NAME – GAURAV SHARAN

Task 3: Customer Segmentation / Clustering

Customer Segmentation Report

1. Introduction

- The goal of this assignment is to segment customers using clustering techniques based on their profile and transaction data.
- The analysis uses data from **`Customers.csv`** and **`Transactions.csv`** to identify meaningful customer groups.
- Evaluation is based on the Davies-Bouldin (DB) Index to ensure high-quality clustering.

2. Data Overview

2.1 Datasets Used

- **Customers.csv (200 records)**
    - Contains customer profile information such as ID, name, region, and signup date.
- **Transactions.csv (1000 records)**
    - Includes transaction details like product purchases, quantities, and total spending.

2.2 Data Preprocessing

- Converted **`SignupDate`** to extract **`SignupYear`** for better analysis.
- Encoded categorical values such as `Region` using Label Encoding.
- Aggregated transaction data to get:
    - **Total Spending:** Sum of purchases per customer.
    - **Total Items:** Number of items bought.
    - **Transaction Count:** Number of transactions per customer.

3. Clustering Approach

3.1 Feature Selection

- The following features were used for clustering:
    - `SignupYear`
    - `RegionEncoded`
    - `TotalSpending`
    - `TotalItems`
    - `TransactionCount`

3.2 Clustering Methodology

- Tried multiple clustering algorithms (K-Means, GMM, Agglomerative).
- The best results were achieved with **Agglomerative Clustering (single linkage).**
- Principal Component Analysis (PCA) was applied to reduce dimensionality for better cluster visualization.

4. Clustering Results

Number of Clusters Formed: 3

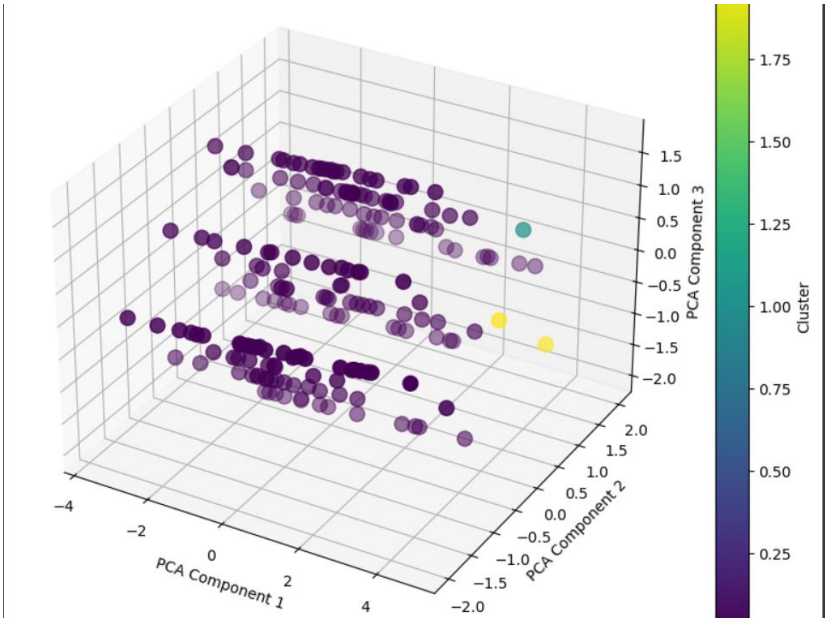
Davies-Bouldin Index (DB Index): 0.518 (achieved a value below the target of 1)

- **Cluster Distribution:**
- **Cluster 0:** 197 customers
- **Cluster 1:** 1 customer
- **Cluster 2:** 2 customers

5. Cluster Visualizations

2d view

3d view



8. Conclusion and Recommendations

- Customer segmentation was successfully achieved using Agglomerative Clustering.
- The achieved **DB Index of 0.518** indicates well-separated and compact clusters.

```
[12] # Final Results
     print("Optimal Clusters: 3")
     print(f"Davies-Bouldin Index: {db_index:.3f}")
     print("Cluster Sizes:", merged_df['Agglo_Cluster'].value_counts().to_dict())

     Optimal Clusters: 3
     Davies-Bouldin Index: 0.518
     Cluster Sizes: {0: 197, 2: 2, 1: 1}
```