Customer Segmentation / Clustering Report

1. Introduction

The purpose of this analysis is to segment customers based on their transaction data and profile information. Customer segmentation helps businesses understand different customer behaviors and tailor their marketing strategies accordingly. In this report, I used clustering techniques to group customers based on their spending patterns and other relevant features.

2. Data Overview

The dataset consists of the following:

- **Customers Data**: Contains customer IDs, names, regions, and signup dates.
- **Transactions Data**: Contains details of customer transactions including product categories and total spending.
- **Products Data**: Contains information about product categories and their IDs.

3. Data Overview

The data underwent several preprocessing steps:

- **Merging Datasets**: Transaction data was merged with product categories to calculate how much each customer spent on each category.
- **Handling Missing Values**: Missing category spend data was replaced with 0.
- **Feature Engineering**: Total spend per customer and recency (days since signup) were calculated and added as features.
- **Standardization**: The features were standardized using **StandardScaler** to have zero mean and unit variance.

4. Clustering Methodology

4.1. Choosing Clustering Algorithm

I chose **KMeans** clustering for customer segmentation. KMeans is a widely used clustering algorithm that assigns each data point to one of the predefined number of clusters.

4.2. Feature Selection

The features used for clustering included:

- Total spend per customer
- Recency (days since signup)
- Spend in different product categories (e.g., Books, Electronics, Clothing, Home Decor)

4.3. Dimensionality Reduction

To improve visualization and reduce complexity, **Principal Component Analysis (PCA)** was applied to reduce the data to 2 components. This allowed us to visualize the clusters in a 2D space.

4.4. Choosing the Number of Clusters

The optimal number of clusters was determined using two methods:

1. **Elbow Method**: The Elbow Method helps identify the point where increasing the number of clusters no longer significantly reduces inertia (the sum of squared distances of samples to their closest cluster center).
2. **Silhouette Score**: The Silhouette Score evaluates how well-separated the clusters are. A higher silhouette score indicates better-defined clusters.

4.5. Clustering Evaluation Metrics

- **Silhouette Score**: The average silhouette score for the clustering was **0.33**, indicating moderate cohesion and separation between clusters.
- **Davies-Bouldin Index (DBI)**: The Davies-Bouldin Index is **0.98**, indicating that the clustering solution is relatively good, though there is some overlap between clusters.

5. Results

5.1. Number of Clusters

Based on the analysis from the Elbow Method and the Silhouette Score, the optimal number of clusters was determined to be **3**.

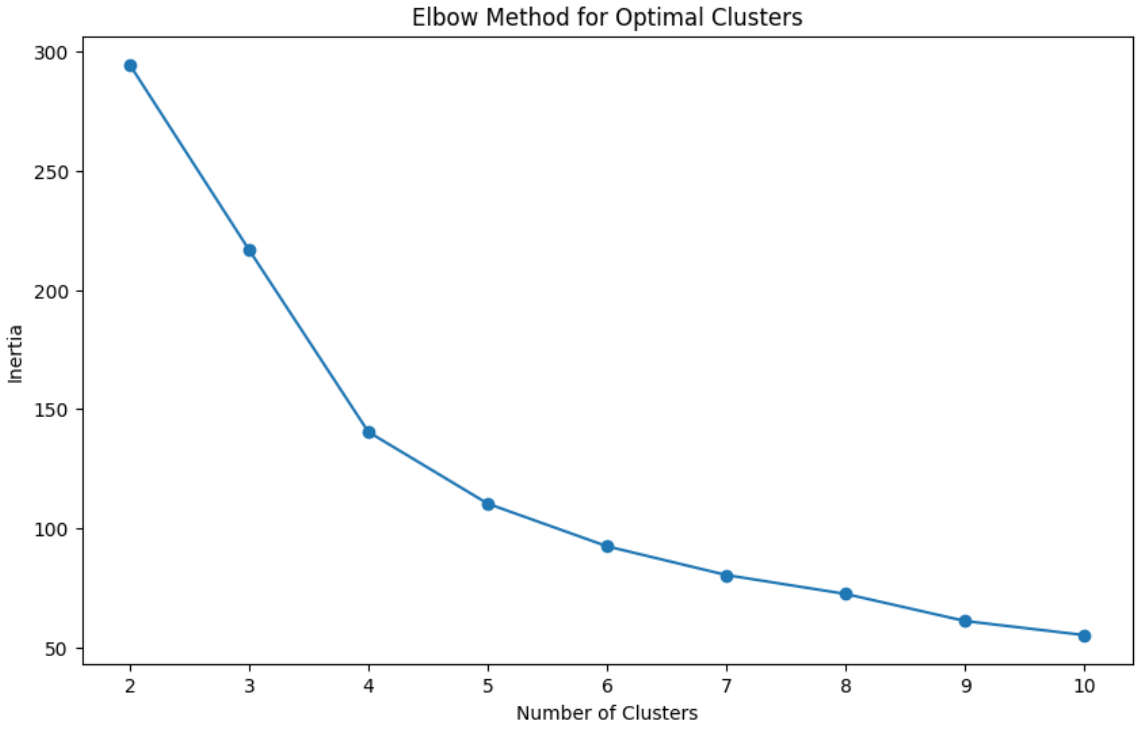
5.2. Clustering Metrics

- **Silhouette Score**: **0.33**
- **Davies-Bouldin Index (DBI)**: **0.98**

5.3. Visualizations

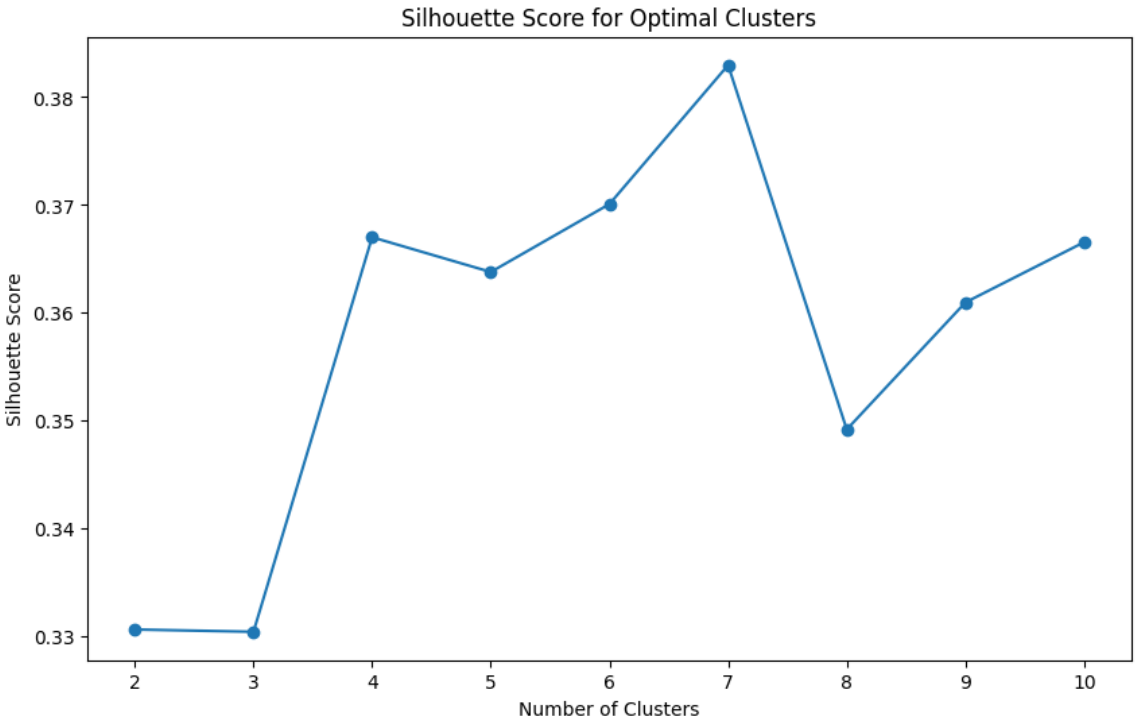
Elbow Method for Optimal Clusters

The graph below shows the **Elbow Method** for determining the optimal number of clusters. It plots the **Number of Clusters** (x-axis) against the **Inertia** (y-axis). The "Elbow" point suggests that 3 clusters are optimal.



Silhouette Score for Optimal Clusters

The graph below shows the **Silhouette Score** for different cluster counts. A higher silhouette score indicates better-defined clusters. The optimal number of clusters (3) is chosen based on this analysis.



Customer Segments using KMeans (n_clusters=3)

The following scatter plot visualizes the customer segments after applying **KMeans** clustering. The customers are represented in the 2D PCA space, with each color representing a different cluster.

6. Conclusion

In this analysis, we successfully performed customer segmentation using clustering techniques. The results show that 3 clusters provide a reasonable segmentation of customers based on their spending behaviors and recency. The clustering results were evaluated using the **Silhouette Score** and **Davies-Bouldin Index**, which confirmed the adequacy of the chosen clustering solution.

Report Prepared by:

Meet Mistry