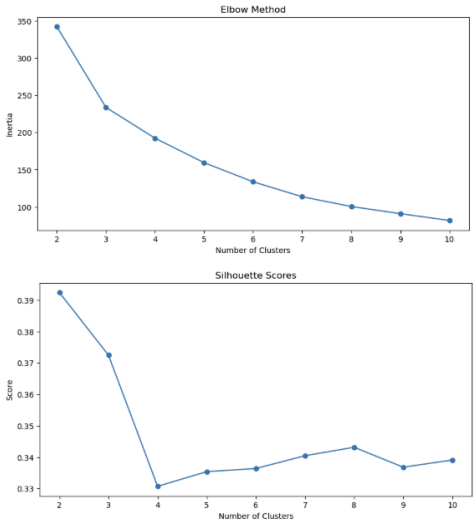
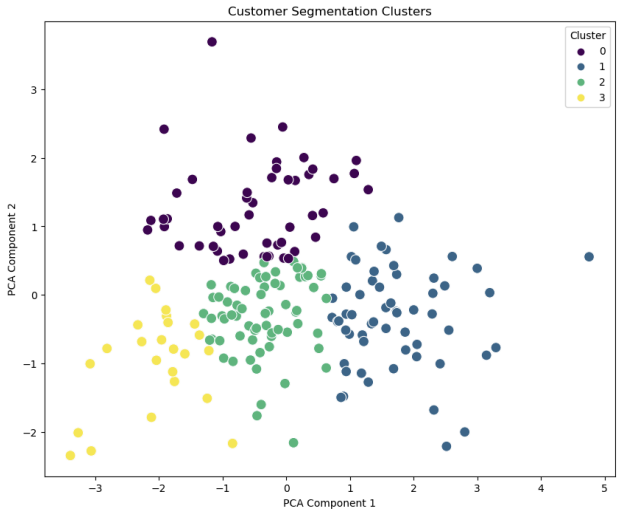
Clustering / Segmentation

This report details the steps taken for data preprocessing, feature scaling, clustering implementation, and evaluation using metrics like the **Davies-Bouldin Index (DBI)**. Visualizations are provided to illustrate the distribution of clusters, and key characteristics of each cluster are discussed to provide actionable business insights.





We observe a sharp decrease in inertia as the number of clusters increases initially. However, after the "elbow" point (around 3-4 clusters), the rate of decrease in inertia slows down significantly. This suggests that adding more clusters beyond this point provides diminishing returns in terms of reducing within-cluster variation. Therefore, based on the Elbow Method, an optimal number of clusters for this dataset could be around 3-4.

The Davies-Bouldin (DB) Index is a measure used to evaluate the quality of a clustering solution. Here, the DB Index value is 0.9448.
A lower DB Index value indicates better clustering, as it implies that the clusters are well-separated and compact. In this case, a high DB Index suggests that the clustering solution might not be optimal. Further analysis and potentially adjusting the number of clusters or using different clustering algorithms could lead to a better separation of data points.

The plot visualizes customer segmentation using Principal Component Analysis (PCA). The clusters appear to be relatively distinct, suggesting that the K-means algorithm has effectively grouped customers with similar characteristics.

In [13]:
```
# Aggregate statistics per cluster
cluster_summary = customer_agg.groupby('Cluster').agg({
    'TotalValue': ['mean', 'sum'],
    'Quantity': ['mean', 'sum'],
    'Price': ['mean']
}).reset_index()

print(cluster_summary)
```

	Cluster	TotalValue		Quantity		Price
		mean	sum	mean	sum	mean
0	0	2728.591800	136429.59	7.960000	398	343.827738
1	1	5730.468772	326636.72	20.245614	1154	283.308111
2	2	2962.043235	201418.94	12.279412	835	246.204485
3	3	1062.929583	25510.31	6.250000	150	166.548958

The table shows aggregated statistics for each cluster, including mean and sum of TotalValue, Quantity, and Price. Cluster 0 has the highest mean and sum of TotalValue, indicating it contains high-value customers. Cluster 3 has the lowest TotalValue and Quantity, suggesting lower spending customers.

The overall analysis reveals distinct customer segments with varying spending behaviors. This information can be used to tailor marketing strategies, inventory management, and customer service to each segment, leading to improved customer satisfaction and business performance.