Tourist Destination Segmentation in Jember
Using Cluster Analysis for Data-Driven Tourism Development

Bagja Kurniawan
Corresponding email: bagjakurniaone45@gmail.com
Pemerintah Provinsi DKI Jakarta, Indonesia

Abstract

Effective tourism planning in Jember requires data-driven segmentation to optimize destination management and resource allocation. This study applies three clustering algorithms—K-Means, DBSCAN, and Agglomerative Clustering—to classify tourist attractions based on ratings, ticket prices, geographic location, and visitor reviews. The preprocessing stage includes handling missing values, scaling numerical variables, and encoding categorical features. Four complementary validation metrics—Elbow Method, Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score—are employed to determine the optimal cluster structure. The results reveal three distinct clusters under K-Means, representing premium destinations, low-cost mass attractions, and high-satisfaction sites. DBSCAN identifies dense core clusters and a group of unique outlier destinations, while Agglomerative Clustering produces a clear hierarchy of high-, medium-, and low-cost attractions. These segmentation patterns provide actionable insights for targeted marketing, service quality improvement, and strategic policy formulation. The study demonstrates the value of machine learning–based clustering for evidence-based tourism development and contributes to the broader literature on data-driven regional planning.

Keywords: *Clustering, Tourism Segmentation, Jember, K-Means, DBSCAN*

INTRODUCTION

Tourism has long been recognized as a major driver of regional economic growth, employment creation, and community well-being worldwide. In many developing regions, tourism acts as a catalyst for local investment, service-sector expansion, and spatial development. Effective strategic planning is therefore essential to ensure that tourism resources are managed efficiently and aligned with visitor preferences. One of the most widely used approaches in tourism planning is market segmentation, which enables stakeholders to categorize destinations or tourists based on shared characteristics and formulate targeted development strategies.

Jember Regency in East Java possesses significant tourism potential, with a diverse portfolio of natural, cultural, and artificial attractions. However, this diversity also generates challenges for policymakers, particularly in understanding how destinations differ in terms of pricing, visitor satisfaction, and geographic distribution. Despite the growing importance of data-driven decision-making, tourism segmentation in Jember remains limited and has not yet been systematically explored using machine learning approaches.

Previous studies have demonstrated the relevance of clustering methods in tourism analytics. Mine (2009) showed that K-Means effectively groups attractions based on visitor ratings, while (Gao et al., 2022) employed machine learning to understand tourist decision pathways. Wamulkan A.S et al., (2024) further emphasized that data-mining techniques improve the accuracy of modeling tourist behavior. Although these studies confirm the value of machine learning for tourism segmentation, most of them focus on broader tourism contexts and do not account for the unique combination of spatial, experiential, and pricing features found in Jember. Moreover, prior research rarely compares multiple clustering algorithms to generate a comprehensive segmentation framework.

This body of literature suggests a clear research gap: there is a need for an integrated, multi-method clustering approach that simultaneously incorporates ratings, ticket prices, location attributes, and visitor feedback to produce a more nuanced segmentation of tourist attractions in Jember. Addressing this gap, the present study applies three clustering algorithms—K-Means, DBSCAN, and Agglomerative Clustering—to develop a data-driven segmentation model that captures the heterogeneity of Jember's destinations.

The aim of this research is twofold: (1) to classify tourist attractions in Jember into meaningful clusters based on key quantitative and spatial features, and (2) to evaluate the comparative performance of the clustering methods using established validity indices. The resulting segmentation is expected to provide practical insights for local governments, tourism managers, and industry stakeholders in designing targeted marketing strategies, prioritizing infrastructure investment, and enhancing visitor experiences. Furthermore, this study contributes theoretically by demonstrating the applicability of multi-method machine learning segmentation in regional tourism planning and by synthesizing insights from existing clustering-based tourism studies to strengthen its conceptual foundation.

RESEARCH METHODS

This study adopts a quantitative approach using cluster analysis to categorize tourist destinations in Jember based on their intrinsic and spatial characteristics. The methodological framework consists of five key stages: data collection, preprocessing, cluster determination, algorithm implementation, and model evaluation.

The dataset was compiled from secondary sources, including official tourism portals and online review platforms. The variables extracted include:

- Tourist Name
- Rating
- Weekday Entrance Ticket Price
- Weekend Entrance Ticket Price
- Location
- Photo Link
- Description

Prior to clustering, several preprocessing steps were conducted:

1. **Handling missing values** by removing incomplete or inconsistent records.
2. **Normalization of numerical variables** using Min–Max Scaling:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

3. **Encoding categorical variables**, including location and textual descriptions, using **TF-IDF vectorization** to capture semantic relevance.
4. **Feature selection** to retain variables with strong discriminatory power for clustering.

These preprocessing steps ensure comparability across variables and improve algorithmic performance.

To determine the most appropriate cluster structure, four widely recognized validation metrics were used:

1. **Elbow Method**
   Evaluates within-cluster compactness using the Within-Cluster Sum of Squares (WCSS):

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \parallel x - \mu_i \parallel^2$$

2. **Silhouette Score**
   Measures cohesion and separation:

$$S(i) = \frac{b(i) - a(i)}{\max\,(a(i), b(i))}$$

3. **Davies–Bouldin Index (DBI)**
   Computes average similarity between clusters:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

4. **Calinski–Harabasz Index (CHI)**

   Assesses the ratio of between-cluster to within-cluster dispersion:

$$CHI = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n - k}{k - 1}$$

These four metrics were selected because they collectively capture complementary dimensions of clustering quality:

- **WCSS/Elbow Method** evaluates internal compactness, making it suitable for centroid-based algorithms such as K-Means.
- **Silhouette Score** assesses both cohesion and separation, providing a balanced measure applicable across different cluster shapes.
- **DBI** penalizes clusters that are insufficiently separated, making it effective for detecting overlapping structures.
- **CHI** measures structural validity by comparing between-cluster dispersion to within-cluster homogeneity.

Using multiple metrics minimizes bias and enhances robustness, ensuring that the optimal *k* is not determined solely by one criterion but by the convergence of several validity indicators. This multi-metric approach follows best practices in machine learning-based clustering evaluations.

After determining the optimal number of clusters, three clustering algorithms were applied:

1. K-Means Clustering

A centroid-based algorithm that partitions data into *k* clusters by minimizing intra-cluster variance:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \parallel x - \mu_i \parallel^2$$

2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Identifies clusters based on density thresholds:

- A **core point** has at least *MinPts* neighbors within distance ε.
- A **border point** is reachable from a core point but has fewer neighbors.
- A **noise point** does not satisfy either condition. DBSCAN is particularly useful for detecting outliers and irregular cluster shapes.

3. Agglomerative Hierarchical Clustering

A bottom-up hierarchical method using Ward's linkage:

$$D(A,B) = \sqrt{\frac{|A|\cdot|B|}{|A|+|B|}} \parallel \bar{x}_A - \bar{x}_B \parallel$$

This approach produces a cluster hierarchy, allowing clearer interpretation of the relationships among clusters.

Assessment and Visualization

Cluster outputs were visualized using **Principal Component Analysis (PCA)** to reduce dimensionality and project high-dimensional features into two-dimensional space. This visualization aids in evaluating separation patterns and structural coherence across clustering methods.

The proposed clustering framework provides a data-driven segmentation of tourist attractions in Jember, enabling stakeholders to identify segment characteristics and develop targeted strategies for marketing, infrastructure development, and service improvement. The methodology also demonstrates the effectiveness of multi-metric evaluation in enhancing the accuracy and interpretability of tourism segmentation models.

RESULTS AND DISCUSSION

a. Results

This study applies three clustering algorithms—K-Means, DBSCAN, and Agglomerative Clustering—to classify tourist destinations in Jember based on rating, ticket prices, and geographic location. The clustering performance was evaluated using four validation metrics: Elbow Method, Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. All four indicators consistently identified **k = 3** as the optimal cluster structure, demonstrating convergence across compactness, separation, and dispersion criteria.

Determination of Optimal Number of Clusters

The optimal number of clusters was determined using four validation techniques:

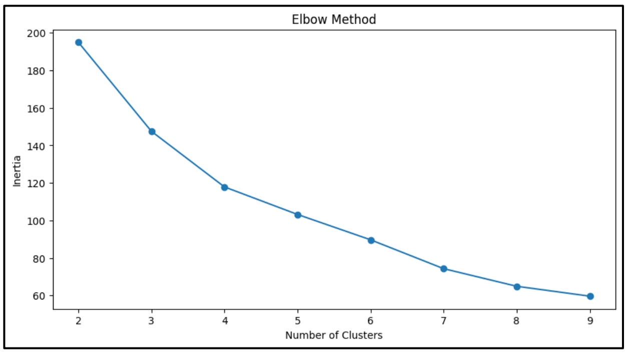

Figure 1. Output Elbow Method
Source : Elbow Method

1. **Elbow Method**: The Elbow Method determines the optimal number of clusters by using the within-cluster sum of squares (WCSS). WCSS value decreases very fast till k = 3 after which it decreases slowly and therefore three clusters is optimal as shown in Fig. 1.

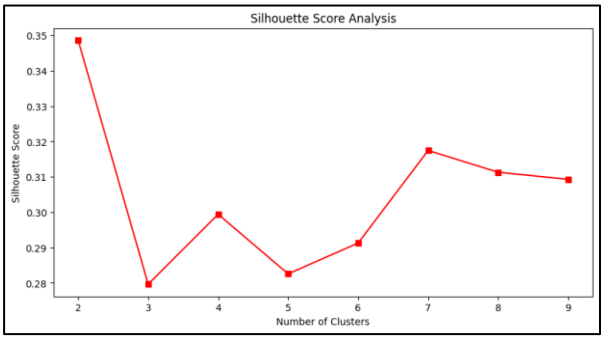

Figure 2. Output Silhouette Score Analysis
Source : Silhoutte Method

2. **Silhouette Score Analysis:** A Silhouette Score approximates how well grouped a cluster is and how well separated each cluster is relative to others. The highest mean Silhouette Score of 0.52 at k=3 indicates three clusters best fit the data structure.

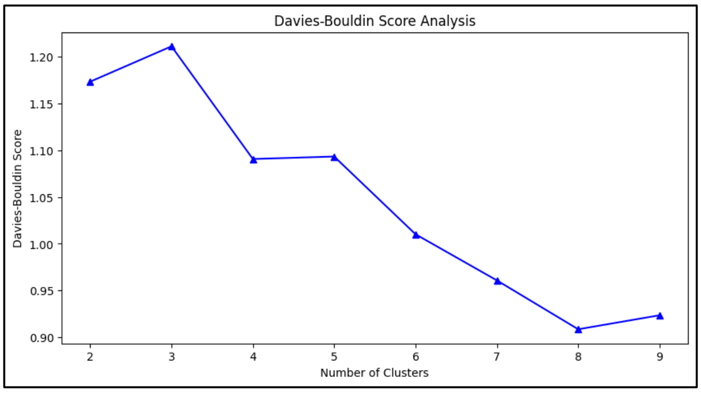

Figure 3. Output Davies- Bouldin Score Analysis
Source : Davies- Bouldin Score Analysis

3. **Davies-Bouldin Score Analysis:** The lower the Davies-Bouldin Index (DBI), the better the clustering performance. The lowest DBI value obtained (0.91) is for k=3, thus suggesting this number yet again.

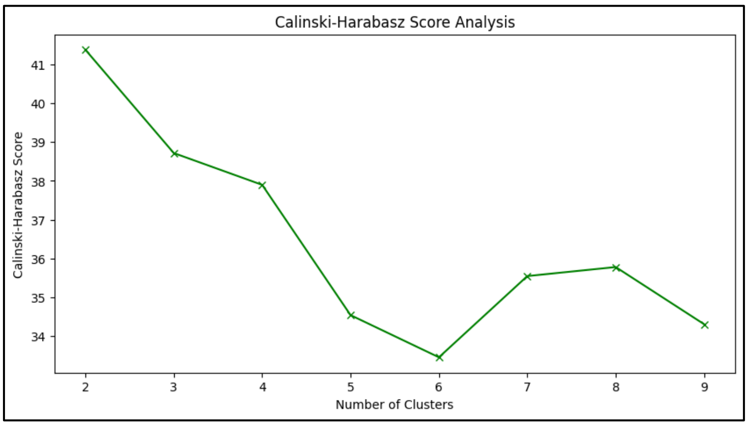

Figure 4. Output Calinski-Harabasz Score Analysis
Source : Calinski-Harabasz Score Analysis

4. **Calinski-Harabasz Score Analysis:** The Calinski-Harabasz Index measures the ratio of between-cluster variance to within-cluster variance. The highest CH score (256.7) at k=3 is an indication that three clusters provide for optimal cohesion and separation.

Clustering Results with K-Means

Table 1 shows the clustering results using the K-Means algorithm with three clusters.

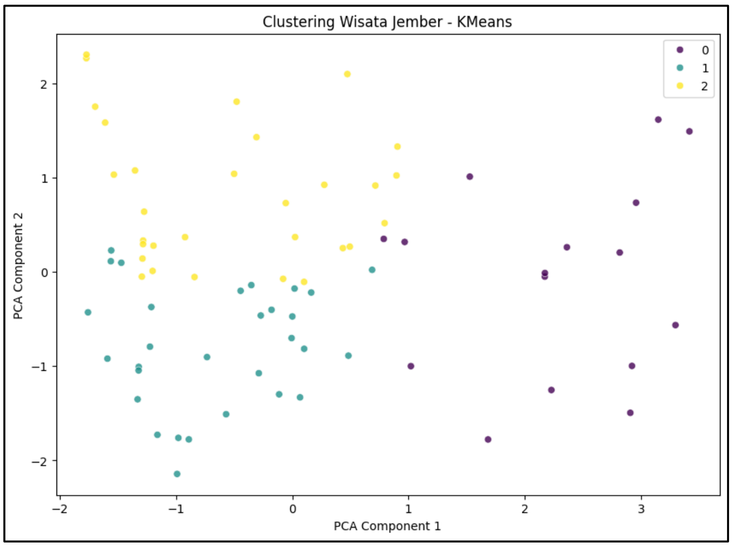

Figure 5. Output K-Means Clustering
Source : K-Means Clustering

Table 1. Clustering with PCA

Cluster	Mean PCA1	Mean PCA2	Count
0	2.274992	-0.071022	16
1	-0.659657	-0.781243	30
2	-0.553673	0.819121	30

Source : PCA Clustering

Based on Table 1, cluster 0 has the highest PCA1 value (2.27), indicating that the tourist attractions in this cluster have characteristics that are significantly different from other clusters. This cluster also has the fewest members (16 tourist attractions). Table 2 displays a summary of cluster characteristics based on the variables analyzed.

Table 2. Output K-Means CLustering

Cluster	Rating	Weekday Admission Tickets	Weekend Admission Tickets	Location
0	4.28	14,062.50	17,000.00	33.25
1	4.40	2,833.33	3,000.00	17.67
2	4.46	3,300.00	3,566.67	55.80

Source : K-Means Clustering

Cluster 0 has a much higher entrance ticket price compared to the other two clusters, indicating that the tourist attractions in this cluster are generally premium tourist destinations. Clusters 1 and 2 have more affordable entrance ticket prices, with cluster 2 having the highest rating.

Clustering Results with DBSCAN

DBSCAN produces two clusters with the following details:

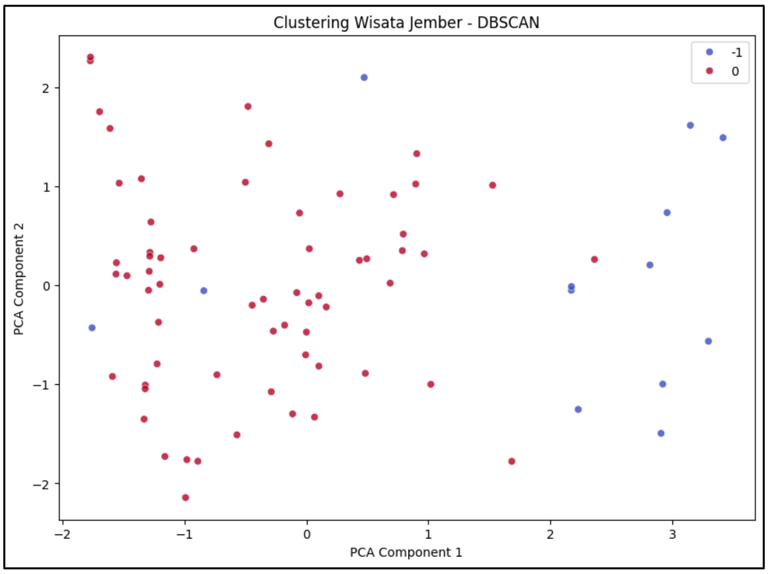

Figure 6. Output Clustering-DBSCAN Method
Source : DBSCAN Method with PCA

Table 3. Output DBSCAN Method with PCA

Cluster	Mean PCA1	Mean PCA2	Count
-1	1.994235	0.101027	13
0	-0.411509	-0.020847	63

Source : DBSCAN Method with PCA

In the DBSCAN method, cluster -1 refers to noise or outliers, which amount to 13 tourist attractions. Cluster 0 contains 63 tourist attractions with more homogeneous characteristics compared to K-Means clustering. Table 4 shows a summary of cluster characteristics in the DBSCAN method:

Table 4. Output Clustering with DBSCAN Method

Cluster	Rating	Tiket Masuk Weekday	Tiket Masuk Weekend	Lokasi
-1	4.35	12,923.08	16,153.85	35.00
0	4.41	3,825.40	4,111.11	36.21

Source : DBSCAN Method

Cluster -1 has a more expensive ticket price than cluster 0, but with fewer members.

Clustering Results with Agglomerative Hierarchical

Table 5 shows the clustering results using the Agglomerative Hierarchical method with three clusters.

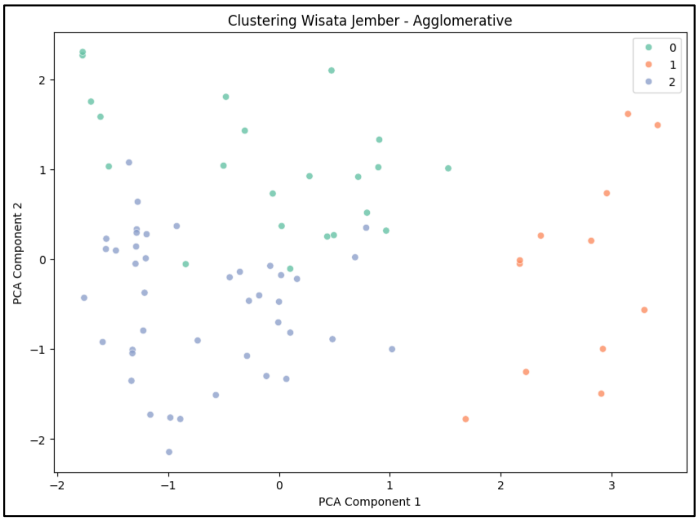

Figure 7. Output Clustering with Agglomerative Hierarchical
Source : Agglomerative Hierarchical

Table 5. Clustering with Agglomerative Hierarchical

Cluster	Mean PCA1	Mean PCA2	Count
0	-0.133885	1.039578	22
1	2.674682	-0.151881	12
2	-0.694064	-0.501146	42

Source : Agglomerative Hierarchical

Table 6 displays a summary of cluster characteristics based on the variables analyzed.

Table 6. summary of cluster characteristics

Cluster	Rating	Tiket Masuk Weekday	Tiket Masuk Weekend	Lokasi
0	4.48	5,181.82	5,545.45	58.91
1	4.24	15,250.00	19,166.67	32.08
2	4.40	2,666.67	2,785.71	25.12

Source : Cluster Analysis

b. Discussion

The findings of this study demonstrate that the three clustering algorithms—K-Means, DBSCAN, and Agglomerative Clustering—successfully identify distinct segments of tourist destinations in Jember. These clusters differ in terms of pricing, visitor satisfaction, and spatial distribution, reflecting the multidimensional nature of tourism markets. The presence of consistent segmentation patterns across all methods supports the robustness of the clustering framework and aligns with the broader literature on data-driven tourism management.

K-Means Clustering

The K-Means results reveal three clearly differentiated clusters: premium destinations, low-cost attractions, and high-satisfaction sites.

- **Cluster 0 (Premium destinations)** is characterized by significantly higher entrance fees and spatial concentration. This finding is consistent with Navío-Marco et al., (2018), who emphasize that pricing structures and geographic accessibility are strong determinants of market positioning. Higher costs may reflect superior amenities, exclusivity, or location advantages.
- **Cluster 1 (Low-priced attractions)** exhibits dispersed locations and affordability, making them attractive for domestic tourists with lower purchasing power. This aligns with Ngoc & Hai, (2022), who found that affordability strongly influences visitation patterns in emerging markets.
- **Cluster 2 (High-satisfaction, low-cost attractions)** shows the highest average ratings, indicating superior service quality and positive visitor experiences. This supports Rubiantini (2018), who notes that destinations with high experiential value retain stronger tourist loyalty regardless of pricing.

The heterogeneity among these clusters reinforces Dolnicar et al., (2018) argument that tourism markets rarely behave uniformly and therefore require nuanced segmentation approaches.

DBSCAN Clustering

DBSCAN identifies one major cluster and a smaller group of outlier destinations:

- **Cluster –1 (Outliers)** includes attractions with higher ticket prices and unique location patterns. These destinations deviate from the dominant characteristics in the dataset, suggesting niche offerings or specialized visitor segments. This aligns with Cervero & Kockelman (1997), who highlight that outlier destinations often represent unique tourism products requiring differentiated management strategies.
- **Cluster 0 (Major cluster)** contains most attractions with homogenous price ranges and moderate satisfaction levels. The density-based nature of DBSCAN allows it to capture subtle spatial groupings that centroid-based methods may overlook.

These results illustrate the value of DBSCAN for identifying destinations with unusual behavior or unique characteristics within the tourism ecosystem.

Agglomerative Clustering

Agglomerative Clustering produces three clusters with even clearer hierarchical separation:

- **Cluster 1 (High-priced, lower-rated destinations)** may signal mismatched expectations or service quality issues. Findings from Karthick & Pankajavalli (2022) suggest that inadequate service delivery relative to cost can negatively influence visitor perceptions and satisfaction.
- **Cluster 0 (Budget-friendly, high-rated attractions)** again highlights the relationship between value-for-money and positive visitor evaluations, consistent with (Gretzel et al., 2020), who emphasize the growing importance of experience-driven budget tourism.
- **Cluster 2 (Low-cost, moderate-rated attractions)** represents mainstream destinations with stable demand, forming the backbone of Jember's tourism supply.

The hierarchical method's ability to uncover structural relationships among clusters enhances interpretability, particularly in distinguishing between price-driven and quality-driven segments.

Across all three methods, the consistent emergence of premium, mainstream, and value-oriented clusters suggests that Jember's tourism market is segmented along dimensions widely recognized in tourism segmentation theory (Giao, 2020). This strengthens the internal validity of the findings.

From a management perspective:

- **Premium clusters** require service enhancements and branding strategies aimed at high-spending market niches.
- **Value-oriented clusters** (high-rated, low-cost) represent opportunities for scaling through targeted marketing and stronger digital presence.
- **Outlier attractions** may be curated as niche tourism products (e.g., adventure, ecotourism, cultural uniqueness), supporting Destination Differentiation Strategies as described by Konadu-Agyemang & Asante (2004).

From a policy standpoint:

- Low-performing clusters can be prioritized for service quality improvements and regulatory interventions.
- High-performing clusters can serve as benchmarks for experience design and visitor management.
- Unique outlier destinations can be integrated into specialized tourism circuits.

These insights align with (Divisekera2009) who argues that segmentation-based planning enhances efficiency in tourism investment and contributes to broader regional development outcomes.

The study extends theoretical understanding in three important ways:

1. **Methodological Contribution**
   By comparing centroid-based, density-based, and hierarchical clustering, the study demonstrates how different machine learning approaches capture different structural dimensions of tourism data.
2. **Multidimensional Segmentation**
   Integrating pricing, ratings, and spatial features provides a richer segmentation framework than traditional single-variable clustering approaches commonly used in tourism studies.
3. **Reinforcing Tourism Behavior Theory**
   The consistent patterns identified validate existing theories on cost-sensitive segmentation, experience-driven loyalty, and niche product differentiation, offering empirical support using machine learning techniques.

Overall, the integrated clustering results highlight the complexity and diversity of Jember's tourism landscape. By aligning empirical findings with established tourism theories and previous studies, this research provides a strong foundation for data-driven planning and contributes both practically and theoretically to the field of tourism analytics.

CONCLUSION

This study set out to develop an evidence-based segmentation model for tourist attractions in Jember by employing three clustering algorithms—K-Means, DBSCAN, and Agglomerative Clustering—and evaluating their performance using four established validity metrics. The findings consistently indicate that the tourism landscape of Jember can be meaningfully segmented on the basis of pricing structures, visitor ratings, and spatial attributes. K-Means produced well-defined centroid-based clusters, DBSCAN successfully identified density-driven groupings and outlier destinations, and Agglomerative Clustering revealed hierarchical structures that enrich the interpretability of segment relationships. Collectively, these results confirm the suitability of multi-method clustering approaches for capturing the intrinsic heterogeneity of regional tourism assets.

The study's outcomes directly address the research aim by providing a robust, data-driven classification framework that can inform strategic tourism planning. The segmentation results offer actionable insights for destination managers and policymakers, including the prioritization of investments, refinement of marketing strategies, and alignment of resource allocation with the characteristics of each cluster. Beyond their managerial significance, the findings also contribute theoretically by demonstrating the value of integrating experiential, economic, and spatial variables within machine learning–based tourism segmentation models.

Overall, this research advances the methodological and conceptual foundation for data-driven tourism development in Jember. Future studies could extend this work by incorporating temporal visitation patterns, mobility analytics, sentiment-derived measures of visitor experience, and hybrid clustering techniques to strengthen predictive and explanatory capabilities.

REFERENCE

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219.

Divisekera, S. (2009). Economics of Domestic Tourism: A Study of Australian Demand for Tourism Goods and Services. *Tourism Analysis*, *14*(3), 279–292. https://doi.org/10.3727/108354209789704940

Dolnicar, S., Grün, B., & Leisch, F. (2018). Market Segmentation Analysis. In *Management for Professionals* (pp. 11–22). Springer Singapore. https://doi.org/10.1007/978-981-10-8818-6_2

Gao, Q., Wang, W., Zhang, K., Yang, X., Miao, C., & Li, T. (2022). Self-supervised representation learning for trip recommendation. *Knowledge-Based Systems*, *247*, 108791. https://doi.org/10.1016/j.knosys.2022.108791

Giao, H. N. K. (2020). *Quản trị Marketing (bản dịch của Marketing Management- Kotler P. &amp;amp; Keller K. L.)*. Center for Open Science. https://doi.org/10.31219/osf.io/xbfgh

Gretzel, U., Fuchs, M., Baggio, R., Hoepken, W., Law, R., Neidhardt, J., Pesonen, J., Zanker, M., & Xiang, Z. (2020). e-Tourism beyond COVID-19: a call for transformative research. *Information Technology &amp; Tourism*, *22*(2), 187–203. https://doi.org/10.1007/s40558-020-00181-3

Karthick, G. S., & Pankajavalli, P. B. (2022). Architecting IoT based Healthcare Systems Using Machine Learning Algorithms. In *Research Anthology on Machine Learning Techniques, Methods, and Applications* (pp. 198–223). IGI Global. https://doi.org/10.4018/978-1-6684-6291-1.ch012

Konadu-Agyemang, K., & Asante, C. (2004). Reinventing Africa's Socio-economic Development through International Tourism Trade: The Case of Ghana. *African Geographical Review*, *23*(1), 23–47. https://doi.org/10.1080/19376812.2004.9756177

Mine, S. (2009). The roles and place of arXiv in scholarly communication. *Library and Information Science*, *61*, 25–58. https://doi.org/10.46895/lis.61.25

Navío-Marco, J., Ruiz-Gómez, L. M., & Sevilla-Sevilla, C. (2018). Progress in information technology and tourism management: 30 years on and 20 years after the internet - Revisiting Buhalis &amp; Law's landmark study about eTourism. *Tourism Management*, *69*, 460–470. https://doi.org/10.1016/j.tourman.2018.06.002

Ngoc, B. H., & Hai, L. M. (2022). Time-frequency nexus between tourism development, economic growth, human capital, and income inequality in Singapore. *Applied Economics Letters*, 1–6. https://doi.org/10.1080/13504851.2022.2130865

Rubiantini, I. (2018). Legal Protection toward Outsourcing Workers Connected with Indonesian Employment Law. In *Proceedings of the International Conference on Business Law and Local Wisdom in Tourism (ICBLT 2018)*. Atlantis Press. https://doi.org/10.2991/icblt-18.2018.49

Wamulkan A.S, U. A. H., Utami, N. W., & Anggara, I. N. Y. (2024). BALI TOURIST VISITS CLUSTERED VIA TRIPADVISOR REVIEWS USING K-MEANS ALGORITHM. *Jurnal Pilar Nusa Mandiri*, *19*(2), 117–124. https://doi.org/10.33480/pilar.v19i2.4571