Customer Segmentation Using K-Means Clustering

Kushagra Agrawal

January 25, 2025

1 Introduction

This report describes an exploratory data analysis and clustering task on a retail dataset consisting of:

- `Customers.csv`
- `Transactions.csv`

The goal was to perform customer segmentation via K-Means clustering and determine the optimal number of clusters using Davies-Bouldin index.

2 Data Preparation and Merging

2.1 Dataset Description

- **Customers.csv**: Contains demographic or identifying information for each customer, such as `CustomerID`.
- **Transactions.csv**: Contains transactional information including `TransactionID`, `CustomerID`, and `TotalValue`.

These two files were merged on the column `CustomerID`. Then, the data was aggregated to engineer features suitable for clustering. Specifically, for each customer, we calculated:

- `TotalValue`: The sum of the transaction amounts for each customer.
- `AverageTransactionValue`: The mean transaction amount for each customer.
- `Frequency`: The number of transactions per customer.

2.2 Feature Selection and Scaling

The features used for clustering were:

- `TotalValue`
- `AverageTransactionValue`
- `Frequency`

These features were standardized using `StandardScaler` so that each has mean 0 and standard deviation 1, which is important for distance-based algorithms such as K-Means.

3 Methodology

3.1 K-Means Clustering

K-Means clustering was performed on the scaled feature set to partition the customers into different segments. We varied the number of clusters k from 2 to 10 to explore different grouping possibilities.

3.2 Clustering Quality Metric: Davies-Bouldin Index

The Davies-Bouldin (DB) Index is a metric used to evaluate clustering results based on the average of similarity measures between each cluster and its most similar one. A lower DB Index indicates better separation between clusters.

$$\mathrm{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

where

- S_i is the average distance of all points in cluster i to its centroid,
- M_{ij} is the distance between the centroids of clusters i and j.

4 Results

4.1 Optimal Number of Clusters

Using the Davies-Bouldin Index values for k ranging from 2 to 10, the minimum DB Index was identified at:

Optimal number of clusters: 5

with a Davies-Bouldin Index value of approximately:

0.8524813520

Hence, the clustering solution chosen was $k = 5$.

4.2 Relevant Clustering Metrics

In addition to the Davies-Bouldin Index, possible further evaluations might include:

- **Inertia (Sum of Squared Distances)**: A measure of how internally coherent each cluster is. This was minimized for the chosen k, subject to the DB Index criteria.
- **Interpretability of Clusters**: Examining each resulting cluster's centroid reveals distinct customer segments based on how frequently they purchase, how large transactions are on average, and their total purchase amounts.

4.3 Cluster Visualization

A scatter plot was generated showing `TotalValue` vs. `AverageTransactionValue`, colored by cluster assignment. This visualization helps interpret differences among clusters. For instance:

- One cluster might contain frequent shoppers who spend moderately each time.
- Another might contain customers who make large purchases but infrequently.

5 Conclusion

- **Number of clusters formed**: 5
- **Davies-Bouldin Index**: 0.8525
- **Interpretation of Clusters**:
  - Each cluster differs by frequency of purchases, total spend, and average transaction size.
  - Clusters can be used to tailor marketing strategies, loyalty programs, or highlight high-value customers.

Future improvements could involve using additional features (e.g., demographic data), employing more robust clustering algorithms, or running more in-depth validation metrics such as Silhouette Score, Calinski-Harabasz Index, or domain-specific cost-benefit analyses. Overall, the five-cluster solution provides a balanced and interpretable segmentation of the customers in this dataset.