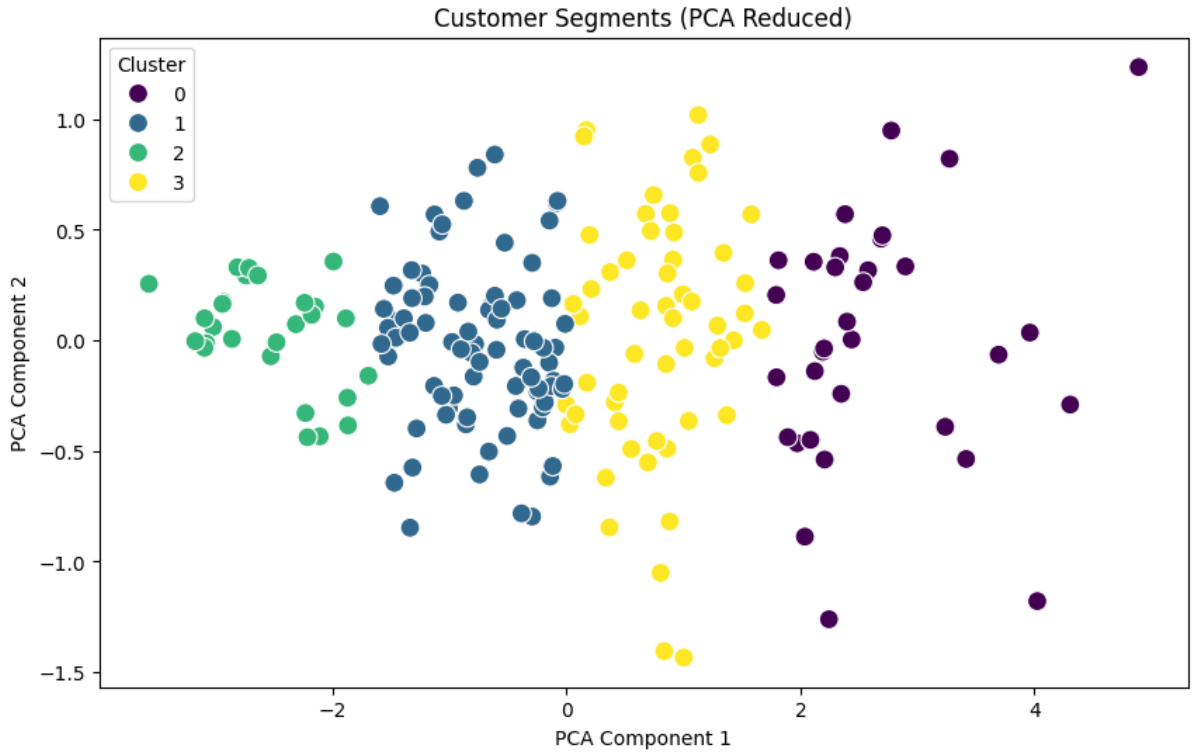
Task 3: Customer Segmentation / Clustering:



This visualization likely represents the clustering of data points that were reduced to two principal components using **Principal Component Analysis (PCA)** for easier plotting in 2D space. The process of clustering for this plot can be broken down into the following steps:

Steps Involved in Creating the Clusters

1. **Data Collection and Preprocessing**:
    - The original dataset (e.g., customer profile or transaction data) was collected and cleaned.
    - The features were scaled (e.g., using StandardScaler) to ensure uniform contribution from each feature.
2. **Dimensionality Reduction (PCA)**:
    - The data was reduced to two dimensions (PCA Component 1 and PCA Component 2) using PCA.
    - PCA helps reduce the complexity of the data while retaining most of its variance, making it easier to visualize clusters.
3. **Clustering Algorithm**:
    - A clustering algorithm (likely **K-Means**) was applied to the original or transformed data.
    - For **K-Means**, the following steps would have been executed:
        - **Step 1**: Select the number of clusters (k=4 in this case).
        - **Step 2**: Initialize cluster centroids (randomly or using other methods).
        - **Step 3**: Assign each data point to the nearest centroid (using Euclidean distance).
        - **Step 4**: Update centroids by calculating the mean of all points assigned to each cluster.
        - **Step 5**: Repeat Steps 3 and 4 until the centroids stabilize (convergence).

4. **Cluster Assignment**:
    - Each data point was assigned a cluster label (1,2,3,4 in this case).
    - These labels were used to color the data points in the visualization.
5. **Visualization**:
    - After clustering, the data points were plotted in 2D space (using PCA-reduced features).
    - Each cluster was given a unique color for clarity, and a legend was added to indicate cluster labels.

The no of clusters formed: 4

What Clusters Represent

Clusters are formed based on the similarity of customers in terms of:

- **Total Spending**: How much they spend.
- **Total Quantity**: Number of items they purchase.
- **Transaction Count**: How often they make purchases.

Each cluster groups customers with similar profiles and behaviors. For example:

1. **Cluster 1 (Purple)**:
    - Could represent **low-spending, low-frequency customers**.
    - These customers make fewer transactions and spend less overall.
2. **Cluster 2 (Blue)**:
    - Might represent **moderate-spending, regular customers**.
    - These customers are consistent but do not spend excessively.
3. **Cluster 3 (Green)**:
    - Likely represents **high-spending but infrequent customers**.
    - They purchase in large amounts occasionally.
4. **Cluster 4 (Yellow)**:
    - Could represent **high-frequency, high-spending customers**.
    - These are loyal or priority customers who purchase often and spend a lot.

1. **Davies-Bouldin Index (DB Index):**

   The **Davies-Bouldin Index (DB Index)** evaluates the quality of clustering by comparing the separation and compactness of clusters. It is based on the ratio of the **within-cluster scatter** (how compact a cluster is) to the **between-cluster separation** (how far apart clusters are).

- **Value**: 0.8052437830269734
- **Explanation**:
  - The DB Index measures the average "similarity" between clusters.
  - Lower values indicate better clustering (i.e., well-separated clusters with compact data points within each cluster).
- **Interpretation**:
  - A DB Index of 0.8052 is generally **good**, showing that the clusters are relatively distinct but could potentially improve.
  - A **lower DB Index** indicates that the clustering algorithm has done a good job separating clusters.

2. **Silhouette Score:**

   The **Silhouette Score** measures how well a data point fits into its cluster compared to other clusters. It evaluates both the compactness of a cluster and how distinct it is from other clusters.

- **Value**: 0.39004223332536625
- **Range**: [-1, 1]
  - **1**: Data points are perfectly matched to their cluster.
  - **0**: Data points are on the boundary between clusters.
  - **-1**: Data points are assigned to the wrong cluster.
- **Explanation**:
  - Measures how well data points fit within their clusters versus the nearest neighboring cluster.
- **Interpretation**:
  - A score of 0.39 is considered **moderate**, suggesting that the clusters are reasonably well-separated but with some overlap.
  - The **Silhouette Score** shows that while most points fit well into their clusters, there may be some ambiguity (e.g., overlapping clusters or points near boundaries).