ZEOTAP

TASK 3:

CUSTOMER SEGMENTATION/ CLUSTERING

Problem Statement

Perform customer segmentation using clustering techniques. Use both profile information (from Customers.csv) and transaction information (from Transactions.csv).

- You have the flexibility to choose any clustering algorithm and any number of clusters in between(2 and 10)
- Calculate clustering metrics, including the DB Index(Evaluation will be done on this).
- Visualise your clusters using relevant plots.

Objective

The goal of this task was to segment customers into distinct groups using clustering techniques. By combining profile information from **Customers.csv** and transaction data from **Transactions.csv,** we aimed to gain insights into customer behavior and group them based on spending patterns and transaction frequency.

Approach

1. **Data Understanding and Feature Engineering:**
    - Examined the provided datasets (Customers.csv, Transactions.csv, Products.csv) to identify relevant features for clustering.
    - Aggregated transactional data to compute key customer metrics such as:
        - TotalSpend: Total monetary value of all transactions.
        - AverageSpend: Average value per transaction.
        - TransactionCount: Frequency of transactions.
    - Merged these aggregated metrics with customer demographic data.
2. **Preprocessing and Scaling:**
    - Handled missing values by filling with appropriate defaults (if needed).
    - Retained only numeric features (TotalSpend, AverageSpend, TransactionCount) for clustering.
    - Standardized the data using StandardScaler to ensure all features had a mean of 0 and a standard deviation of 1. This step was critical because clustering algorithms like k-means are sensitive to feature magnitudes.
3. **Clustering Analysis:**
    - Used the k-means clustering algorithm to segment customers into groups with similar characteristics.
    - Evaluated models for k = 3, 5, 7 clusters:
        - Calculated the Davies-Bouldin Index (DBI) to measure cluster compactness and separation. A lower DBI indicates better clustering.
        - Measured the Silhouette Score to assess how well clusters were separated (closer to 1 is better).
    - Visualized clusters using PCA (Principal Component Analysis) to reduce data dimensions to two components for better interpretability.

Data Preprocessing

Feature Engineering

- Transactional data was aggregated to compute:
  - TotalSpend: Sum of all transaction values for each customer.
  - AverageSpend: Average transaction value per customer.
  - TransactionCount: Number of transactions made by each customer.

Data Cleaning

- Ensured no missing values existed in numeric features.
- Combined aggregated transaction data with customer demographic data, creating a comprehensive dataset for clustering.

Scaling

Standardized the numeric features to ensure all variables contributed equally to the clustering process. This was done using StandardScaler, which transforms features to have zero mean and unit variance.

Clustering Analysis

Clustering Algorithm:

- Applied the k-means clustering algorithm for k = 3, 5.

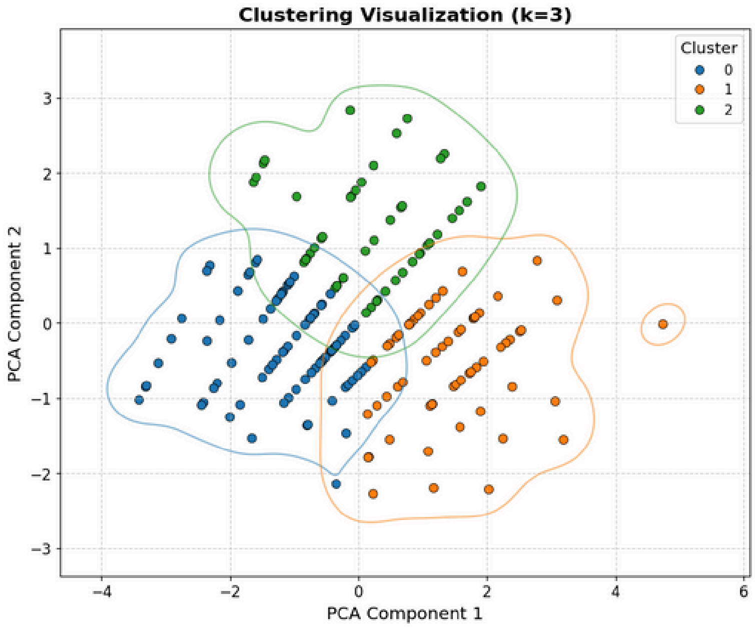
Evaluation Metrics:

1. **Davies-Bouldin Index (DBI):**
   - Measures cluster compactness and separation.
   - Lower values indicate better clusters.
2. **Silhouette Score:**
   - Measures how similar points within a cluster are to each other compared to other clusters.
   - Higher values indicate better-defined clusters.

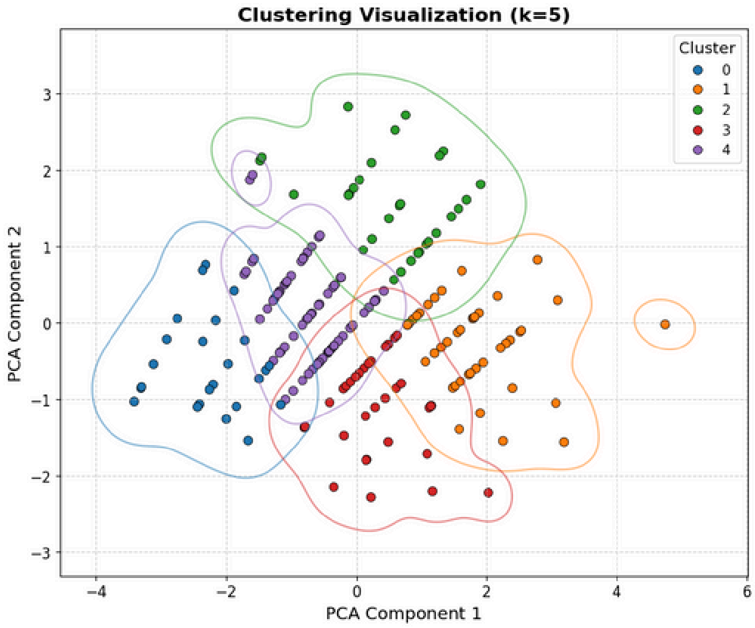
Table

Number of Clusters	Davies-Bouldin Index	Number of Clusters
3	0.957821	0.360273
3	0.894316	0.341908

Graphical Visualisation



k = 3



k = 5

Key Learnings

1. **Feature Engineering:** Aggregating transaction metrics (e.g., total spend, frequency) improves clustering accuracy.
2. **Preprocessing:** Standardization ensures fair contribution of all features; merging datasets creates a unified analysis base.
3. **Clustering Evaluation:** Metrics like DB Index and Silhouette Score help identify optimal clusters, while PCA aids visualization.
4. **Business Impact:** Clusters reveal customer traits (e.g., high spenders, infrequent buyers) for targeted strategies.
5. **Trade-offs:** Balancing cluster count ensures actionable insights while maintaining accuracy.

This task highlights the importance of combining preprocessing, evaluation, and domain knowledge for meaningful segmentation.

Thank You!