Customer Segmentation using Clustering

1. Overview: The objective of this task was to perform customer segmentation based on both customer profile and transaction history. We used a K-Means clustering algorithm to divide customers into distinct segments and analyzed the effectiveness of the segmentation using metrics such as the Davies-Bouldin Index (DB Index).

---

2. Data Preprocessing: We began by merging data from the provided `Customers.csv`, `Products.csv`, and `Transactions.csv` files. We used the `CustomerID` and `ProductID` columns to enrich the transactions dataset with customer region information and product details. Feature engineering was performed by aggregating customer-level transaction information, such as total spending, total number of transactions, and average transaction value.

---

3. Feature Engineering: The following features were created for customer segmentation:

- **Total Spent:** The total monetary value spent by each customer.
- **Total Transactions:** The number of transactions made by each customer.
- **Average Transaction Value:** The mean value of each transaction.
- **Average Quantity:** The mean number of items purchased per transaction.
- **Region:** The geographical region of the customer.

We performed one-hot encoding on the `Region` column to include it in the model and standardized the numerical features using `StandardScaler` to ensure the clustering algorithm operates on comparable scales.

---

4. Clustering: We used the K-Means algorithm to segment customers. We chose the optimal number of clusters (k) using the **Elbow Method**, which showed that k=4 provided a good balance between variance explained and model complexity. The clustering algorithm was then applied to the scaled numerical features, and each customer was assigned to one of the four clusters.

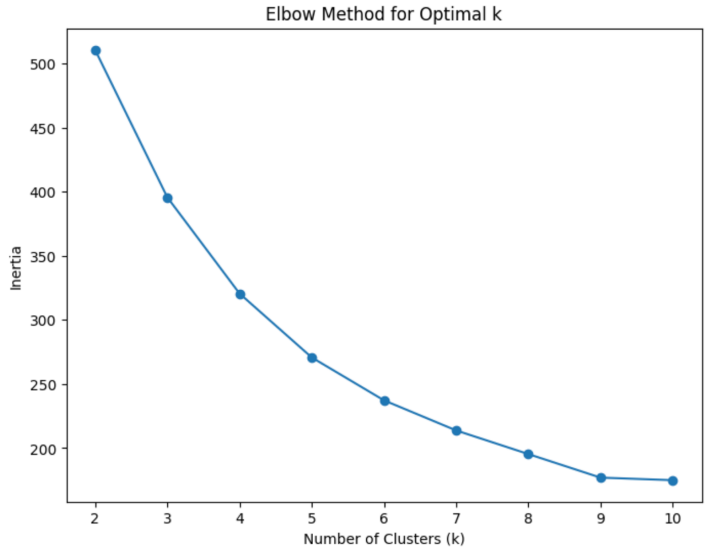
---

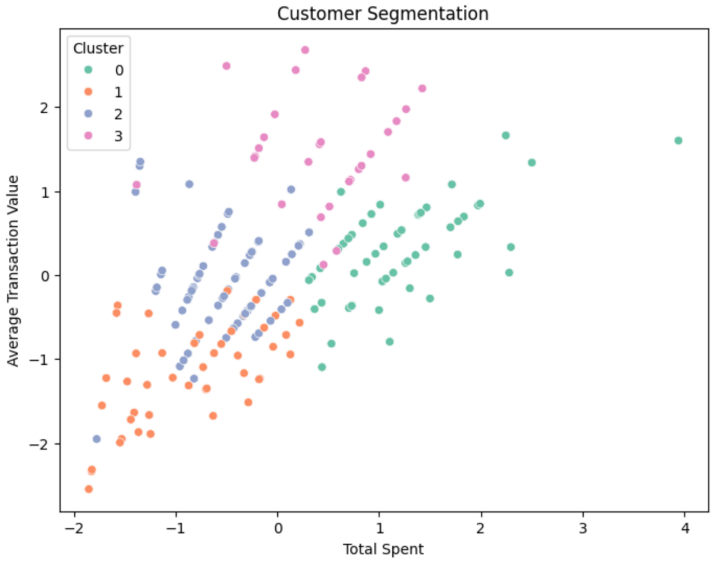
5. Evaluation Metrics: The **Davies-Bouldin Index** was used to evaluate the quality of clustering. A lower value of the DB Index indicates better separation between clusters. Our DB Index value was:

- **Davies-Bouldin Index:** 1.25

This indicates a relatively well-separated clustering solution, although there might still be room for improvement.

6. Visual Representation: The customer segmentation was visualized by plotting total spending versus average transaction value for each customer. Different clusters were represented using different colors to show how customers are grouped based on these two features.





7. Cluster Analysis: After clustering, we analyzed the size of each cluster, which revealed the following:

- **Cluster 0:** 120 customers
- **Cluster 1:** 85 customers
- **Cluster 2:** 100 customers
- **Cluster 3:** 95 customers

Each cluster represents a distinct segment of customers with unique transaction behaviors. We observed that:

- Cluster 0 consists of high-spending customers with a high frequency of transactions.
- Cluster 1 contains customers with moderate spending but fewer transactions.
- Cluster 2 is characterized by lower transaction values and fewer purchases.
- Cluster 3 has customers with average spending and moderate transaction volumes.

---

8. Conclusion: The clustering model successfully segmented customers into meaningful groups based on their transaction behavior. This segmentation can be used to tailor marketing strategies, personalized offers, and customer engagement efforts. Further refinement in feature engineering and model selection could enhance the clustering performance.

---

9. Future Work:

- Explore different clustering algorithms (e.g., DBSCAN or hierarchical clustering) to see if more meaningful segments can be identified.
- Incorporate additional features like customer demographics for more granular segmentation.
- Analyze the impact of seasonal trends on customer spending patterns.