Samarth A

Customer Segmentation Using K-Means Clustering

The primary aim of this clustering is to segment customers based on their transactions and demographic data, enabling the company to target different customer groups more effectively and optimize business strategies.

Data Preprocessing:

1. **Data Cleaning:** Ensured data columns signupdate and transactiondate were correctly parsed as datetime objects.

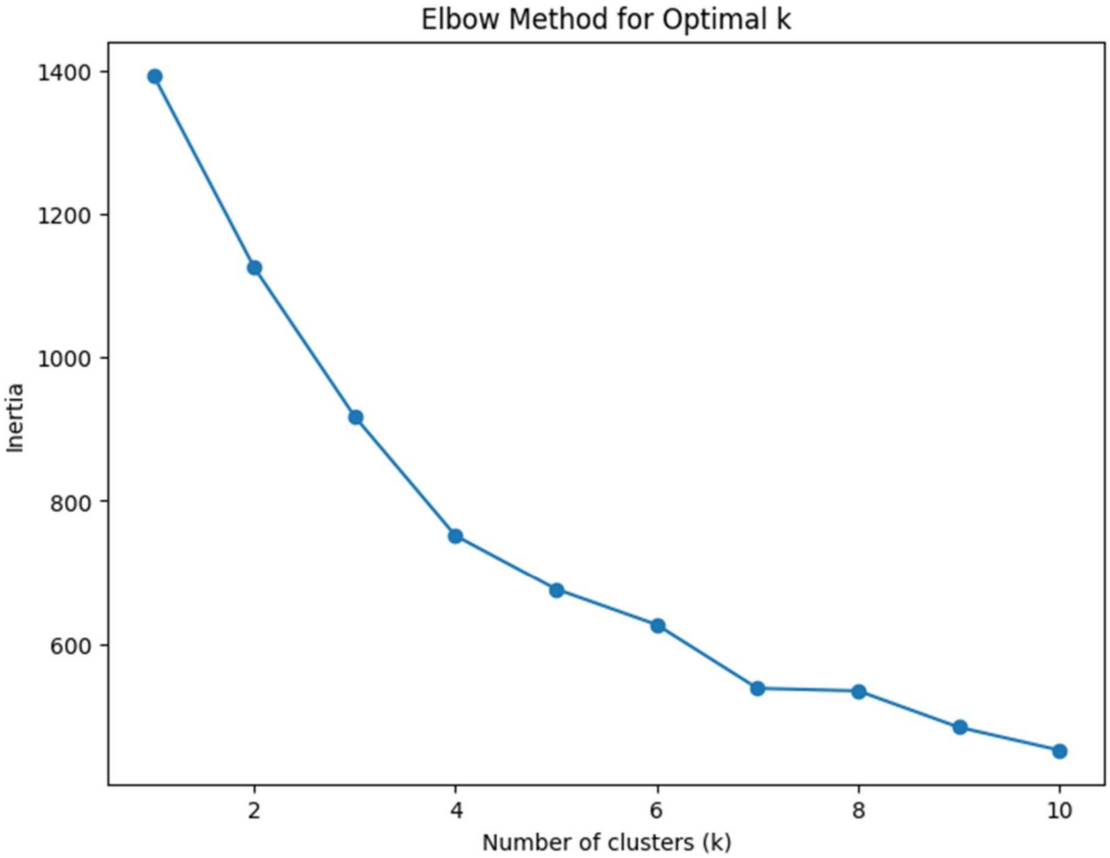
2. **Feature Engineering:**
   - **Customer Tenure:** Calculated the number of days since each customer signed up difference between the current date and signup date.
   - **Transaction Aggregation:** Aggregate transaction data for each customer:
     - **Total Spending:** Sum of all transaction values.
     - **Average Transaction:** Mean transaction values.
     - **Transaction Count:** Total number of transactions.

3. **Feature Encoding:** Categorical variable "**Region**" was encoded using one – hot encoding to create binary columns for each region.

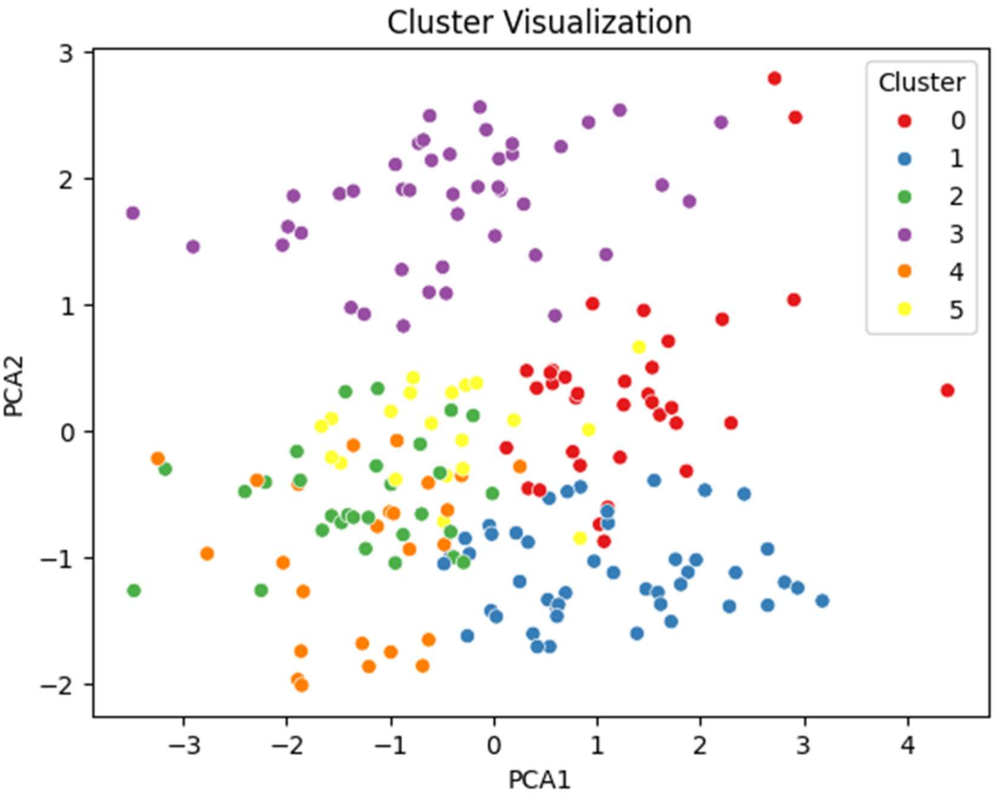
4. **Feature Selection:**
   - Selected key features for clustering:
     - **Tenure**: Represents customer loyalty by calculating the number of days since the customer signed up. Longer tenure generally indicates more established relationships and possibly higher loyalty.
     - **TotalSpending**: The sum of all transaction values for each customer. This measures the overall financial contribution of a customer to the business, highlighting high-value customers.
     - **AvgTransactionValue**: The mean amount spent per transaction by a customer. It provides insight into the purchasing behaviour of customers, helping to identify those who tend to make larger or more frequent purchases.
     - **TransactionCount**: The total number of transactions made by each customer. This feature reflects customer engagement and can indicate how often customers interact with the business.
     - **Region Encodings**: Geographic data encoded into binary variables to capture regional differences in customer behaviour, such as spending patterns or preferences specific to different locations. This helps understand how location influences customer activities.

Clustering Details:

1. **Number of Clusters:** The K-Means clustering algorithm was applied with 6 clusters (k=6), as specified for this analysis.
2. **Evaluation Metrics:**
   - **Davies-Bouldin Index – DBI:** Value: 1.3045
     The DBI is a measure of clustering quality, where lower values indicate better-defined clusters. A value of 1.238 suggests well-separated and compact clusters.
   - **Silhouette Score:** Value: 0.253
     A score of 0.253 is moderate, suggesting some overlap between clusters but still meaningful distinctions.
3. **Cluster Characteristics:** Each cluster represents a group of customers with similar behaviour s and characteristics based on the following variables:
   - **Tenure:** Represents customer loyalty.
   - **Total Spending:** Indicates overall financial contribution.
   - **Average Transaction Value:** Captures purchasing behaviour.
   - **Transaction Count:** Reflects transaction frequency.
   - **Region Information:** Provides location-based segmentation.



The Elbow Method identifies the optimal number of clusters as from **4**, where the inertia curve significantly flattens, balancing cluster compactness and model simplicity. Adding clusters beyond this point provides diminishing returns in reducing intra-cluster variance. Thus, using from 4 clusters ensures meaningful segmentation without overfitting, effectively capturing key customer behaviours.



The PCA visualization illustrates the clustering of customers into 6 clusters using K-Means. Each color represents a distinct cluster, with varying densities and separation. Cluster 3 – purple is well-separated, indicating a unique group of customers with distinct characteristics. Other clusters, such as Cluster 2 - green and Cluster 5 - yellow, show some overlap, suggesting shared traits among these segments. Cluster 4 - orange and Cluster 0 - red are relatively dense, reflecting homogeneity, while dispersed clusters like Cluster 4 indicate variability within that group. This segmentation provides actionable insights to develop targeted strategies, such as personalized marketing or retention programs for specific clusters.

The clustering analysis effectively segments customers into distinct groups based on key behavioural and transactional features. From the PCA visualization, clusters like Cluster 3 – purple are well-separated, indicating unique customer groups, while others, such as Cluster 2 – blue and Cluster 5 - yellow, exhibit some overlap, suggesting shared characteristics that may require further analysis. The Elbow Method recommended k=4, but the analysis used k=6 to capture finer distinctions among customers.

The clustering quality, supported by a DB Index of 1.238 and a moderate Silhouette Score of 0.25, shows that the segmentation is meaningful but could benefit from additional feature refinement or advanced techniques. These clusters provide actionable insights for targeted marketing, customer retention strategies, and personalized offerings. Further analysis could focus on reducing overlaps and enhancing separation for even more precise segmentation.