Customer Segmentation using Clustering Techniques

K-Means Clustering and Optimal Cluster Formation

In this analysis, K-Means clustering was applied to a combined dataset, formed by concatenating numeric_df and categorical_df into a new DataFrame called final_data, using the pd.concat([numeric_df, categorical_df], axis=1) command, and the Elbow Method was employed to determine the optimal number of clusters by evaluating inertia values across a range of cluster numbers (k) from 2 to 10, with the inertia values plotted to identify the point where the rate of decrease in inertia slows, suggesting the ideal number of clusters, and after visual inspection, the optimal number of clusters was found to be 4, which was subsequently used to fit a K-Means model with 4 clusters, and the resulting cluster assignments were added to the original dataset as a new column labeled 'Cluster', indicating which of the 4 clusters each data point belongs to.

Evaluating the Clustering Performance

To evaluate the quality of our clustering results, we used two common metrics: the Davies-Bouldin Index (DB Index) and the Silhouette Score. The DB Index measures the average similarity ratio between each cluster and its most similar cluster, with a lower value indicating better clustering. In our case, the DB Index was calculated to be 0.4606, suggesting relatively well-separated clusters. The Silhouette Score measures how similar each point is to its own cluster compared to other clusters, with a higher score indicating better-defined clusters. Our model achieved a Silhouette Score of 0.7103, indicating that the clusters are fairly well-separated and well-defined. Together, these metrics give us a strong indication of the effectiveness of the clustering model.

Visualizing and Analyzing Customer Segments

To visualize the customer segments formed by KMeans clustering, Principal Component Analysis (PCA) was applied to reduce the data to two dimensions for a 2D scatter plot, with points color-coded by their cluster assignment. This helped assess the separation and effectiveness of the clustering. Additionally, a 3D scatter plot was created using the first three features, allowing further insight into the distribution of the clusters. Finally, a correlation matrix heatmap was generated to visualize the relationships between the numerical features, helping identify correlations and potential multicollinearity.