incorporates both customer profile information (from **Customers.csv**) and transaction details (from **Transactions.csv**) to identify meaningful customer groups. The primary objective is to gain deeper insights into customer behavior and demographics, enabling targeted marketing strategies and improved customer engagement.

2. Methodology

2.1 Data Preparation

- The **Customers.csv** and **Transactions.csv** datasets were merged using the CustomerID column to create a comprehensive dataset combining demographic information and purchasing behavior.
- Feature engineering was performed to calculate relevant features, such as:
  - **Purchase Frequency:** Total number of transactions made by each customer.
  - **Total Spending (TotalValue):** Aggregated value of all transactions for each customer.

2.2 Feature Selection

The features selected for clustering were:

- **TotalValue:** Represents a customer's total spending.
- **PurchaseFrequency:** Represents the number of transactions made by a customer.

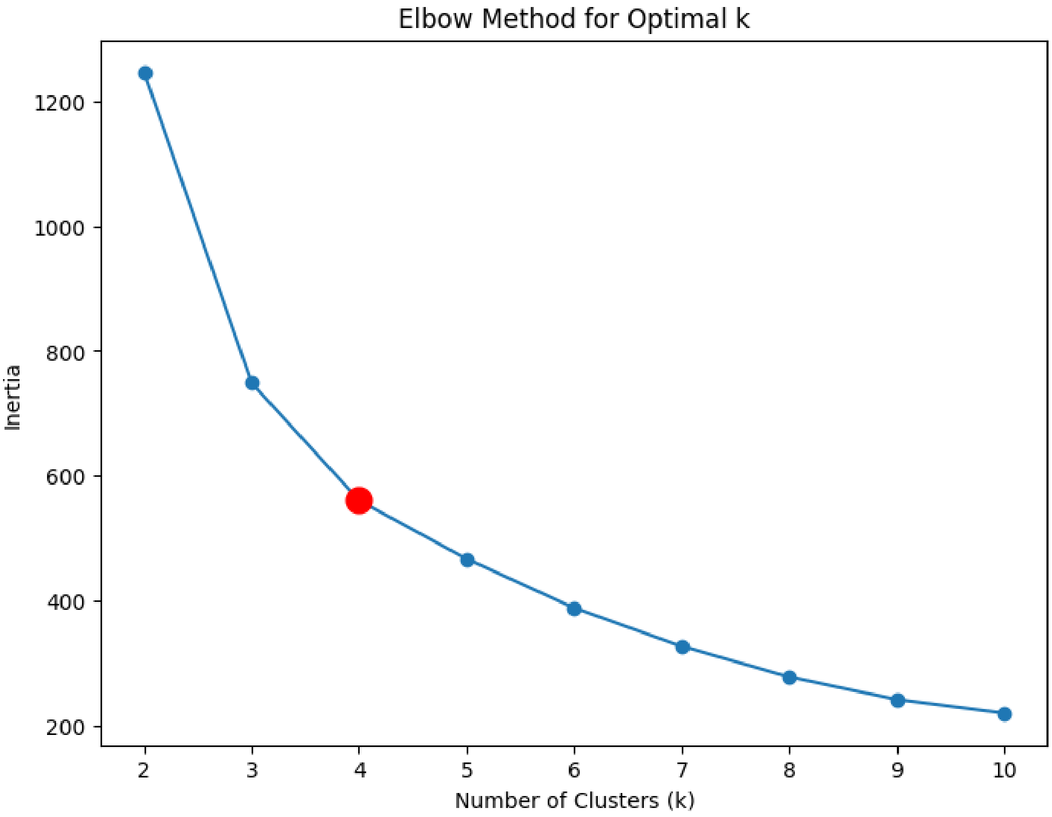
These features capture core dimensions of customer behavior: spending levels and engagement with the business.

2.3 Data Scaling

- Features were standardized using **StandardScaler** to ensure equal contribution to clustering.
- This process transforms the data to have a mean of zero and a standard deviation of one, avoiding biases caused by varying feature scales.

2.4 Clustering Algorithm and Cluster Number Selection

- The **K-Means clustering algorithm** was selected for its effectiveness and interpretability.
- To determine the optimal number of clusters (kkk), the **Elbow Method** was employed. This technique evaluates the within-cluster sum of squares (inertia) for different kkk values and identifies the "elbow point," where inertia reduction slows significantly. The optimal kkk was determined to be **4**.



2.5 Cluster Analysis and Evaluation

- Clustering metrics were used to assess the quality of the results:
  - **Davies-Bouldin Index (DB Index):** Measures cluster compactness and separation (lower is better). Value: **0.89**.

  - **Silhouette Score:** Measures how distinct and well-separated clusters are (higher is better). Value: **0.55**.

---

3. Results

3.1 Number of Clusters Formed: 4

3.2 Cluster Profiling

Each cluster was profiled based on the average values of the selected features (**TotalValue** and **PurchaseFrequency**), providing insights into customer segments:

Cluster	Average TotalValue	Average PurchaseFrequency	Characteristics
0	706.86	4.41	Low spenders with minimal transactions.
1	1,736.53	8.14	Moderate spenders with occasional engagement.
2	3,523.59	13.03	High spenders with frequent transactions.
3	6,733.70	20.22	Premium customers with the highest spending and engagement.

3.3 Cluster Visualization

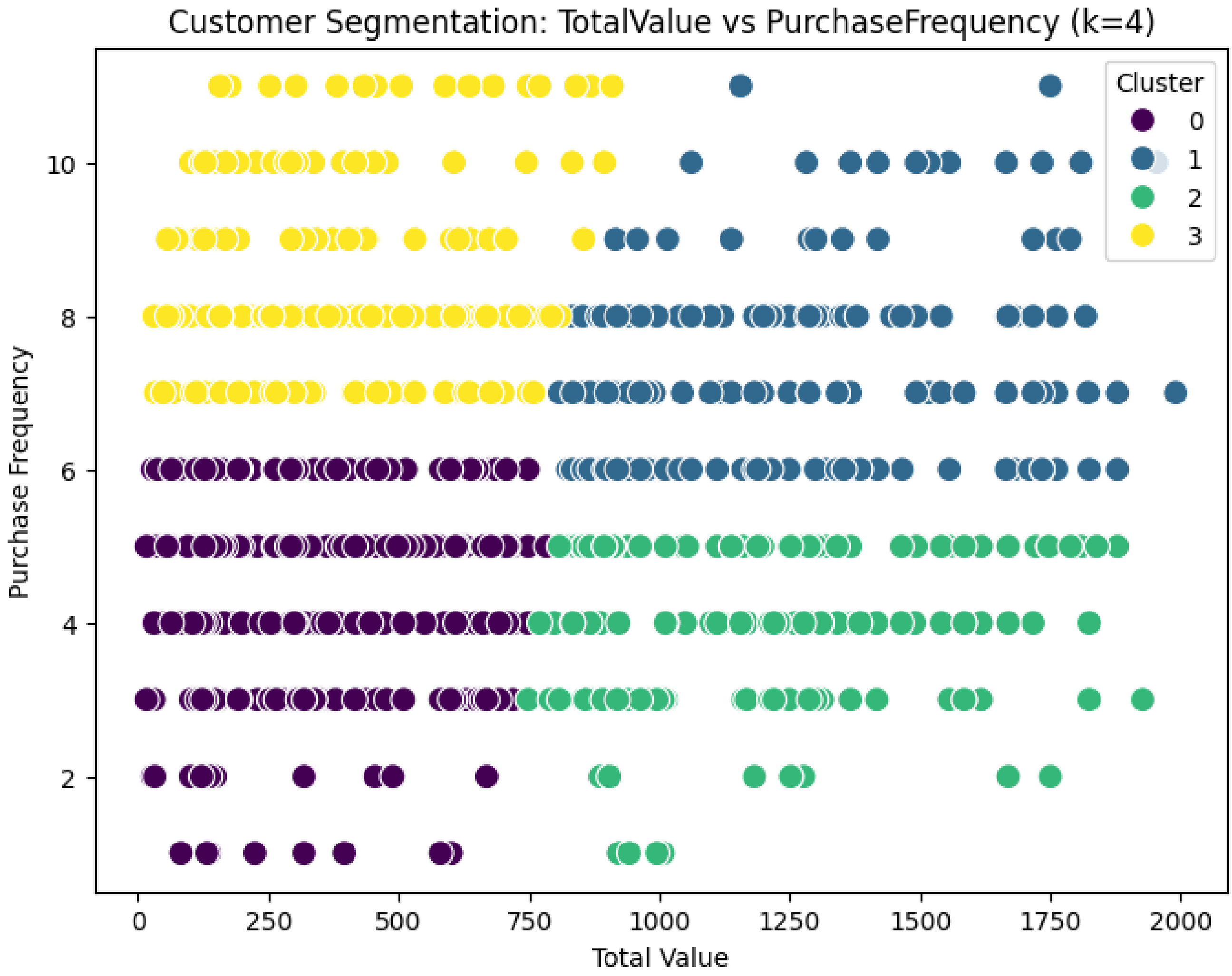
- **Scatter Plot:** Displays TotalValue vs. PurchaseFrequency, illustrating clear separation of clusters based on spending and frequency.
- **Pairplot:** Highlights pairwise relationships between features, emphasizing distinct cluster groupings.
- **Box Plots:** Show the distribution of TotalValue and PurchaseFrequency within each cluster, providing insights into variance and central tendencies.

Heatmap: Shows feature correlations, offering additional insights into the relationship between TotalValue and PurchaseFrequency.

---

4. Conclusion

Distinct Customer Segments Identified: The analysis identified four customer segments, each with unique spending and engagement patterns.



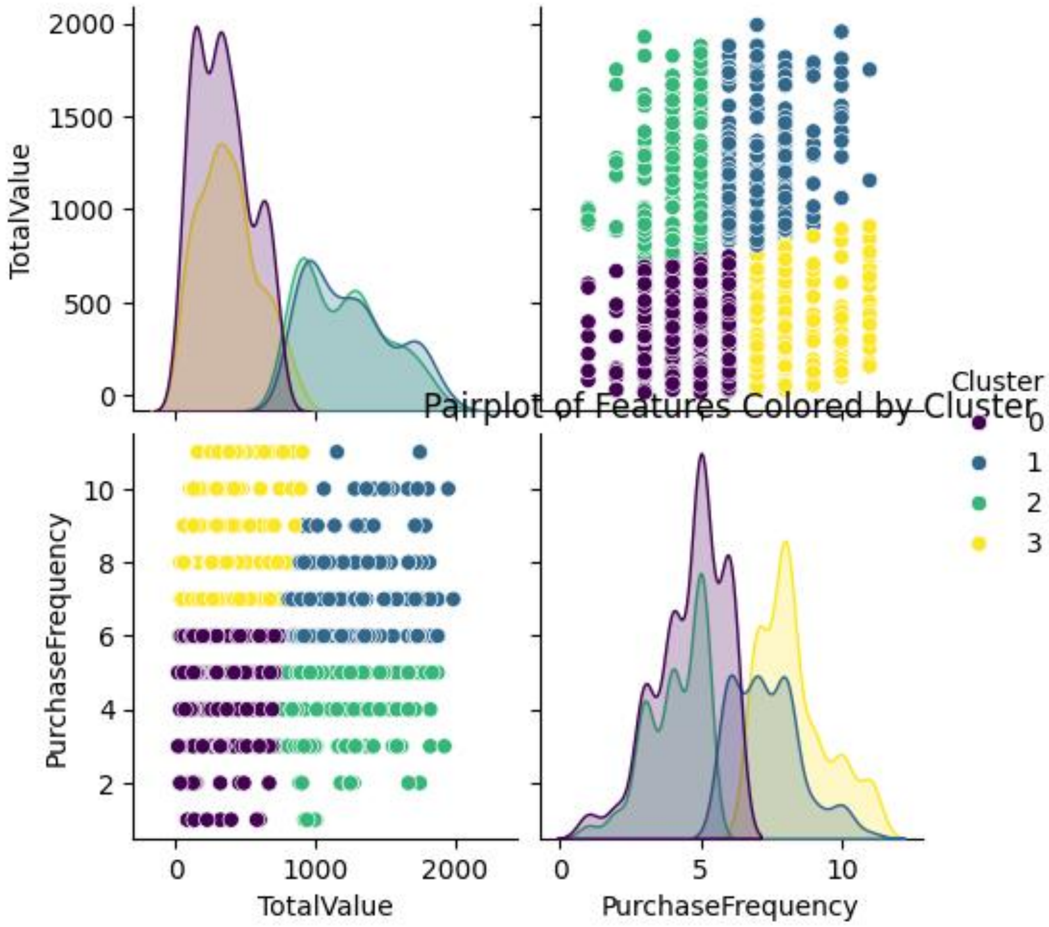
- **Clustering Metrics:**

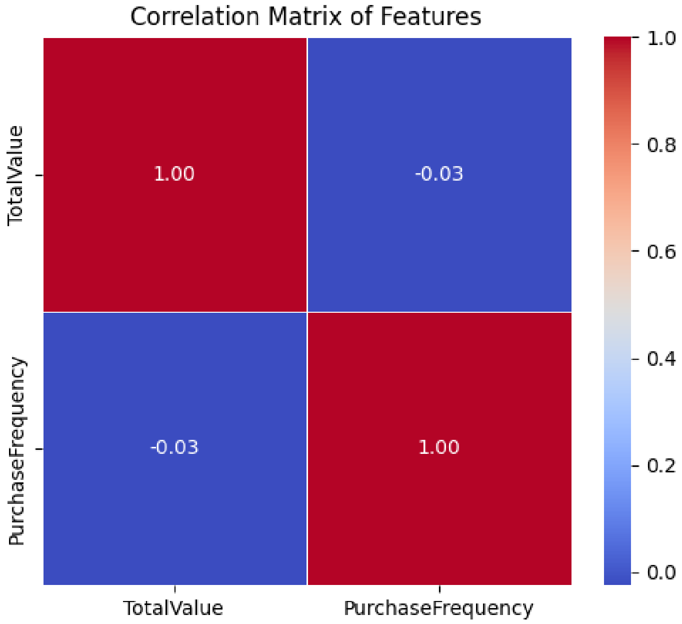
1. The **Davies-Bouldin Index (0.89)** and **Silhouette Score (0.55)** suggest reasonable clustering quality, though improvements may be possible.

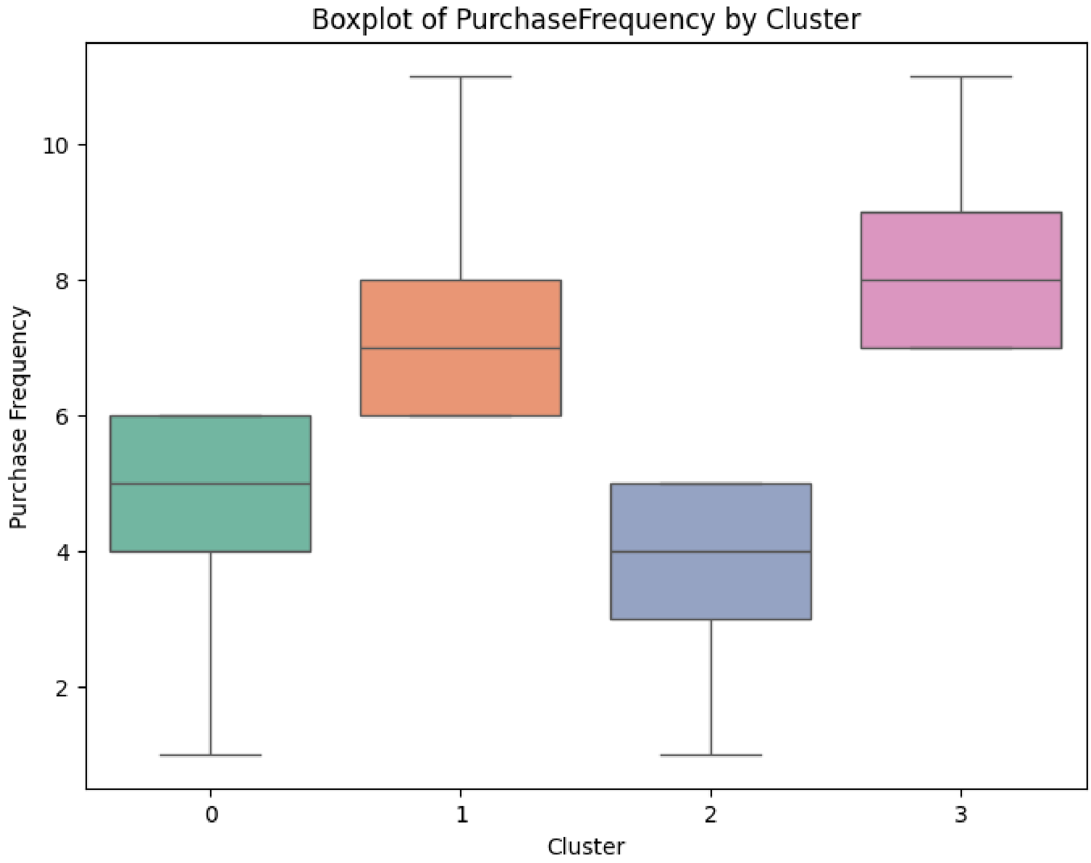
- **Actionable Insights:**

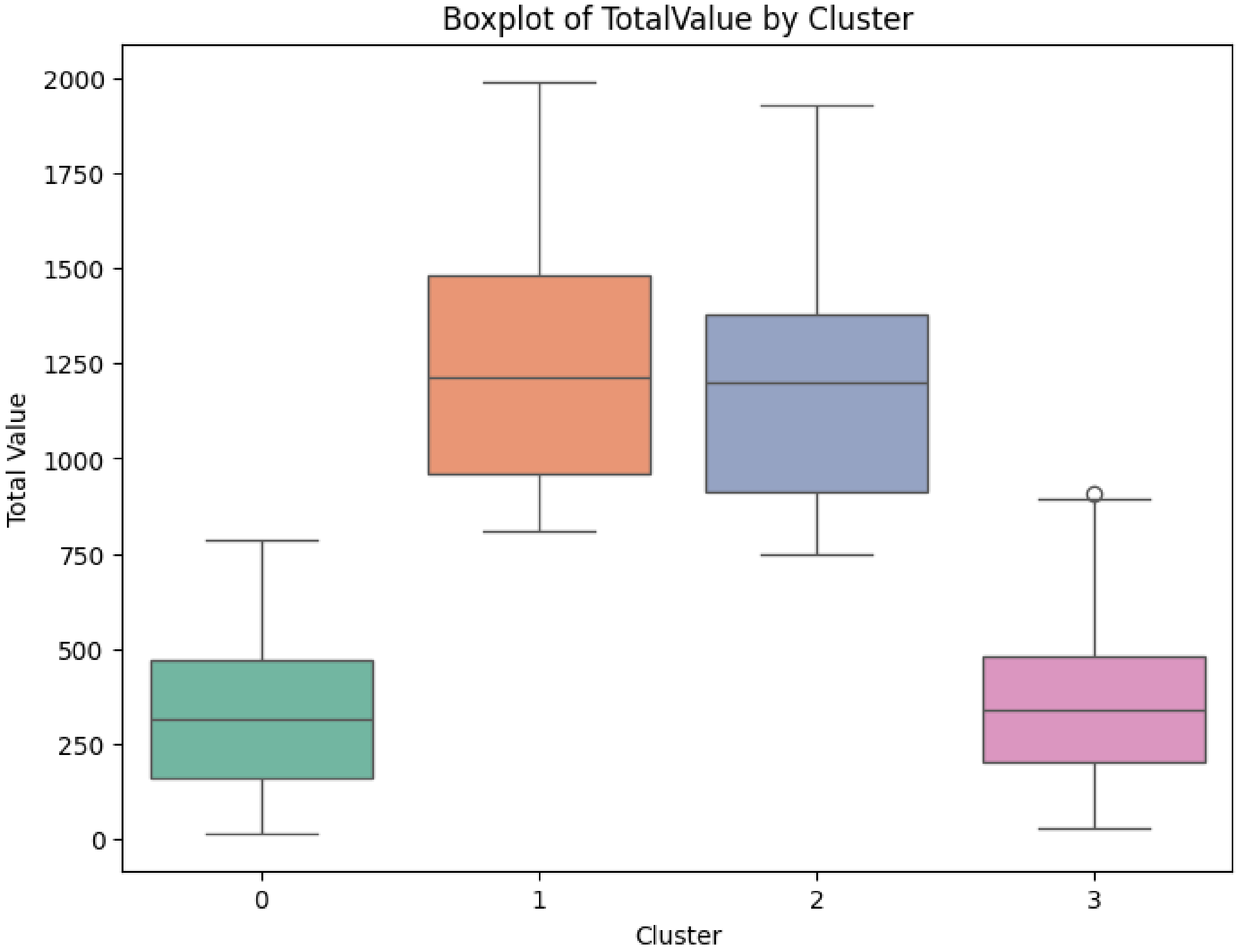
- **Cluster 3 (Premium Customers):** These customers contribute the most revenue and should be targeted with loyalty programs and premium offerings.
- **Cluster 0 (Low Spenders):** Strategies such as discounts or introductory offers could encourage increased engagement.
- **Cluster 1 & 2:** Moderate and high spenders could be nurtured into premium segments through personalized marketing campaigns.

The segmentation results provide a foundation for targeted marketing, personalized promotions, and efficient resource allocation. Further interpretation and strategic implementation of these insights can drive better business outcomes.






Boxplot of PurchaseFrequency by Cluster
Purchase Frequency
10
8
6
4
2
0
1
2
3
Cluster


Boxplot of TotalValue by Cluster
Total Value
2000
1750
1500
1250
1000
750
500
250
0
0
1
2
3
Cluster