Customer Segmentation Report

1. Introduction

The goal of this task was to perform customer segmentation by applying clustering techniques on customer profile and transaction data. The dataset consisted of customer details, product information, and transactional data. The objective was to identify distinct customer segments based on both their demographic profile and transaction history, which could be useful for targeted marketing strategies.

2. Data Preprocessing and Feature Engineering

Before performing clustering, the following steps were taken:

- **Data Merging**: Merged Customers.csv, Transactions.csv, and Products.csv based on common identifiers (CustomerID and ProductID).
- **Feature Engineering**: Aggregated the transactional data by customer, including features such as:
    - Total amount spent (total_spent)
    - Total quantity purchased (total_quantity)
    - Average transaction value (avg_transaction_value)
    - Number of transactions (num_transactions)
    - Most purchased product category (most_purchased_category)

These features were used to represent the customer behavior and profile.

- **Scaling**: Standardization of numerical features was performed using StandardScaler to ensure the clustering algorithm treated all features equally, preventing any feature from dominating due to different scales.

3. Clustering Approach

3.1 Clustering Algorithm Used

We used the **K-Means** clustering algorithm due to its efficiency and ability to handle large datasets. The K-Means algorithm partitions the data into a predefined number of clusters by minimizing the within-cluster variance.

3.2 Number of Clusters

We evaluated various values for k (the number of clusters) ranging from 2 to 10 using clustering evaluation metrics. The optimal number of clusters was selected based on the lowest **Davies-**

Bouldin Index and highest **Silhouette Score**. The final clustering was performed using k=10 clusters, which provided the best trade-off in terms of cluster cohesion and separation.

4. Evaluation Metrics

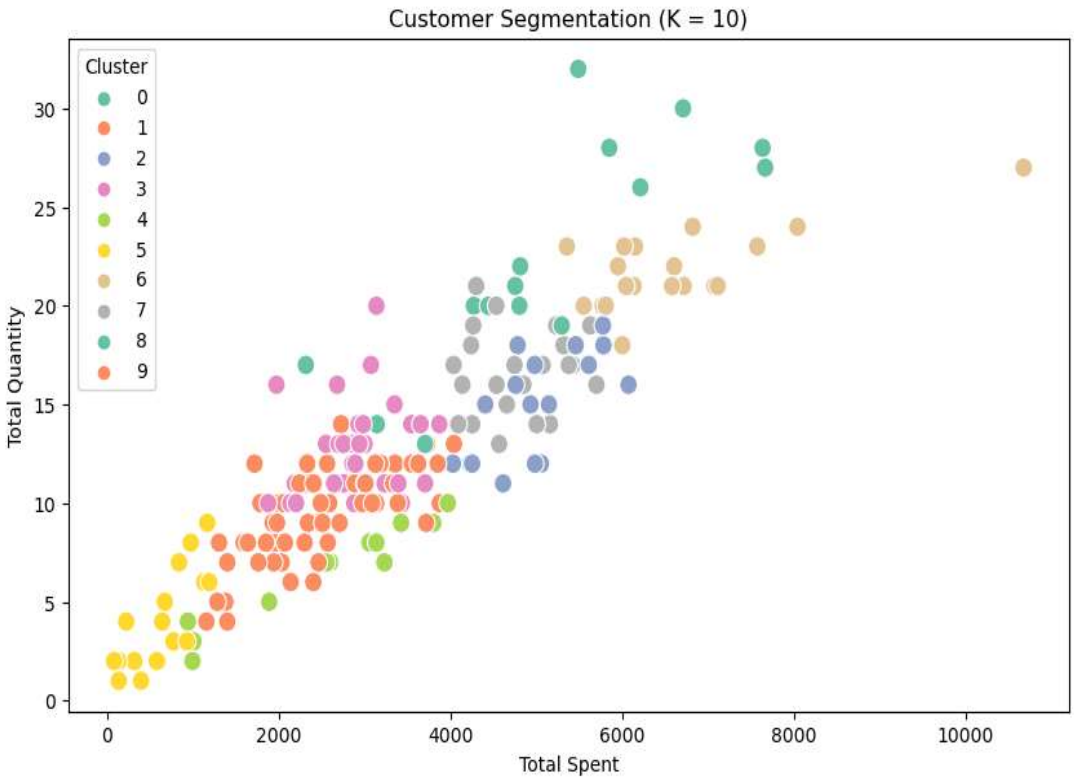
4.1 Davies-Bouldin Index (DB Index)

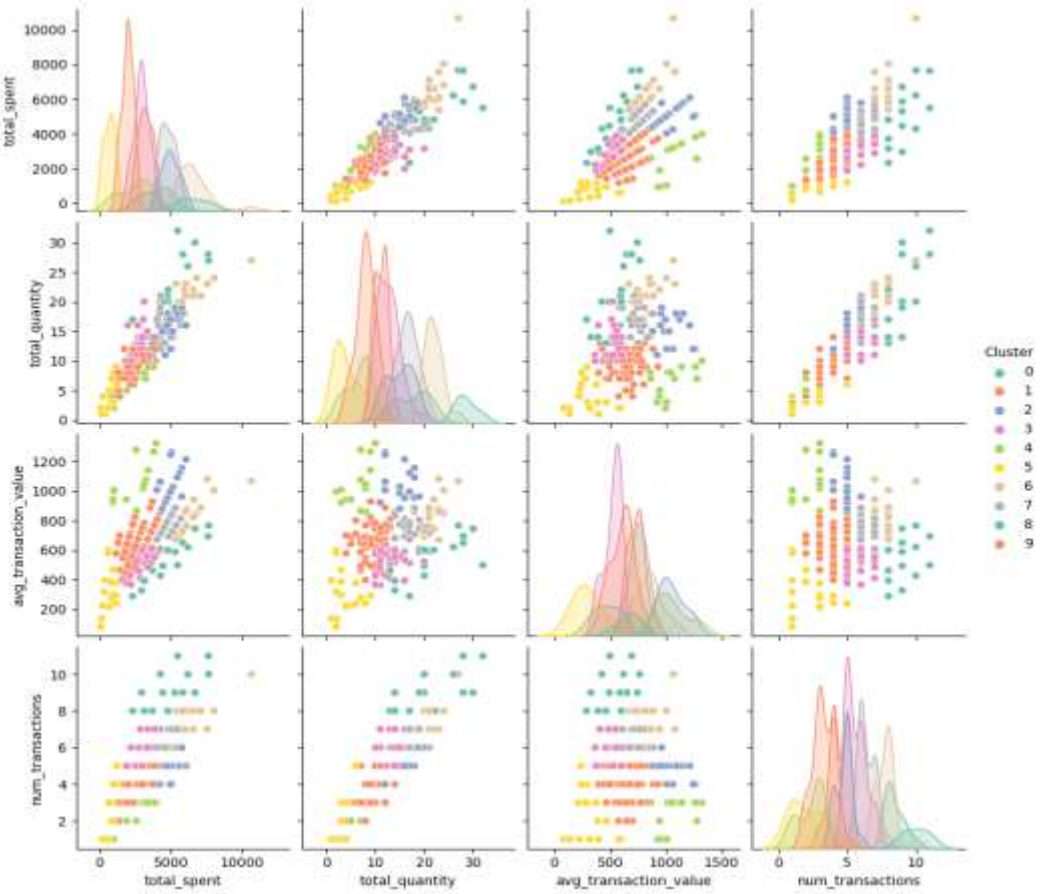
The **Davies-Bouldin Index** (DB Index) is a measure of clustering quality. Lower values indicate better clustering, where the clusters are well-separated and compact. The DB Index for the selected number of clusters (k=10) was **0.91**, suggesting that the clusters are fairly distinct and well-separated.

4.2 Silhouette Score

The **Silhouette Score** measures how similar a point is to its own cluster compared to other clusters. A score closer to 1 indicates well-separated and clearly defined clusters. The Silhouette Score for k=10 was **0.30**, indicating that the clusters are moderately well-formed but still some overlap exists.

5. Visualization





6. Conclusion

In summary, K-Means clustering with k=10 provided meaningful customer segments based on spending behavior and product preferences. The clusters can be used for targeted marketing, where each group can be offered personalized product recommendations or promotional campaigns.

- **DB Index**: 0.91
- **Silhouette Score**: 0.30