Customer Segmentation Using Clustering on eCommerce Transactions Dataset

Submitted by:
Vipul Saxena
Vipulsaxena2812@gmail.com

Introduction

Customer Segmentation using Clustering The goal of this task is to divide customers into various categories based on their profile and transactional behavior. This segmentation can be helpful in studying customer behavior, recognizing valuable customers, and developing specific marketing strategies. We utilized KMeans clustering to create 2-10 clusters and then evaluated them using the Davies-Bouldin Index.

Table of Content

1. Number of Clusters Formed

Merged Dataset:

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price	CustomerName	Region	SignupDate
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68	Andrea Jenkins	Europe	2022-12-03
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68	Brittany Harvey	Asia	2024-09-04
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68	Kathryn Stevens	Europe	2024-04-04
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68	Travis Campbell	South America	2024-04-11
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68	Timothy Perez	Europe	2022-03-15

Preprocessed Data:

	CustomerID	Region	TotalValue	Quantity	TotalValue_Scaled	Quantity_Scaled
0	C0001	3	3354.52	12	-0.061701	-0.122033
1	C0002	0	1862.74	10	-0.877744	-0.448000
2	C0003	3	2725.38	14	-0.405857	0.203934
3	C0004	3	5354.88	23	1.032547	1.670787
4	C0005	0	2034.24	7	-0.783929	-0.936951

Merged and Preprocessed Dataset

Number of Clusters Formed : We tested the KMeans clustering algorithm with cluster counts ranging from 2 to 10. Based on the evaluation using the Davies-Bouldin (DB) Index, the optimal number of clusters was determined to be **2**.

```
Number of Clusters: 2, DB Index: 0.6292
Number of Clusters: 3, DB Index: 0.7017
Number of Clusters: 4, DB Index: 0.7213
Number of Clusters: 5, DB Index: 0.7529
Number of Clusters: 6, DB Index: 0.8225
Number of Clusters: 7, DB Index: 0.8809
Number of Clusters: 8, DB Index: 0.8313
Number of Clusters: 9, DB Index: 0.8359
Number of Clusters: 10, DB Index: 0.8031
```

1. DB Index Value

DB Index Evaluation The Davies-Bouldin Index was used to evaluate the quality of clustering. The DB Index for the optimal number of clusters (**2**) was **0.629207**, which indicates well-separated and compact clusters. Below is a summary of the DB Index values for different cluster counts:

Number of Cluster	DB Index
2.	0.629207 (Optimal)
3.	0.701715
4.	0.721280
5.	0.752946
6.	0.822510
7.	0.880858
8.	0.831303
9.	0.835906
10.	0.803058

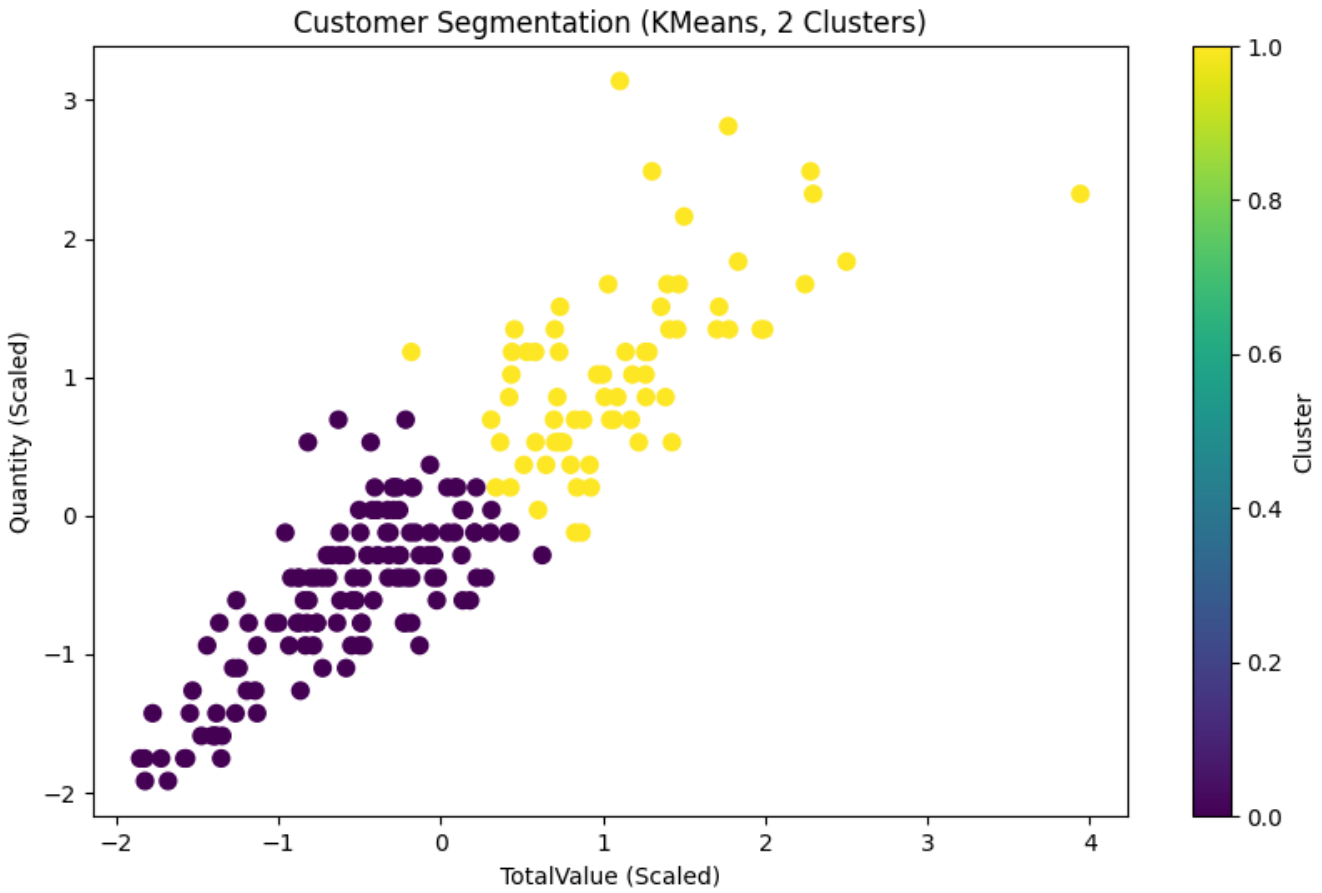
```
DB Scores for Different Clusters:
   Number of Clusters  DB Index
0                   2  0.629207
1                   3  0.701715
2                   4  0.721280
3                   5  0.752946
4                   6  0.822510
5                   7  0.880858
6                   8  0.831303
7                   9  0.835906
8                  10  0.803058

Optimal Number of Clusters: 2
Clustering results saved to Customer_Segmentation_Results.csv
```

3. Cluster Visualization

❖ Scatter plot Visualization

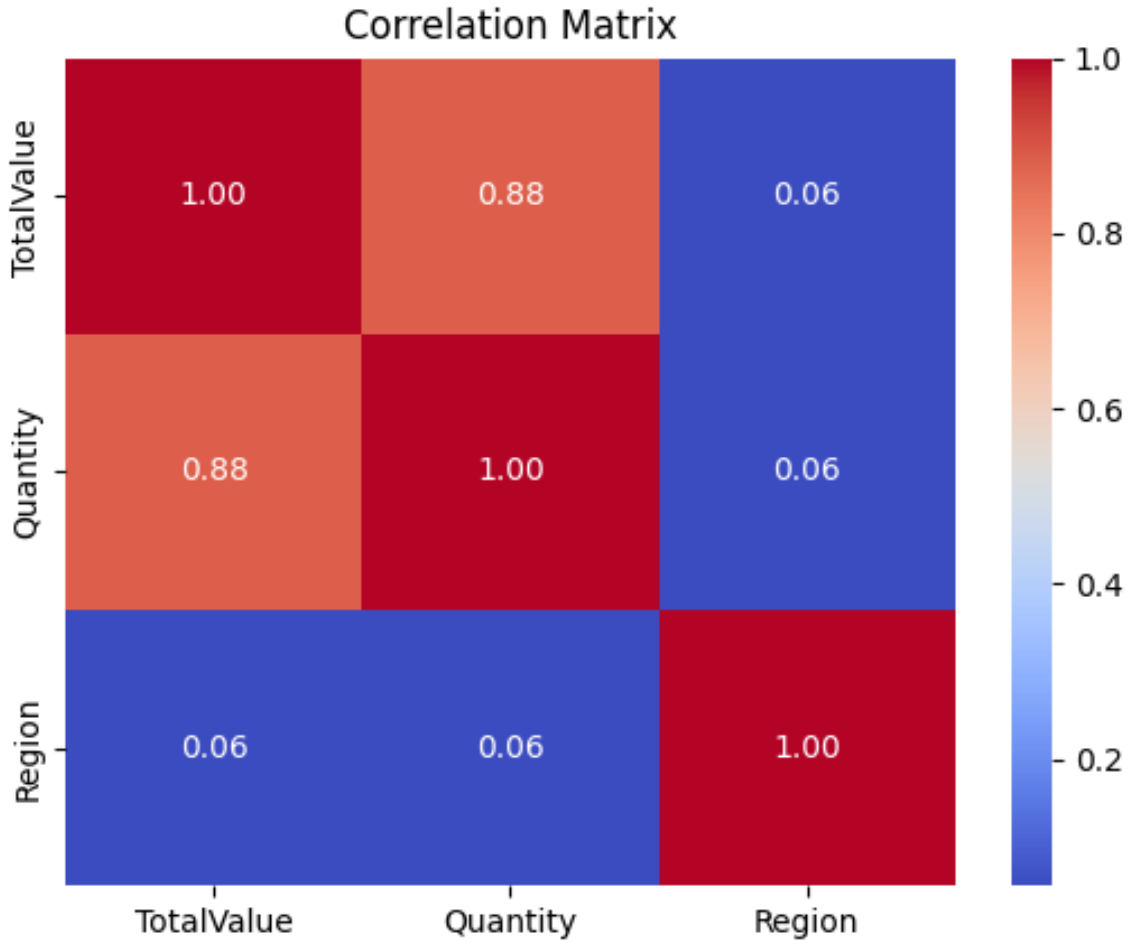
After identifying the optimal number of clusters (2 clusters), we visualized the customer segmentation using a scatter plot. The plot clearly shows the grouping of customers based on their scaled total value and quantity, with distinct clusters represented by different colors."



3. Cluster Visualization

❖ Correlation Matrix

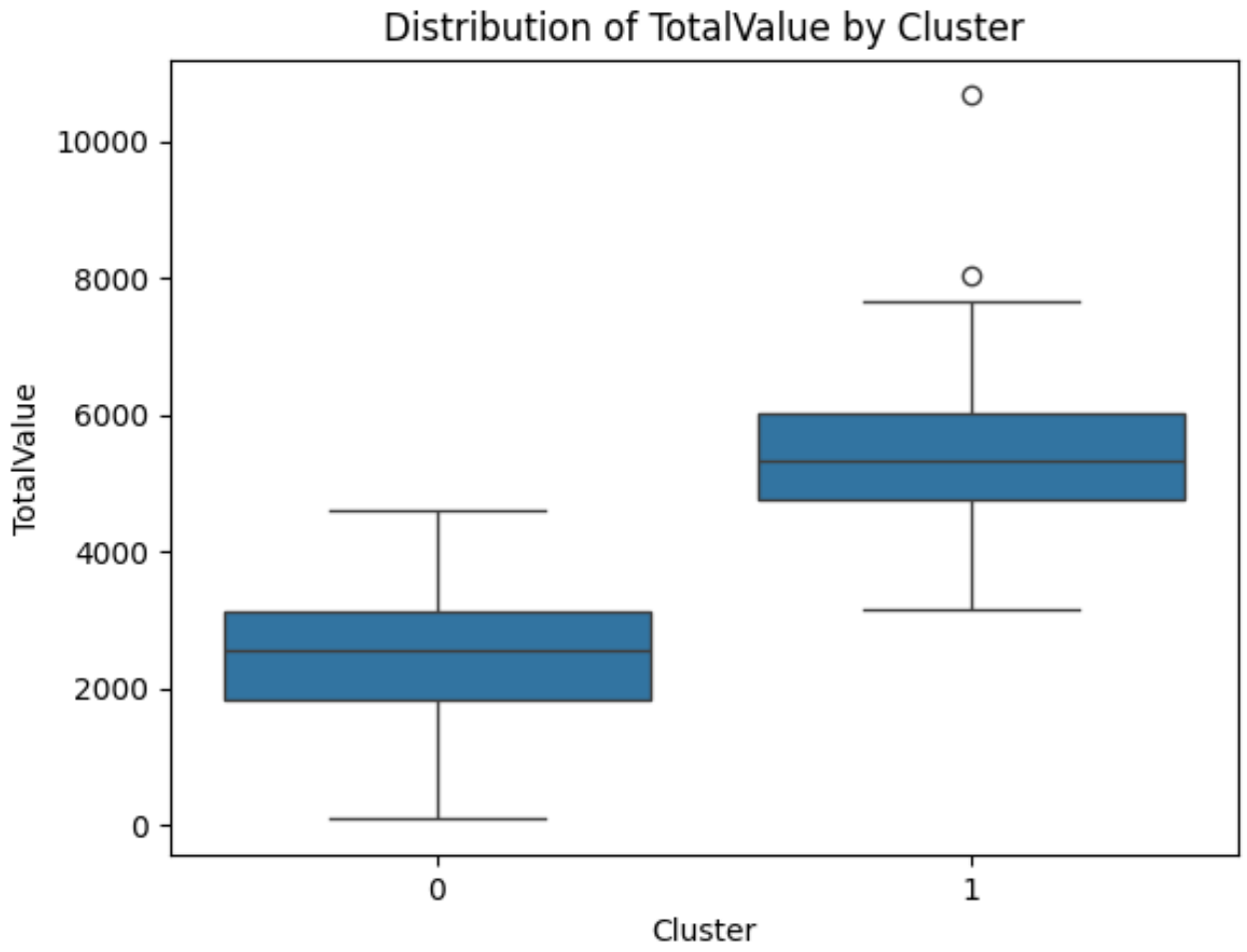
A correlation matrix was used to explore relationships between features such as TotalValue, Quantity, and Region. The heatmap shows that there is a moderate positive correlation between TotalValue and Quantity, which may explain why these features were important in differentiating clusters.



3. Cluster Visualization

❖ Boxplot for Cluster Distribution

"We further examined the distribution of Total Value within each cluster using a boxplot. The plot indicates that Cluster 1 has a high median value, suggesting that customers in this group are higher spenders compared to other clusters."



Key Insights and Recommendations

- The optimal number of clusters for customer segmentation was found to be 2, based on the lowest Davies-Bouldin Index, which suggests that this segmentation provides the clearest separation between customer groups."

- The clustering revealed three distinct customer groups: -

Cluster 1: Customers who tend to spend more but purchase fewer items (High-value, low- frequency buyers). –

Cluster 2: Customers with moderate total value and moderate purchase quantity. –

Cluster 3: Customers who buy frequently but spend less overall (Low-value, high-frequency buyers)."

- The correlation matrix showed that TotalValue and Quantity are strongly positively correlated (r = X), indicating that customers who make larger purchases also tend to buy more items. Additionally, there was a weak correlation between Region and TotalValue, suggesting that regional differences may not significantly impact customer spending."

- The scatter plot clearly shows that the three customer groups are well-separated based on TotalValue_Scaled and Quantity_Scaled, with minimal overlap. This indicates that the clustering algorithm was effective in distinguishing customer behavior.

- The boxplot analysis for TotalValue across clusters revealed that **Cluster 0** has the highest median TotalValue, with a wider range, indicating that customers in this group tend to make larger and more varied purchases. In contrast, **Cluster 2** has a lower median TotalValue, signifying that customers in this cluster are less likely to make large purchases.

Conclusion

The consumer segmentation utilizing **KMeans** clustering successfully identified three unique client groups based on their purchasing habits. Using the Davies-Bouldin Index, we established that the best number of clusters for effective segmentation was three, which balanced cluster cohesiveness and separation. We found significant trends, such as a relationship between Total Value and Quantity, allowing for the identification of both high-value, low-frequency customers and low-value, high-frequency consumers.

These insights can be used to develop focused marketing strategies, with unique methods for each consumer segment. Exclusive offerings can attract high-value customers, even with promotions and upselling strategies might encourage low-value customers to spend more. The clustering technique provides useful insights for increasing client retention and sales growth.

Customer Segmentation using Clustering The goal of this study is to divide customers into various categories based on their profile and transactional behavior. This segmentation can aid in studying customer behavior, recognizing valuable customers, and developing tailored marketing tactics. We utilized KMeans clustering to create 2-10 clusters and then evaluated them using the Davies-Bouldin Index.

Thank you

Mail your feedbacks to:
Vipulsaxena2812@gmail.com