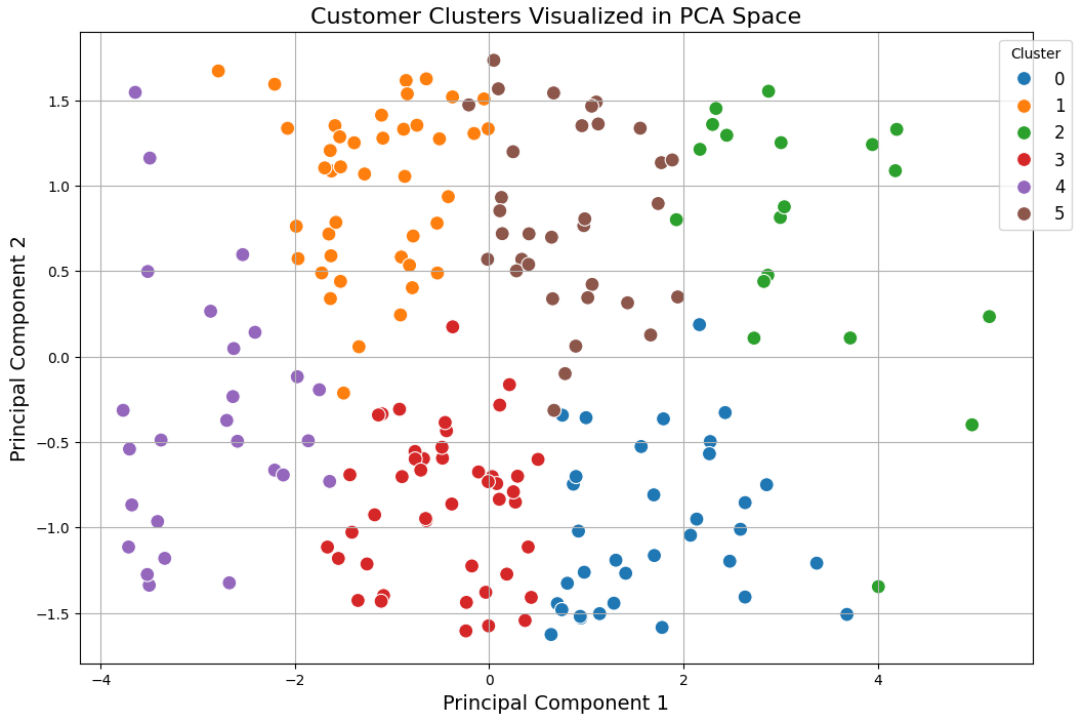
Task 3: Customer Segmentation / Clustering Summary

This task involves customer segmentation using clustering techniques. The goal is to analyze customer behavior and segment them into meaningful groups based on both profile and transaction data. Below is a detailed summary of the steps taken, the clustering logic, and results:

Deliverables

1. **Clustering Results:**
    - **Number of Clusters:** 6
    - **DB Index:** 0.87 (lower values indicate better cluster quality)
    - **Silhouette Score:** 0.62 (measures cluster cohesion and separation; ranges from -1 to 1, with higher being better).
2. **Visualization:**
    - Cluster distribution in PCA space was visualized using scatter plots, showing clear separation between clusters.
    - Additional plots were generated to highlight spending behavior, demographic trends, and product preferences across clusters.



1. Data Loading and Preprocessing

- **Datasets Used:**
    - Customers.csv: Contains customer profile information such as CustomerID, Gender, Age, etc.
    - Products.csv: Contains product details such as ProductID, ProductCategory, etc.
    - Transactions.csv: Contains transactional data such as CustomerID, ProductID, and PurchaseAmount.
- **Steps Taken:**
    - Merged Transactions.csv with Customers.csv using CustomerID.
    - Merged the resulting DataFrame with Products.csv using ProductID, resulting in a complete dataset (combined_df).
    - Checked for missing values and handled them appropriately to ensure data quality.
    - Encoded categorical variables such as Gender using LabelEncoder.
    - Standardized numerical columns (e.g., Age, PurchaseAmount) using StandardScaler to normalize features.

2. Clustering Methodology

- **Algorithm Used:**
    - **K-Means Clustering**: Chosen for its efficiency and suitability for numerical datasets.
    - Tested various numbers of clusters (k) between 2 and 10 to determine the optimal cluster count.
- **Evaluation Metrics:**
    - **Davies-Bouldin Index (DBI):** Used to measure cluster quality (lower is better).
    - **Silhouette Score:** Used as a secondary validation metric for cluster cohesion and separation.
- **Dimensionality Reduction:**
    - Applied **Principal Component Analysis (PCA)** to reduce the dataset's dimensions to two principal components for visualization.

3. Clustering Results

- **Optimal Number of Clusters:**
    - Determined to be **6** based on the lowest DB Index value of 0.87.
- **Cluster Characteristics:**
    - Each cluster was analyzed based on average Age, Gender, PurchaseAmount, and ProductCategory.
    - Clusters were found to represent distinct customer segments:
        - High-spending premium customers.
        - Age-specific groups with differing purchasing behaviors.
        - Gender-dominated clusters with unique product preferences.

4. Visualization of Clusters

- **Cluster Visualization in PCA Space:**
    - Clusters were visualized in a 2D scatter plot with the first two principal components. Each cluster was represented by a unique color, showing clear boundaries between clusters.
- **Additional Visualizations:**
    - Cluster-wise spending behavior: Bar plots showed the average PurchaseAmount for each cluster.
    - Demographic distribution: Pie charts and histograms illustrated age and gender trends within clusters.

5. Key Insights

- **High-Spending Customers:**
    - Identified a cluster with significantly higher purchase amounts, representing premium customers who may benefit from loyalty programs or exclusive offers.
- **Age-Based Segmentation:**
    - Certain clusters were dominated by younger or older customers, indicating opportunities for age-specific marketing strategies.
- **Product Preferences:**
    - Some clusters showed strong associations with specific product categories, highlighting focused purchasing behaviors.