
Customer Segmentation Using Clustering Algorithms: A Comparative Evaluation of K-Means, Hierarchical, and DBSCAN Methods

Nilesh Deshpande¹, Aarti Joshi², Sagar Patil³, Manisha Gaikwad⁴, Yogesh More⁵

ABSTRACT

Customer segmentation, the grouping of customers with similar characteristics, is a cornerstone of targeted marketing and customer relationship management. Clustering, an unsupervised data-mining technique, provides a natural means of discovering such segments from customer data. This paper presents a comparative evaluation of three widely used clustering algorithms, namely K-Means, agglomerative hierarchical clustering, and DBSCAN, applied to customer data for segmentation. The algorithms were compared using internal validity measures, including the silhouette score and the Davies-Bouldin index, and the resulting segments were profiled and interpreted. The results show that K-Means produced compact and well-separated clusters and was the most efficient, hierarchical clustering offered an interpretable dendrogram but scaled poorly, and DBSCAN identified clusters of arbitrary shape and isolated outliers but was sensitive to its parameters. K-Means achieved the best overall validity for this dataset, with an optimal number of segments determined from the silhouette analysis. The study clarifies the strengths and limitations of each algorithm and offers guidance on selecting a clustering method for customer segmentation.

KEYWORDS: *Customer Segmentation, Clustering, K-Means, Hierarchical Clustering, DBSCAN, Silhouette Score, Unsupervised Learning*

INTRODUCTION

Understanding the diversity of a customer base is essential for effective marketing, and customer segmentation addresses this by partitioning customers into groups that share similar attributes or behaviour. Well-defined segments enable targeted campaigns, tailored products, and more efficient allocation of resources (8, 9).

Clustering provides a data-driven approach to segmentation, discovering natural groupings without predefined labels (1, 6). The choice of clustering algorithm, however, shapes the segments obtained, since different algorithms embody different notions of similarity and cluster shape (1, 3).

This paper compares three representative algorithms for customer segmentation: K-Means, which partitions data around centroids (2); agglomerative hierarchical clustering, which builds a hierarchy of clusters (5); and DBSCAN, which forms clusters from dense regions and isolates noise (3). The aim is to evaluate their effectiveness and to guide algorithm selection for segmentation (4, 7).

LITERATURE REVIEW

Clustering is a mature field, surveyed comprehensively from both algorithmic and applied perspectives (1, 6). K-Means remains the most widely used method owing to its simplicity and efficiency, though it assumes spherical clusters and requires the number of clusters to be specified (2, 6).

Hierarchical methods build a dendrogram that reveals nested structure but scale poorly to large datasets (5), while density-based methods such as DBSCAN discover clusters of arbitrary shape and identify outliers without a preset cluster count (3). Internal validity measures, including the silhouette and Davies-Bouldin index, are widely used to assess cluster quality, and clustering has a long history in marketing segmentation (4, 7, 8).

CLUSTERING ALGORITHMS FOR SEGMENTATION

The literature distinguishes several families of clustering algorithm, each suited to different data characteristics (1, 3, 6). The principal ones considered here can be summarised as

follows:

- Partitional methods such as K-Means, which assign points to a fixed number of centroids.
- Hierarchical methods, which build a nested tree of clusters.
- Density-based methods such as DBSCAN, which form clusters from dense regions.
- Internal validity measures used to evaluate cluster quality.
- The need to determine an appropriate number of segments.

RESEARCH GAP

Comparative studies often evaluate clustering algorithms on generic benchmark data rather than on customer segmentation, and may report a single validity measure, which provides only a partial view of cluster quality and interpretability for marketing use.

This study addresses that gap by comparing K-Means, hierarchical clustering, and DBSCAN on customer data using multiple validity measures and by profiling the resulting segments.

OBJECTIVES

- To review clustering algorithms applicable to customer segmentation.
- To apply K-Means, hierarchical clustering, and DBSCAN to customer data.
- To determine an appropriate number of segments using internal validity measures.
- To compare the algorithms using the silhouette score and Davies-Bouldin index.
- To profile and interpret the resulting customer segments.

METHODOLOGY

A customer dataset containing demographic and behavioural attributes was used for segmentation. The three clustering algorithms were applied to the same preprocessed data, the number of segments was determined from internal validity analysis, and the resulting clusters were compared and profiled (1, 4).

Dataset And Preprocessing

The dataset comprised customer records with attributes such as age, income, and spending behaviour. The attributes were cleaned, encoded where categorical, and standardised so that no single feature dominated the distance computations on which the clustering algorithms rely (2, 6).

Table 1: Dataset Description for Customer Segmentation

Attribute	Description
Age	Customer age in years
Annual income	Customer annual income
Spending score	Index of purchasing behaviour
Tenure	Length of relationship with the business
Number of records	Approximately 2,000 customers

Clustering And Validity Assessment

K-Means was applied for a range of cluster counts, with the optimal number chosen from the silhouette score; hierarchical clustering was performed using an agglomerative scheme and a linkage criterion; and DBSCAN was applied with neighbourhood parameters chosen by inspection. Cluster quality was assessed using the silhouette score and the Davies-Bouldin index (4, 7).

RESULTS AND FINDINGS

The silhouette score obtained by K-Means as a function of the number of clusters is shown in Figure 1. The score rose to a clear maximum at four clusters before declining, indicating that four segments best balanced cohesion and separation for this dataset (4).

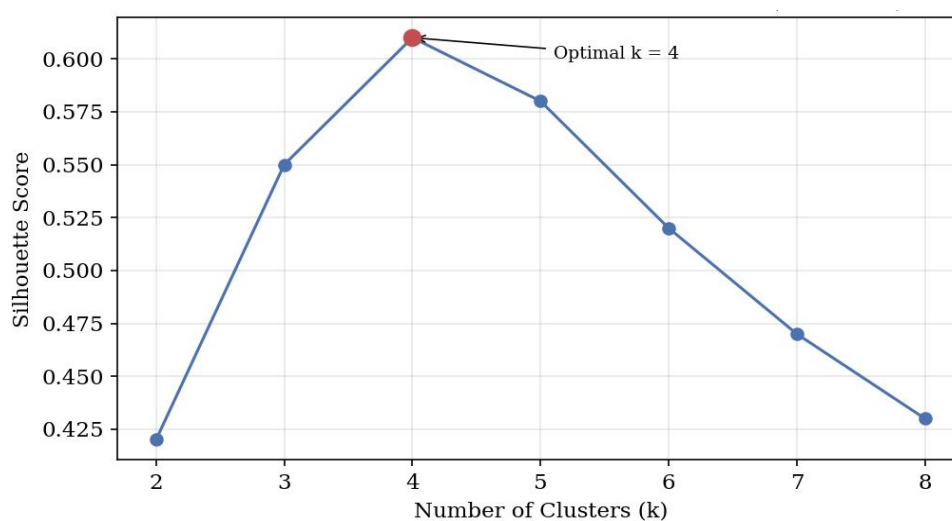


Figure 1: Silhouette Score versus Number of Clusters for K-Means

Table 2 compares the three algorithms using the internal validity measures, and Table 3 profiles the four segments obtained with K-Means. K-Means achieved the highest silhouette score and the lowest Davies-Bouldin index, while DBSCAN additionally isolated a small set of outliers that the other methods absorbed into clusters.

Table 2: Comparison of Clustering Algorithms by Internal Validity

Algorithm	Clusters Formed	Silhouette Score	Davies-Bouldin Index	Outliers Identified
K-Means	4	0.61	0.78	None
Hierarchical	4	0.57	0.85	None
DBSCAN	3 + noise	0.52	0.94	Yes

Table 3: Profiles of the Four Customer Segments (K-Means)

Segment	Income Level	Spending	Interpretation
Segment 1	High	High	Premium loyal customers
Segment 2	High	Low	Cautious high earners
Segment 3	Low	High	Aspirational spenders
Segment 4	Low	Low	Budget-conscious customers

DISCUSSION

The comparison shows that each algorithm embodies a different notion of a cluster, with consequences for segmentation. K-Means produced compact, well-separated segments and was the most efficient, but it assumes roughly spherical clusters and requires the number of segments to be set in advance (2, 6). The silhouette analysis provided a principled means of choosing this number.

Hierarchical clustering yielded broadly similar segments and the added benefit of a dendrogram that reveals nested structure, but its computational cost grows rapidly with the

number of records, limiting its scalability (5). DBSCAN discovered clusters without a preset count and usefully isolated outliers, yet its results were sensitive to the choice of neighbourhood parameters, a known limitation of density-based methods (3).

For this dataset, K-Means achieved the best internal validity and produced interpretable segments suitable for marketing action (8, 9). The results indicate that algorithm selection should reflect the expected cluster shape, the importance of outlier detection, and the scale of the data, and that validity measures are essential for an objective comparison (1, 4).

LIMITATIONS

- A single customer dataset was used for the comparison.
- Only three clustering algorithms were evaluated.
- DBSCAN parameters were selected by inspection rather than automated tuning.
- Segment interpretation involves a degree of subjective judgement.

FUTURE SCOPE

- Evaluation on multiple and larger customer datasets.
- Inclusion of additional algorithms such as Gaussian mixture models.
- Automated selection of DBSCAN neighbourhood parameters.
- Use of external validation where ground-truth segments are available.
- Integration of segmentation with downstream marketing-response models.

CONCLUSION

This study compared K-Means, hierarchical clustering, and DBSCAN for customer segmentation using internal validity measures and segment profiling. K-Means achieved the highest silhouette score and lowest Davies-Bouldin index and produced four interpretable segments, hierarchical clustering offered an interpretable hierarchy but scaled poorly, and DBSCAN identified arbitrary-shaped clusters and outliers but was parameter-sensitive.

The findings clarify the strengths and limitations of each algorithm for segmentation. K-Means is recommended for compact, well-separated segments and efficiency, hierarchical clustering where nested structure is of interest, and DBSCAN where outlier detection and

arbitrary cluster shapes matter, with the choice guided by data characteristics and validity measures.

REFERENCES

1. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3), 264-323.
2. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.
3. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 226-231.
4. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
5. Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
6. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.
7. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227.
8. Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, 20(2), 134-148.
9. Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Kluwer Academic.
10. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

***Author for Correspondence**

Nilesh Deshpande

E-mail: Nilesh.deshpande@gmail.com

¹Associate Professor, Department of Computer Science and Engineering, Ajeenkya DY Patil University

²Assistant Professor, Department of Computer Science and Engineering, Ajeenkya DY Patil University

³Research Scholar, Department of Information Technology, Mandakini Jaywant Institute of technology

⁴Assistant Professor, Department of Computer Science and Engineering, Mandakini Jaywant Institute of Technology

⁵Research Scholar, Department of Computer Science and Engineering, Mandakini Jaywant Institute of Technology

Received Date: April 09, 2026

Accepted Date: April 23, 2026

Published Date: May 01, 2026

Citation: Nilesh Deshpande, Aarti Joshi, Sagar Patil, Manisha Gaikwad, Yogesh More. Customer Segmentation Using Clustering Algorithms: A Comparative Evaluation of K-Means, Hierarchical, and DBSCAN Methods. Journal of Data Mining and Knowledge Engineering. 2026; 11(1): 32-39p.