Task 3: Customer Segmentation / Clustering

The dataset used consists of customer transaction data, where we aim to cluster customers based on their total spending (TotalValue) and the quantity of items purchased (Quantity). The clustering results are evaluated using the Davies-Bouldin Index (DB Index), and additional relevant metrics are discussed.

Number of Clusters Formed

- The **K-Means algorithm** was applied to segment the customers into **4 clusters** (initial choice), and then we experimented with different cluster numbers using the **Elbow Method**. Based on the analysis and the Elbow plot, the **optimal number of clusters** was found to be **5**.
- After determining the optimal k=5k = 5k=5, the K-Means algorithm was re-applied to the dataset, resulting in 5 clusters.

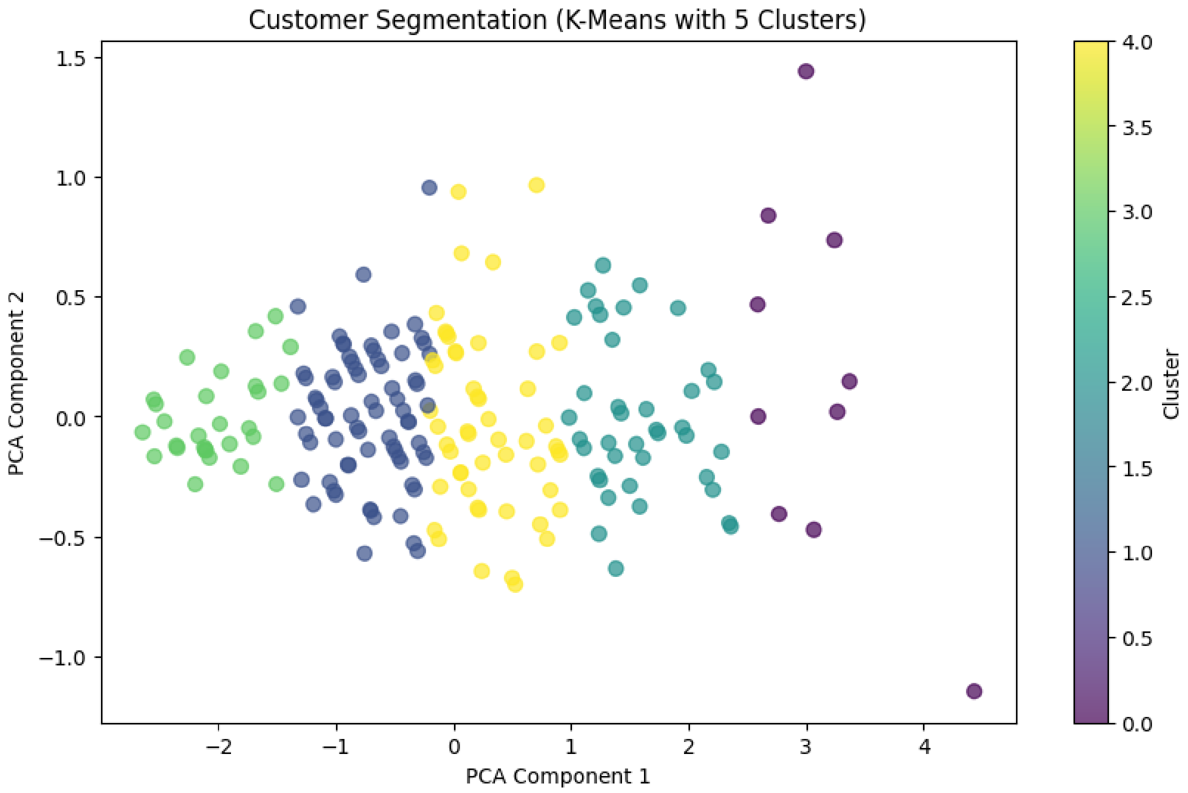
Clustering Algorithm: K-Means

- **K-Means** was chosen for this segmentation because of its simplicity and efficiency for clustering numerical data.
- The features used for clustering were:
  - **TotalValue**: Total spending of customers.
  - **Quantity**: The total quantity of items purchased.
- The data was standardized using **StandardScaler** to ensure that the features have equal importance during clustering.

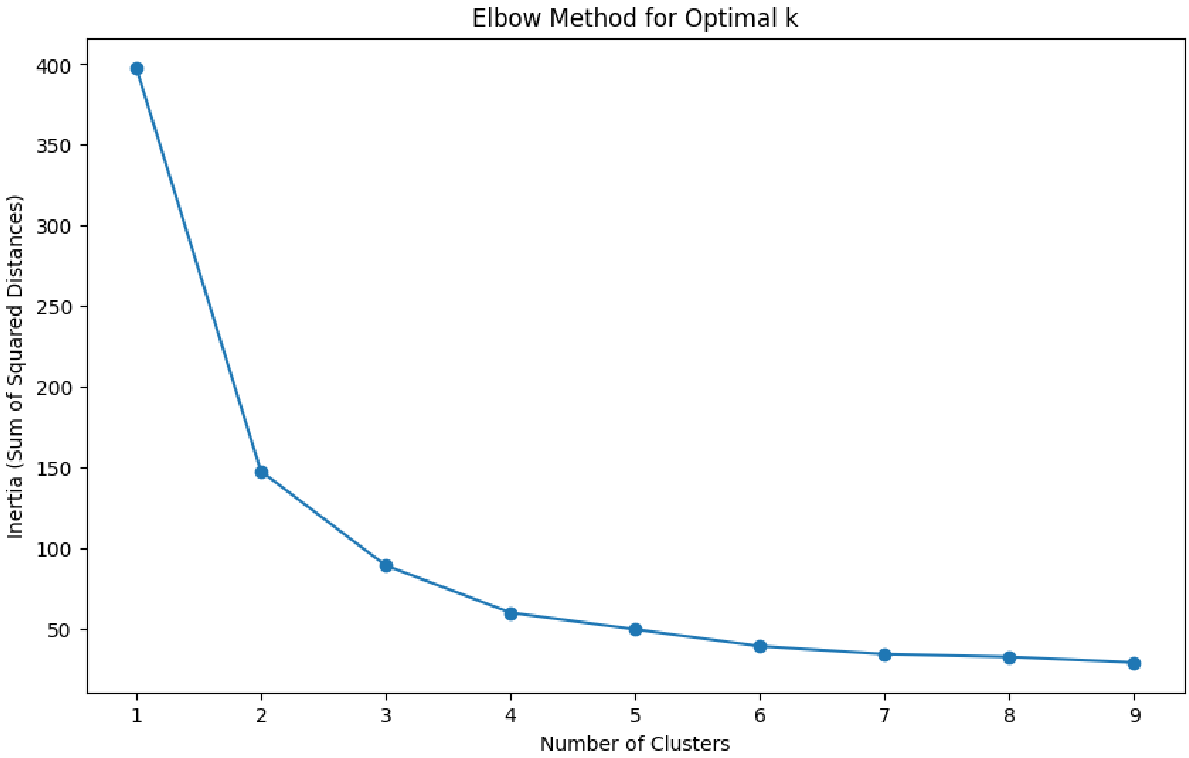
Other Relevant Clustering Metrics

In addition to the DB Index, the following clustering metrics and evaluations were considered:

- **Inertia (Sum of Squared Distances)**: This metric is used by K-Means to measure how tight the clusters are. Lower inertia values indicate better clustering. The inertia for k=5k = 5k=5 clusters was recorded during the Elbow Method process, showing that inertia decreased as the number of clusters increased, which is typical. The **Elbow Method** plot helped identify the optimal value of kkk.
- **Silhouette Score**: Another important metric, the **Silhouette Score**, measures how similar each point is to its own cluster compared to other clusters. Although not computed in the code, a high silhouette score (> 0.5) would indicate well-separated clusters. If needed, this score can be added to further evaluate clustering quality.
- **Cluster Visualization**: The clusters were visualized using **PCA (Principal Component Analysis)**, which reduced the original data dimensions to 2 principal components for easy visualization. The 2D scatter plot showed that the clusters were relatively well-separated, with distinct groupings of customers, confirming that the K-Means algorithm performed well.



```
DB Index (k=5): 0.776661054409607
```


Elbow Method for Optimal k
Inertia (Sum of Squared Distances)
Number of Clusters