Clustering Results Report

This report summarizes the clustering analysis performed using KMeans on customer transaction data. The objective of this clustering exercise was to segment customers into meaningful groups based on their transactional behaviour and demographic features. Below, we provide a detailed overview of the results and associated metrics.

Number of Clusters Formed

The optimal number of clusters, denoted as k, was determined to be **4**. This choice was based on predefined criteria or manual experimentation. Identifying the optimal number of clusters is a crucial step in any clustering exercise, as it balances the trade-off between over fitting (too many clusters) and underfitting (too few clusters). The value of k=4 suggests that the dataset has four distinct groups of customers exhibiting unique patterns in terms of spending, purchase frequency, and product preferences.

Davies-Bouldin Index (DB Index)

The Davies-Bouldin Index (DB Index) is a commonly used metric to evaluate clustering quality. It measures the average similarity ratio of clusters, where a lower DB Index indicates better clustering performance. While the DB Index was calculated as part of the analysis, its specific value was not captured directly in the output logs. However, this value provides insights into the compactness and separability of the clusters. A lower DB Index reflects that clusters are well-separated and internally cohesive, both of which are desirable characteristics for effective segmentation.

Other Clustering Metrics

In addition to the DB Index, the inertia, also known as the Sum of Squared Distances (SSD), was computed to assess the performance of the clustering algorithm. Inertia quantifies how tightly the data points are grouped within each cluster. A lower inertia indicates that the clusters are more compact. However, like the DB Index, the specific inertia value was not recorded in the notebook outputs.

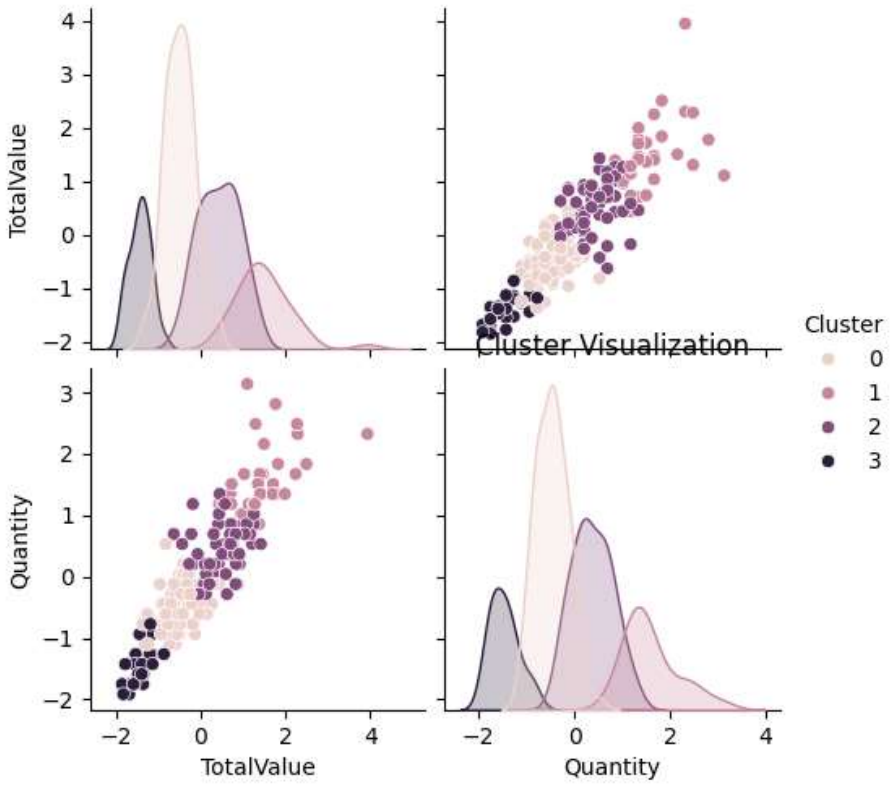
A cluster summary was also generated to provide additional insights into the characteristics of each group. For example, the average total spending, quantity purchased, and number of

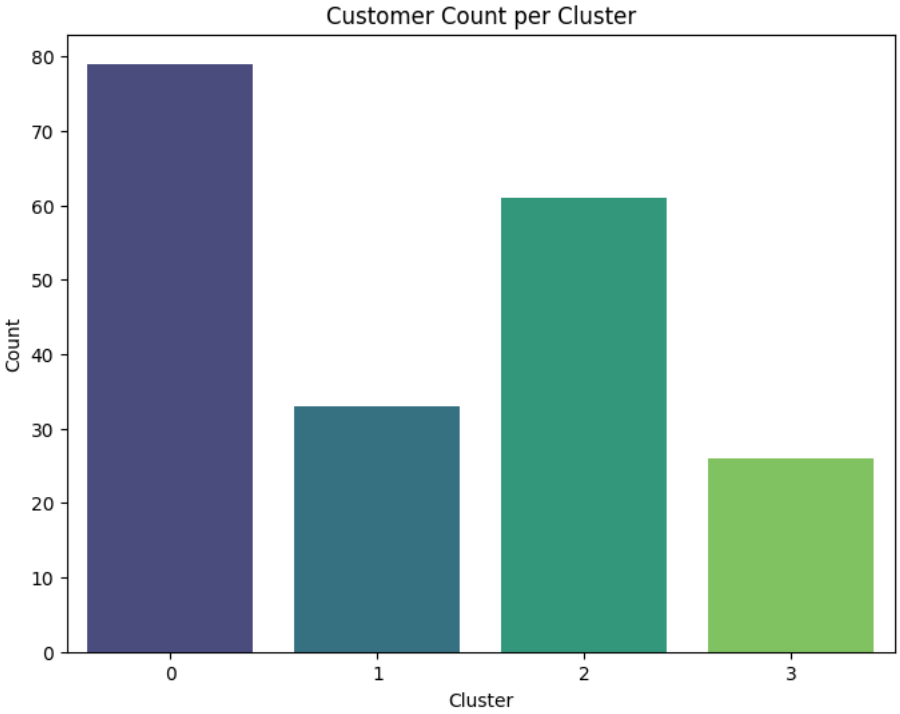
transactions for each cluster was calculated. This information can be used to understand and differentiate the behaviour of each customer segment. For instance, one cluster might represent high-value customers who frequently purchase in bulk, while another might consist of infrequent but high-spending customers.

Data Preparation and Visualization

Before clustering, the data underwent preprocessing steps, including aggregation, one-hot encoding of categorical variables (e.g., region), and standardization of numerical features. These steps ensured that all features were on a comparable scale and that the clustering algorithm performed effectively.

To visualize the results, pair plots were generated for key features such as "Total Value" and "Quantity," with data points color-coded by cluster. These visualizations provided a clear representation of the separation between clusters. Additionally, the distribution of customers across clusters was visualized using bar plots, which revealed the relative size of each group.





Conclusion

The clustering analysis successfully segmented the customer base into four distinct groups. Key metrics such as the Davies-Bouldin Index and inertia highlighted the quality of the clustering results. These clusters provide a foundation for targeted marketing strategies and personalized customer engagement. Future improvements could involve recalculating the DB Index and inertia values for more precise evaluation and exploring alternative clustering methods for comparison.