Customer Segmentation

1. Clustering Overview

In this analysis, customer segmentation was performed using the KMeans clustering algorithm. The dataset was pre-processed by merging customer and transaction data, followed by feature scaling. The optimal number of clusters was chosen to be 4 after evaluating clustering metrics.

2. Key Metrics

- Davies-Bouldin Index: 1.0604

  The Davies-Bouldin Index measures the average similarity ratio of each cluster with its most similar cluster. A lower value indicates better clustering. The value obtained suggests a reasonably good clustering result.

- Silhouette Score: 0.3135

  The Silhouette Score quantifies how similar a point is to its own cluster compared to other clusters. A higher value (close to 1) is better, but a positive score indicates that the clusters are well formed. The obtained score suggests moderate clustering quality.

3. Visualizations

3.1. PCA Visualization of Clusters

The plot below visualizes the customer segments obtained after applying KMeans clustering on the dataset. To facilitate a more understandable and visually interpretable representation of the data, Principal Component Analysis (PCA) was used to reduce the high-dimensional feature space into two principal components. This dimensionality reduction technique captures the most important aspects of the data while simplifying the visualization process.

In the plot, each point represents a customer, and the different colours correspond to different clusters formed by the KMeans algorithm. The clusters are the result of grouping customers based on similarities in both their profile information (such as region and signup date) and transaction behaviour (such as the number of transactions, total spending, and product categories purchased). By reducing the dimensions, the plot highlights how customers with similar purchasing patterns and characteristics are grouped together. This visual representation allows us to gain insights into customer segmentation, which can be valuable for targeted marketing, personalized recommendations, and other business strategies.

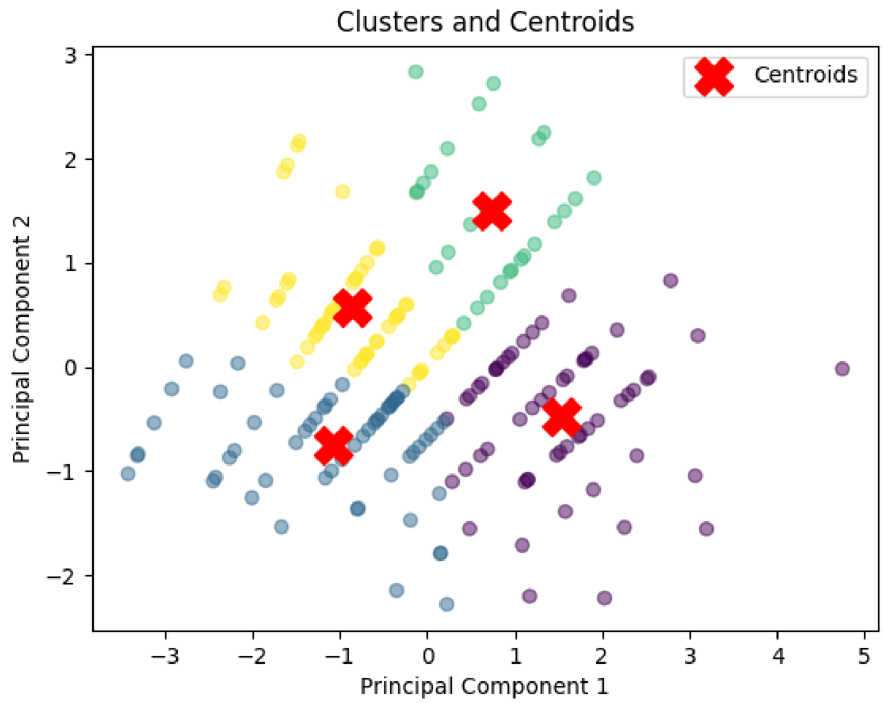
The separation between clusters indicates distinct groups of customers with different buying behaviours, enabling businesses to tailor their strategies to each segment's unique characteristics.

3.2. Clusters with Centroids

This plot displays the same customer clusters as in the previous visualization, but with an added layer of insight: the red 'X' markers represent the centroids of each cluster. A centroid in clustering is the central point or average location of all the data points within that cluster. In this case, the centroids are calculated as the mean of all customer points that belong to a specific cluster in the reduced 2D space.

The centroids serve as the "centre" of each cluster and provide a summary of the typical characteristics of customers within that cluster. By identifying the centroids, we can better understand the general behaviour and profile of the customers in each segment. For example, a centroid closer to the upper right might indicate a cluster of high-spending customers who tend to purchase frequently, while a centroid in the lower left could suggest a group with more moderate spending behaviour.

These central points act as the representative "profiles" for each customer segment, helping businesses identify and target specific groups more effectively. The distance between centroids can also offer valuable insights: clusters that are farther apart indicate more distinct customer segments, whereas clusters that are close together suggest overlapping behaviours and profiles. Understanding these centroids can guide marketing teams in designing tailored campaigns and product offerings that match the characteristics of each cluster.



4. Conclusion

The customer segmentation analysis revealed four distinct clusters based on customer profiles and transaction behaviour. The Davies-Bouldin Index and Silhouette Score indicate reasonable quality clustering, with some room for improvement. These clusters can be used for targeted marketing, customer retention, and product recommendations.