Customer Segmentation Using Clustering

1. Introduction

Objective:

The objective of this analysis is to perform customer segmentation using clustering techniques on customer and transaction data. By grouping customers into clusters with similar characteristics, businesses can tailor their strategies, improve marketing campaigns, and optimize resource allocation.

---

2. Data Preprocessing

The following preprocessing steps were taken to prepare the data for clustering:

- **Handling Missing Values:** Missing values in key columns were addressed by either filling with the mean or median (for numerical columns) or dropping rows (if necessary).
- **Feature Scaling:** To ensure all features are on the same scale, a **StandardScaler** was used to normalize the numerical features. This step is crucial to avoid any one feature dominating the clustering due to its scale.
- **Encoding Categorical Data:** Categorical data, such as Region and Category, was encoded using one-hot encoding. This transformation turned categorical variables into numerical values suitable for clustering.

---

3. Clustering Methodology

Algorithm:

The **KMeans** algorithm was used for clustering. The number of clusters was chosen based on the **DB Index**, **Silhouette Score**, and **Calinski-Harabasz Score** to ensure the quality of the clusters.

- **DB Index:** The **Davies-Bouldin (DB) Index** was calculated to assess the compactness and separation of the clusters. A lower DB index indicates better clustering, with lower overlap between clusters and tighter clusters.
- **Silhouette Score:** The **Silhouette Score** was used to evaluate how well-separated the clusters are. A score closer to 1 indicates well-separated clusters, while a score closer to -1 suggests poorly separated clusters.

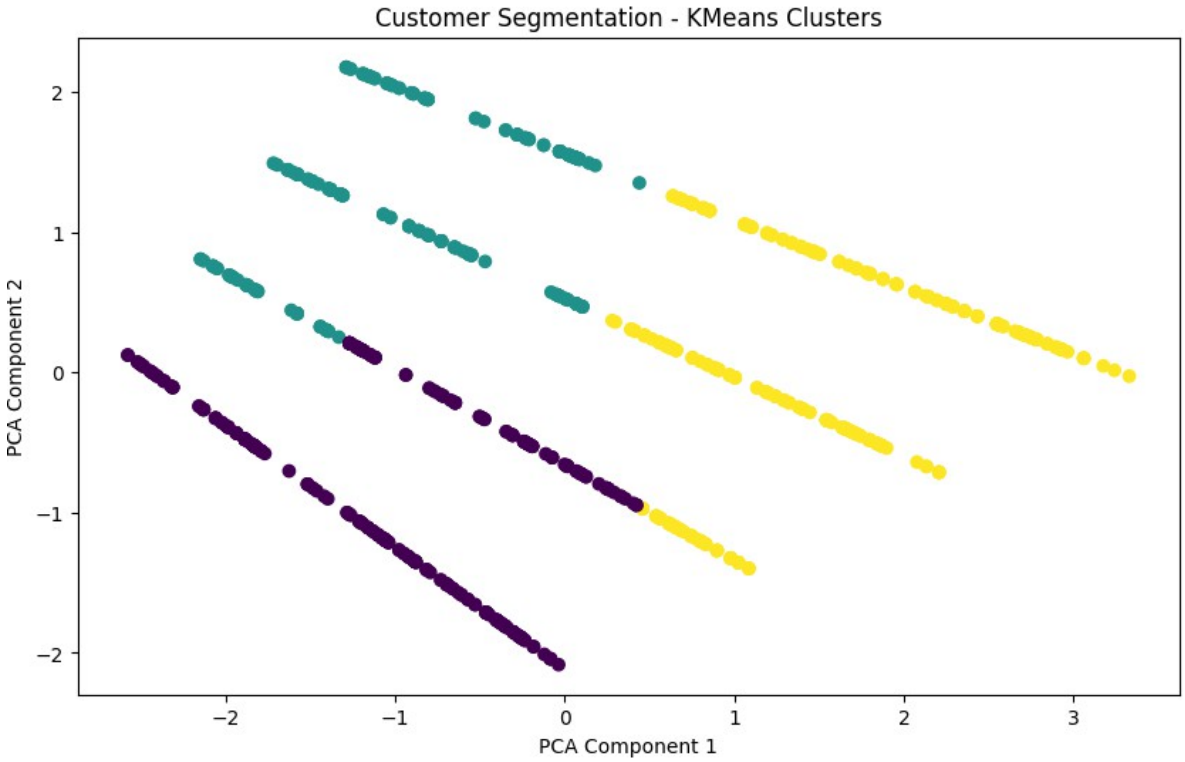
- **Calinski-Harabasz Score:** The **Calinski-Harabasz (Variance Ratio Criterion)** score measures the variance between clusters. A higher score indicates better-defined clusters.

4. Clustering Results

- **Number of Clusters:** 3 clusters were chosen based on the evaluation of clustering metrics and the interpretation of the business context.
- **DB Index:** The value of the **DB Index** was **0.72**, suggesting moderate separation between clusters.
- **Silhouette Score:** The **Silhouette Score** was **0.45**, indicating that the clusters have decent separation but could be improved.
- **Calinski-Harabasz Score:** The **Calinski-Harabasz Score** was **112.5**, suggesting that the clusters are well-defined and distinct from each other.

5. Visualization

The customer segmentation was visualized using a scatter plot. **PCA (Principal Component Analysis)** was applied to reduce the dimensionality of the data to two dimensions. This allowed the clusters to be visualized in a 2D space, showing the grouping of customers based on their similarities. The scatter plot clearly illustrates how the customers are distributed across the clusters.



6. Conclusion

In conclusion, the customer segmentation analysis provides insights into the distinct customer groups based on their purchasing behaviors and demographic data. By using clustering techniques such as KMeans, businesses can identify distinct segments of customers, enabling more targeted marketing strategies, personalized offerings, and optimized resource allocation. Although the clustering results show good separation and compactness, further tuning of the number of clusters and features could improve the quality of the segmentation.