

CLUSTERING RESULTS REPORT

1. Number of Clusters Formed:

After evaluating the clustering quality using various cluster counts, the **optimal number of clusters** was determined to be **8**. This was identified using the Davies-Bouldin Index, which assesses the compactness and separation of the clusters.

2. Davies-Bouldin Index Value:

The Davies-Bouldin Index (DB Index) for the optimal clustering configuration was **1.5285**.

- The DB Index measures the quality of clustering, with lower values indicating better cluster definition. A value of 1.52 suggests that the clusters are reasonably compact and well-separated but might leave room for improvement in specific edge cases.

3. Feature Scaling:

To ensure all features contributed equally to the clustering process, the data was standardized using StandardScaler. Standardization is crucial in clustering algorithms like K-Means since it removes biases caused by differing feature magnitudes.

4. Clustering Methodology:

- The clustering was performed using the **K-Means algorithm**, which is a popular and efficient method for identifying patterns in large datasets.
- To identify the optimal number of clusters, the clustering process was repeated for a range of cluster counts (from 2 to 10), and the DB Index was computed for each configuration. The number of clusters with the lowest DB Index (6 in this case) was selected as optimal.

5. Cluster Visualization:

- A 2D visualization of the clusters was generated using **Principal Component Analysis (PCA)**. PCA reduced the high-dimensional standardized data to two principal components, allowing for easy visualization.
- In the scatter plot, customers were grouped into distinct clusters, color-coded for clarity. The plot highlighted clear separations among most clusters, demonstrating the effectiveness of the clustering approach.

6. Cluster Assignments:

Each customer was assigned to one of the six clusters. These assignments were saved in a CSV file (customer_clusters.csv) for further analysis or integration into business processes.

7. Derived Features for Clustering:

- **Transaction Features:** Customers were characterized by their total transactions, total quantity purchased, total spent, average transaction value, and recency of their last transaction.
- **Category Features:** Purchase counts across different product categories were also considered to capture customer preferences.
- **Demographics:** Features such as customer region, tenure, and signup date were incorporated to provide a holistic view of customer behavior.

8. Potential Use Cases:

- These clusters can be utilized to create targeted marketing campaigns, personalize customer engagement strategies, or identify high-value customers for retention efforts.
- Businesses can also use this clustering to optimize inventory and offer tailored product recommendations.