



Customer clustering analysis in hotel bookings

Análisis de agrupamiento de clientes en reservas de hotel

Jaime Josué Morales-Morales

Doctor en Administración Estratégica. Profesor de la Universidad Autónoma de Sinaloa, México.

ORCID <https://orcid.org/0000-0001-7723-4137>

José Rosario Lara-Salazar*

Doctor en Ciencias de la Administración. Profesor de la Universidad Autónoma de Sinaloa, México.

ORCID <https://orcid.org/0000-0002-7174-4854>

Arturo Yee-Rendón

Doctor en Ciencias Computacionales. Profesor de la Universidad Autónoma de Sinaloa, México.

ORCID <https://orcid.org/0000-0002-9052-6588>

Información del artículo

Recibido:
15/03/2026

Aceptado:
19/07/2026

Publicado:
05/10/2026

***Autor de correspondencia**
jose-larasalazar@uas.edu.mx

Páginas:
088 - 108

<https://rperiplo.uaemex.mx/>

<https://doi.org/10.36677/elperiplo.v0i51.27801>

Abstract

This study aims to conduct a clustering analysis of hotel service customers using unsupervised learning techniques to identify distinct behavioral profiles and inform customer service and loyalty strategies. To this end, the k-means algorithm was applied to group customers based on numerical and categorical variables related to the booking process, customer type, and stay characteristics. The optimal number of clusters (profile groups) was determined using the elbow method, resulting in five clearly differentiated segments, which were validated and interpreted through statistical analysis and t-SNE visualization. Each cluster reflects a distinctive profile, such as families, spontaneous couples, or high-demand couples, with specific patterns in terms of booking lead time, length of stay, group composition, and booking channel. The results provide a solid foundation for designing personalized commercial strategies and optimizing operations in the hospitality sector.

Keywords:

Customer segmentation, cluster analysis, unsupervised learning, k-means algorithm, hotel sector.

Resumen

El presente estudio tiene como objetivo realizar un análisis de agrupamiento de clientes de un servicio hotelero mediante técnicas de aprendizaje no supervisado, con el fin de identificar perfiles diferenciados de comportamiento y mejorar las estrategias de atención y fidelización. Para ello, se utilizó el algoritmo k-medias para agrupar a los clientes a partir de variables numéricas y categóricas relacionadas con el proceso de reserva, el tipo de cliente y las características de la estancia. El número óptimo de clústeres (grupos de perfiles) se determinó mediante el método del codo, identificándose cinco grupos claramente diferenciados, los cuales fueron validados e interpretados mediante análisis estadísticos y visualización con t-SNE. Cada clúster refleja un perfil distintivo, como familias, parejas espontáneas o parejas de alta demanda, que presenta patrones específicos en la anticipación de la reserva, la duración de la estancia, la composición del grupo y el canal de adquisición. Los resultados obtenidos proporcionan una base sólida para el diseño de estrategias comerciales personalizadas y la optimización operativa en el sector hotelero.

Palabras clave:

Segmentación de clientes, análisis de clústeres, aprendizaje no supervisado, algoritmo k-medias, sector hotelero.

Introduction

In the era of digital tourism, data analysis from hotel bookings has taken center stage in the strategic management of the sector. Platforms like TripAdvisor not only facilitate the booking and review process but also generate a rich database that helps hotels better understand their customers. Among the emerging methodologies for leveraging this information, customer clustering stands out—an unsupervised machine learning technique that identifies common patterns among guests based on their booking decisions and reviews (van Leeuwen & Koole, 2021).

Reservations provide valuable information about travelers' habits, such as length of stay, room type, number of people, booking channel, and frequency of visits. This information can be enriched by reviews that users post online, which contain subjective details about the customer experience. Vargas-Calderón et al. (2021) demonstrated that, using natural language processing (NLP) and machine learning techniques, it is possible to automatically evaluate service perception and identify key elements that influence guest satisfaction.

The analysis of this data is based on indicators such as the recency of the last reservation, the frequency of visits, and the amount spent (RFM model), as well as how far in advance the reservation is made and the duration of the stay. Combined with sentiment analysis of texts, these indicators allow hotels to identify profiles such as "spontaneous travelers" or "meticulous planners", thus adjusting their communication and loyalty strategies (Vargas-Calderón et al., 2021).

The application of algorithms such as k-means or hierarchical clustering has been especially useful for effective segmentation. For example, van Leeuwen and Koole (2021) used these methods to segment customers based on their purchasing history and booking behavior, which enabled the design of more personalized campaigns and improved resource use in marketing. In turn, Sánchez-Rivero et al. (2023) used the k-means algorithm to classify tourist accommodations by how they use their web pages, demonstrating that segmentation can also be applied to the analysis of supply, not only demand.

These approaches demonstrate that customer grouping, based on booking data and digital reviews, is an effective tool for anticipating trends, adjusting services, and optimizing the customer experience. The objective goes beyond simply responding to current guest needs; it involves predicting future behaviors to enable the hotel's agile and effective adaptation.

The combination of artificial intelligence techniques, booking data, and customer feedback analysis offers hotels a unique opportunity to reinvent their service model. In an increasingly competitive tourism environment, understanding customers through their bookings is not only a strategic advantage but also a necessity.

Within this framework, this article proposes a clustering analysis of hotel customer profiles using unsupervised learning techniques, based on a real-world booking dataset. This study incorporates both numerical variables (such as booking lead time, length of stay, and group composition) and categorical variables (such as hotel type, distribution channel, and customer type, among others),

aiming to capture a more comprehensive view of customer behavior. Clustering is performed using the k-means algorithm and structurally validated with t-SNE to ensure the consistency and separability of the generated clusters.

The main contribution of this work lies in offering a replicable and operationally interpretable analytical tool that allows the identification of differentiated customer profiles and the guidance of strategic decisions in marketing, loyalty, and operational management, among others, in the organizational field.

Although there is abundant literature on clustering, the novelty of this research lies in combining a broad public sample with a practical segmentation that differentiates demand patterns by hotel type, thereby providing profiles to inform marketing campaigns, pricing strategies, and improvements in the customer experience.

In addition, the proposed methodology facilitates the translation of findings into decisions or operational management, such as promotions aimed at the detected profiles and special services requested by segments. This analysis more clearly shows how the needs vary between the City hotel and the Resort hotel, compared to previous, smaller theoretical works that only allow certain behavioral patterns to be detected.

In this sense, effective clustering and the use of algorithms that identify optimal behavioral patterns allow us to attract and retain customers. Consequently, the practical application of this methodology enables the more precise allocation of technological, financial, and human resources and enhances the competitiveness of hotels in the tourism sector.

The article is structured as follows: Section 2 presents a review of the literature on hotel analysis, with an emphasis on machine learning techniques, particularly k-means. Section 3 describes the data set and details the methodology used for preprocessing, variable selection, determining the optimal number of clusters using the elbow method, and applying k-means. Section 4 presents the results of the clustering analysis, including the numerical and categorical characterization of the clusters (groups of profiles) obtained. Likewise, the interpretation of the identified profile groups and their operational applicability in the context of hotel management are discussed. Section 5 presents specific discussions of the study, its valuation, and replicability; it also outlines the study's conclusions, highlights its strategic implications, and suggests possible directions for future research.

Specialized literature: implications of machine learning and k-means in the hotel sector

The advancement of machine learning, or Machine Learning, has transformed various economic sectors, and the hotel sector has been no exception. Thanks to its ability to identify patterns, automate processes, and generate accurate predictions, machine learning has established itself as a strategic tool in data-driven decision-making (Parvez, 2021). In this context, hotels have begun adopting intelligent analytics technologies to improve the customer experience, optimize operations, and enhance competitiveness in a dynamic digital environment.

Its evolution from rigid programming systems to adaptive models powered by big data has enabled hotels to predict demand, design personalized marketing campaigns, optimize maintenance, and improve fraud detection, among other areas. Furthermore, the integration of machine learning enhances companies' competitiveness in the sector. Therefore, data analysis has become indispensable for the strategy and operation of modern hotels (Cherenkov et al., 2024).

In the current context of high uncertainty, accurate forecasting of hotel demand represents a critical challenge for efficient revenue management. Using clustering algorithms and autoregressive models, this methodology enables more precise anticipation of hotel occupancy (Viverit et al., 2023).

Comments, reviews, and opinions published on digital platforms represent a valuable source of information that, when properly analyzed, can significantly contribute to improving services and the customer experience, as noted by Puh and Bagić (2023), who highlight the strategic potential of data analysis in decision-making within the hotel and tourism sector.

In this regard, and following Bulchand-Gidumal (2022), artificial intelligence has been integrated across the tourism sector through applications such as personalized recommendations and predictive systems, driven by advances in computational capacity.

In a study by Oliveira et al. (2025), machine learning and time-series analysis were applied to predict hotel room prices. They used Prophet, autoregressive, random forest, gradient boosting, extra trees, and neural network models. The case study included 40,000 hotel reservations, and the gradient boosting model yielded the best results, demonstrating its effectiveness in price estimation.

Herrera et al. (2024) developed classification models to predict hotel reservation cancellations. They used dropout layers and K-fold cross-validation to validate the data, and also employed algorithms such as Random Forest, Deep Neural Network, and XGBoost. The latter two provided sufficient data to predict reservation cancellations by hotel service users.

On the other hand, Ampountolas (2025) evaluated machine learning systems using over 209,000 reservations, including more than 46,000 cancellations, from a four-star hotel with historical records from 2015 to 2019. He used several models, including KNN and the XGBoost-Ridge hybrid, which resulted in a low false-negative rate. Therefore, understanding hybridization is necessary for managing hotel revenue. A study by Ampountolas and Legg (2025) that considered short-term forecasts to design pricing strategies for the day after a cancellation, aiming to generate favorable revenue results for the hotel, using LightGBM, XGBoost, Random Forest, and multilayer perceptron artificial neural networks (ANN-MLP), resulted in adaptive strategies within the hotel sector based on historical data to understand customer cancellation patterns.

Sánchez-Medina and C-Sánchez (2020) proposed a method to predict cancellations using machine learning techniques, such as artificial neural networks optimized with genetic algorithms, achieving a cancellation prediction rate of up to 98% and identifying customers with a higher propensity to cancel. This would allow hotels to strengthen their action protocols upon tourist arrival.

Ramzán et al. (2019) analyzed the semantic lexicon of reviews to understand hotel characteristics. Using a big data platform called Hadoop, they can recommend hotel categories to users based on their guest type, whether they are traveling alone or with family.

Therefore, in a changing and increasingly competitive hotel environment, machine learning has become an essential tool, allowing hotels to anticipate events such as cancellations and price fluctuations and to adjust strategies based on the preferences of users with similar profiles. As discussed in previous paragraphs, its use can improve hotel operational efficiency and enhance the guest experience.

The incorporation of advanced models and real-time data analysis allows hotels to make evidence-based decisions, optimize revenue, and design more precise strategies. In this sense, machine learning represents a profound transformation in the management of tourism services.

As documented in previous studies, the use of artificial intelligence in the hotel sector is increasing, particularly to optimize services and support data-driven decision-making. In this context, machine learning algorithms like k-means have proven useful for segmenting customers based on their behavioral patterns, enabling the design of more targeted and therefore more effective strategies. For this reason, it is crucial to delve into recent research to understand its application in the hotel industry.

A study by Eibl et al. (2024) used adaptive conjoint analysis (ACBC) and the k-means algorithm to identify consumer profiles among hotels, based on responses from 801 respondents in Vienna. They analyzed attributes such as price, reviews, location, and sustainability, identifying five user groups with different preferences. These findings allow booking platforms, regardless of their specific platform, to personalize their strategies and thus optimize the user experience.

Kaya et al. (2022) developed a demand forecasting model using Attention-Least-Short-Term Memory (LSTM) that integrates features derived from k-means clustering and neural network representations. This allowed them to capture similarities between hotels based on user criteria, improving demand prediction compared to traditional deep learning models.

In their study, Yildirim et al. (2022) applied the k-means algorithm to segment user and tourist profiles. They identified three key indicators for travel decisions: age, vacation frequency, and the time between booking and check-out. These findings enable hotels to design more personalized offers, thereby improving service quality and competitiveness within the sector.

In this vein, Kumar et al. (2024) examined unsupervised techniques that identify hotel bookings with a propensity for cancellation. Although k-means allowed for initial segmentation, other algorithms, such as DBSCAN and Isolation Forest, were more accurate in detecting booking anomalies in high-risk contexts, such as bookings made many days in advance and bookings without payment guarantees. This reinforces the importance of choosing appropriate models based on the nature of the data obtained from reliable sources.

Alghamdi (2023) applied conventional machine learning techniques, particularly fuzzy k-means, to group data from the TripAdvisor platform and select features based on correlations to determine hotel customer satisfaction for each group. He also used artificial neural networks (ANNs) and decision trees, finding that fuzzy k-means achieved better predictions of customer satisfaction.

Another study that used k-means to group words by sentiment in text was by Zhang and Yu (2017), who applied this technique to classify hotel user reviews, thereby achieving a better textual representation of sentiment. Although k-means proved accurate, the ISODATA method yielded greater precision.

On the other hand, Shu et al. (2023) developed an approach for classifying hotel data from TripAdvisor using k-means and a 2-tuple linguistic model. Their findings suggest that this method can generate denser groups of reviews, allowing hotel organizations to distinguish among customer perceptions and develop strategies more quickly for each group.

In another study analyzing the organizational readiness of hotels, but from the perspective of offering local cuisine, Bondzi-Simpson and Ayeh (2019) used a combination of hierarchical and non-hierarchical clustering algorithms, specifically k-means. The results revealed three clusters that differentiate hotels by their preparedness to offer local dishes. Lower-category hotels showed significantly higher levels of organizational readiness to serve local cuisine compared to higher-category hotels.

In this sense, k-means has proven to be a very effective tool for grouping diverse types of data. This capability allows companies in the hotel industry to identify groups with similar preferences or characteristics, thereby influencing service personalization and strategy design tailored to guests' real needs. Therefore, applying this technique helps optimize the guest experience.

Integrating k-means with other methodologies, such as neural networks or attention models, can also enhance its effectiveness for more complex tasks, such as demand forecasting or reservations. While other algorithms may be more efficient than k-means in certain contexts, it remains a viable and reliable option for handling large volumes of data.

Consequently, its application is not limited to customer analysis but also extends to understanding other organizational aspects, with different applications in the tourism market. In this sense, the use of k-media allows management to adapt to the demands of the hotel sector, thereby driving competitiveness and profitability.

Methods and materials

This section introduces the unsupervised clustering technique as a method of exploratory data analysis, whose purpose is to identify patterns and group similar characteristics. In particular, k-means was used because it can generate clusters (groups) in datasets with N dimensions. This particular method allowed each edge to be assigned to one of the K defined groups, thereby enabling precise interpretation of the structures present in the analyzed data.

Clustering

Clustering analysis is an unsupervised learning technique used to discover hidden patterns or structures in unlabeled datasets. Its main objective is to identify natural groups within the data, optimizing criteria for internal similarity and intergroup dissimilarity (Golalipour et al., 2021). Over time, this methodology has evolved to adapt to diverse contexts, from classic algorithms such as k-means to more sophisticated approaches, including set clustering and robust subset selection.

One of the most significant challenges in applying clustering is the lack of a universal approach for determining which algorithm is best suited to a specific dataset. This difficulty has driven the development of multiple methodological variants, as well as new ways to validate and select features in models. The interpretation of clustering results has gained relevance with the emergence of explainable artificial intelligence, which seeks to clarify how and why algorithms generate certain clusters (Bandyapadhyay et al., 2023).

In addition to its theoretical consolidation, clustering has had a significant practical impact. Its application has extended to strategic sectors such as transportation, health, and tourism, among others, where it is used both for exploratory analysis and data-driven decision-making. In these contexts, clustering does not act as an isolated technique, but rather as part of analytical systems that strengthen the understanding of complex phenomena (Oyewole & Thopil, 2023).

The k-means algorithm

One of the best-known clustering methods is k-means, which involves dividing the data into a fixed number of clusters. This is done by calculating the distance from each data point to the centroid and assigning it to the nearest centroid. The centroids are then updated based on the newly assigned data. This process is repeated until the clusters stop changing or a predefined number of repetitions is reached (Fränti & Sieranoja, 2019; Bagirov et al., 2020).

K-means is popular because it is fast and can be applied to large volumes of data without consuming many resources. However, one of its limitations is the need to decide how many clusters to form. To make this decision, methods such as the elbow method are often used to determine whether the resulting clusters are coherent and well-separated (Schubert, 2023; Rousseeuw, 1987). See Table 1 for the algorithmic description of the elbow method.

Table 1. Step-by-step description of the elbow method algorithm for estimating the optimal number of clusters.

Step	Description
Step 1	Define a range of values of k to evaluate, typically from k = 1 to a reasonable maximum value (for example, k = 10)
Step 2	For each value of k: (a) apply the k-means algorithm to the dataset; (b) calculate the total sum of squared errors within the clusters (within-cluster sum of squares, WCSS), defined as the sum of the squared distances between each point and the centroid of its assigned cluster

Step	Description
Step 3	Store the WCSS values for each k
Step 4	Plot WCSS as a function of k
Step 5	Identify the inflection point (elbow) on the curve, where the decrease in WCSS begins to stabilize. This point represents the optimal value of k, since beyond it, adding more clusters produces only marginal improvement in explained variation.

Source: own elaboration.

In this sense, the k-means algorithm was applied to group the clients into homogeneous groups. It is reiterated that k-means is a clustering method that partitions a set of observations into k groups, assigning each observation to the group whose centroid is closest. It is an unsupervised learning algorithm that groups data based on their similarity (Euclidean distance). Since k-means is based on distances, it is important to scale the variables before applying it; All characteristics were taken to comparable scales so that no one dominates due to its magnitude. In this case, the numerical variables were standardized (by subtracting the mean and dividing by the standard deviation) to ensure they all contribute equally to the calculation of distances.

Dataset

For the present study, the Hotel Booking Demand Dataset (Antonio et al., 2019) was used, which contains 119,390 reservation records across two types of establishments: an urban hotel (City Hotel) and a resort-type hotel (Resort Hotel). Each instance in the set represents an individual reservation and provides detailed information on its characteristics, customer profile, distribution channels used, and final status, including whether it was canceled or completed. Table 2 shows the thematic classification of these variables, grouped by their descriptive function in the data set.

Table 2. Description of the variables contained in the Hotel Booking Demand dataset

Description	Representative variables
Temporal characteristics	arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month
Customer data	adults, children, babies, country, is_repeated_guest
Booking information	lead_time, stays_in_weekend_nights, stays_in_week_nights, booking_changes, total_of_special_requests
Commercial segmentation	market_segment, distribution_channel, customer_type, deposit_type
Outcomes and tracking	is_canceled, reservation_status, reservation_status_date, previous_cancellations, previous_bookings_not_canceled
Internal assignments	reserved_room_type, assigned_room_type, agent, company

Source: own elaboration.

Variables used

Both numerical and categorical variables describing customers' booking behavior were included. Among the numbers considered are:

- Lead_time: Number of days of advance notice with which the booking was made.
- Total_nights: Total nights of stay (sum of weekday nights and weekend nights).
- Adults: Number of adults.
- Children: Number of children (the children and babies columns were combined into a single total_kids variable to represent all minors).
- Total_of_special_requests: Number of special requests made.

The categorical variables included were the following:

- Hotel: Hotel type (City Hotel or Resort Hotel).
- Meal: Meal plan (e.g., BB, HB, etc.).
- Market_segment: Market segment of the booking (Online TA, Offline TA/TO, Corporate, Groups, etc.).
- Distribution_channel: Distribution channel (TA/TO, Direct, Corporate, GDS).
- Customer_type: Customer type (Transient, Transient-Party, Contract, Group).

The selection of variables for cluster analysis was based on criteria of operational relevance, discriminatory power, and prior evidence from consumer behavior studies in the hotel sector. Both numerical and categorical variables were included to capture the multiple dimensions that shape the customer experience and their relationship with the establishment.

Among the numerical variables, lead time allows us to distinguish between customers who plan and those who book spontaneously, a key difference in the implementation of dynamic pricing strategies. Total length of stay (total_nights) and the number of adults and children (adults, total_kids) indicate the group type (individuals, couples, or families). In contrast, the number of special requests (total_of_special_requests) reflects the expected level of personalization, which has direct implications for operations and potential customer lifetime value.

On the other hand, the categorical variables provide essential information about the context and the customer acquisition channel. The hotel type (City Hotel or Resort Hotel) indicates the trip's geographical context and purpose (urban vs. vacation), while the meal plan offers insight into service expectations. Likewise, the market_segment and distribution_channel variables characterize the booking method, which is crucial for evaluating the effectiveness of marketing channels. Finally, the customer_type variable distinguishes between transient, contracted, and group customers, providing information on customer loyalty and demand predictability.

Together, these variables provide a comprehensive view of customer behavior, preferences, and strategic value, enabling the creation of differentiated and actionable profiles.

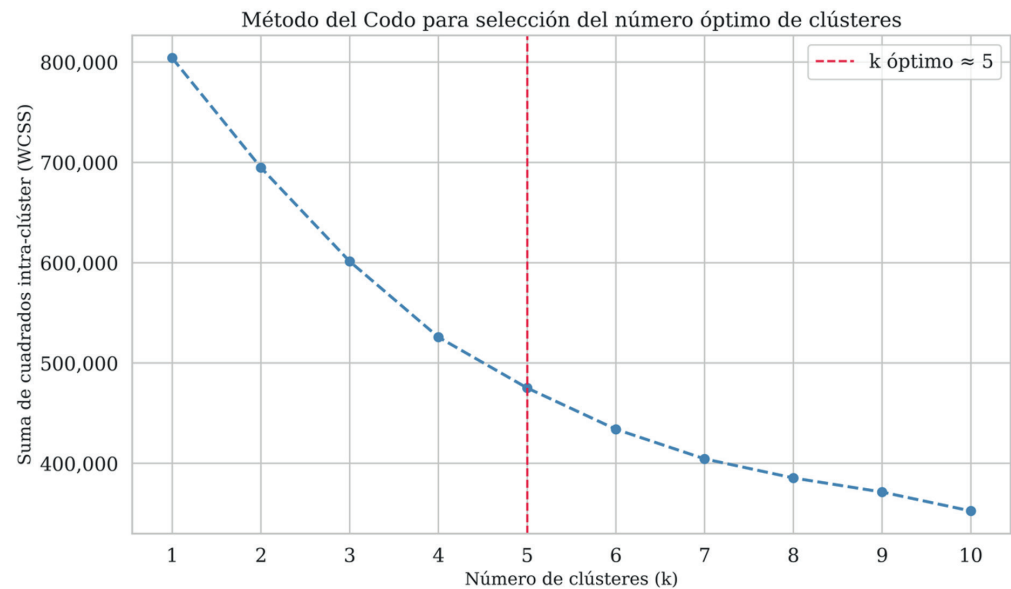
Results

This section presents the results obtained from applying the k-means algorithm, which allowed for the identification of significant clusters within the dataset. These groups or clusters reflect homogeneous behavioral patterns among customers, highlighting relevant similarities in their booking decisions and stay characteristics. As a starting point, the procedure for determining the optimal number of clusters, which forms the basis for subsequent grouping, is explained.

Determining the optimal number of clusters

To determine the number of groups, *k*, the elbow method was used. This method evaluates the inertia (sum of squared errors within groups, WCSS) for different values of *k* and identifies the point at which adding more groups yields diminishing returns. Figure 1 shows the inertia for values of *k* from 1 to 10: Intra-cluster inertia for different values of *k* (elbow method).

Figure 1. Elbow method plot



Source: own elaboration.

As shown in Figure 1 of the elbow method, the within-cluster sum of squares (WCSS) shows a sharp decline between *k* = 1 and *k* = 5, after which the rate of reduction slows significantly. This behavior suggests an inflection point at *k* = 5, where the marginal benefit of increasing the number of clusters diminishes.

Since the elbow method is heuristic, choosing *k* = 5 as the optimal number of clusters is justified by its ability to capture most of the data's underlying structure without introducing unnecessary complexity. Beyond this value, improvements in explained variability are minimal, indicating

diminishing returns to adding more clusters. Based on this analysis, the k-means model was trained with $k = 5$, and each observation was assigned to a cluster. The following sections describe the average characteristics of each group and interpret the resulting customer profiles.

Average numerical characteristics by cluster

Once the clusters(groups)were defined, each group was characterized using a set of key variables, both numerical and categorical, to interpret their main behavioral patterns. This characterization enables the identification of distinct customer profiles by booking lead time, length of stay, group composition, and level of requested personalization. Table 3 summarizes the averages of the most representative numerical variables for each cluster, providing an initial quantitative approximation of the identified segments.

Table 3. Averages of numerical variables by group

Group	Lead Time (days)	Nights of stay	Adults	Children	Special requests
0	~135	~8.5	1.96	0.01	0.53
1	~277	~2.89	1.93	0.0	0.15
2	~70.46	~3.08	~2.09	0.0	1.43
3	~47	~2.43	~1.62	0.0	0.11
4	~85.9	~3.79	~1.96	1.44	0.93

Source: own elaboration.

Table 3 shows that Group 0, designated as long-stay couples, has an average booking time of 136 days and an average extended stay of 8.5 nights. These bookings are for two adults, with no significant number of children, and an average of 0.5 special requests. These clients make extended stays, with moderate planning and low expectations, likely motivated by extended vacations or corporate trips.

Group 1, identified as occasional couples, has the longest booking time among the clusters, averaging 277 days. Despite their high level of planning, their stays are shorter (2.9 nights) and are almost always for two adults, without children. Special requests are minimal (0.15 on average), suggesting bookings managed through agencies, probably as part of tour packages or occasional trips organized well in advance.

In group 2, corresponding to couples with high demand, the booking period is medium (70 days), and stays average 3.1 nights. Reservations include, on average, two adults without children, but stand out for their high level of special requests, with 1.4 per reservation, making this the most demanding group in terms of personalization. These customers may be motivated by romantic getaways or high-end leisure experiences.

Group 3, called spontaneous travelers, has the shortest booking time at just 47 days and the shortest stay at 2.4 nights. Most of these reservations are made by adults traveling alone or as a couple, without children, and with very few special requests (an average of 0.11). This segment comprises the most spontaneous customers, possibly drawn to last-minute promotions or short trips without extensive planning.

Finally, group 4, identified as families, comprises guests who book an average of 86 days in advance and stay an average of 3.8 nights. This is the only group with a significant number of children (1.4 per booking) and exhibits a high level of personalization, with 0.9 special requests. This pattern is characteristic of families who require additional services such as connecting rooms, cribs, or special meals, and who represent an important segment for strategies focused on the family experience.

This grouping allows for the identification of distinct behavioral patterns among guests, which is highly useful for designing segmentation strategies, personalized service, and targeted marketing in the hotel sector. The groups vary not only in length and booking time but also in their demand levels and family composition—key factors for optimizing service offerings and strengthening customer loyalty.

Distribution of categories by group

In addition to numerical analysis, relevant categorical variables were considered to enrich the understanding of each profile group, including hotel type, distribution channel, meal plan, and customer profile. These variables provide key information on the preferences and behaviors of each group, enabling the identification of distinct patterns in how customers access the service, their loyalty, and their expectations. Table 4 presents the relative distribution of these categories by group or cluster, thereby complementing the general characterization of the identified profiles and reinforcing their applicability to strategic decision-making in the hotel sector.

Table 4. Distribution of hotel type and booking channel by group, along with the percentage of repeat guests (TA/TO = Travel Agency/Tour Operator)

Cluster	% City Hotel	% Resort Hotel	% Agency (TA/TO)	% Direct	% Repeat guests
0	24%	76%	86%	12%	1%
1	80%	20%	93%	4%	1%
2	70%	30%	92%	7%	2%
3	71%	29%	70%	18%	6%
4	59%	41%	80%	20%	1%

Source: own elaboration.

Table 4 shows that group or cluster 0, called long-stay couples, is characterized by a strong preference for resort hotels, with 76% of their bookings in this type of establishment, compared to 24% in city hotels. The majority of their bookings are made through travel agencies or tour operators (TA/TO), at 86%, suggesting traditional intermediation. Regarding loyalty, only 1% of the clients in this group are repeat customers, indicating low retention and an opportunity to implement loyalty strategies geared towards extended stays.

Cluster 1, identified as occasional couples, shows a strong preference for city hotels, with 80% of their guests staying at such establishments. This group makes 93% of its bookings through TA/TO, while only 4% use direct channels. Similar to the previous cluster, only 1% of customers are repeat visitors, reinforcing their profile as infrequent travelers, albeit those who plan well in advance.

In cluster 2, which corresponds to couples with high demand, the behavior is more balanced: 70% of their guests choose city hotels, while 30% opt for resorts. Intermediation remains dominant, with 92% of bookings made through agencies. This segment shows a slight improvement in loyalty, with 2% of customers being repeat visitors. Their combination of personalized demand and use of traditional channels suggests a group with high potential value for differentiated services and VIP offers.

Cluster 3, called spontaneous travelers, is distinguished by its dynamism. 71% stay in City Hotels, and while the majority still use travel agencies (70%), this group shows the greatest use of direct channels, with 18% of their bookings managed without intermediaries. Furthermore, this cluster stands out for having the highest relative loyalty among all the clusters, with 6% repeat guests. This profile suggests a segment sensitive to price and time, potentially influenced by last-minute promotions or flexible travel decisions.

Finally, cluster 4, called families, presents a more balanced distribution: 59% choose city hotels and 41% prefer resorts. The majority (80%) book through agencies but also show notable use of direct bookings (20%), the second-highest among the groups. However, their loyalty remains low, with only 1% of guests being repeat guests. This group represents a key opportunity for family retention strategies focused on logistical needs (connecting rooms, cribs, children's menus) and repeat guest benefits.

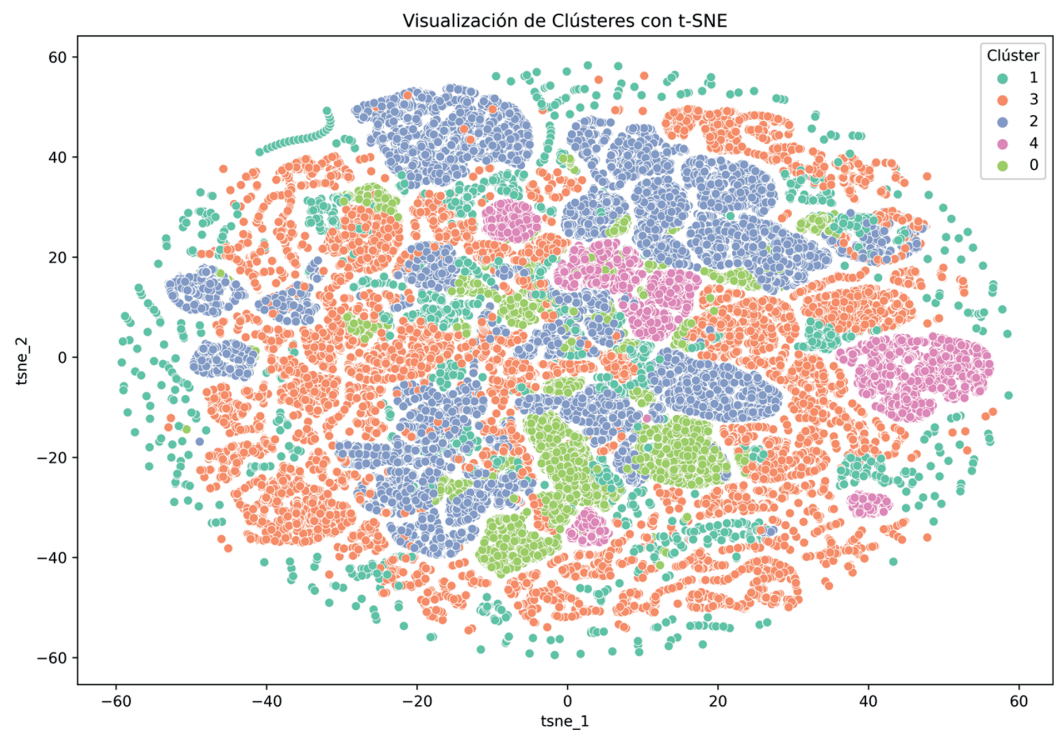
Overall, this analysis reveals distinct patterns of categorical behavior that can guide strategic decision-making in marketing, sales, and customer service within the hotel sector.

Cluster structure analysis

As a complement to the numerical and categorical analyses of the identified groups, the t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality-reduction technique was used to visualize the cluster structure in two dimensions. This technique, widely used in exploratory analysis of multivariate data, allows the projection of high-dimensional observations onto a 2D plane, while primarily preserving local proximity between points (Maaten & Hinton, 2008).

In the context of customer clustering, t-SNE visualization provides a powerful tool for validating the structural consistency of the resulting clusters. Unlike linear methods such as Principal Component Analysis (PCA), t-SNE captures nonlinear relationships and complex neighborhood structures, allowing for a clearer observation of the density, dispersion, and overlap between groups. This visual representation facilitates the qualitative interpretation of the clustering and supports the empirical validity of the clusters generated by the k-means algorithm (see Figure 2).

Figure 2. Distribution of the five customer clusters projected in 2D via t-SNE



Source: own elaboration.

Figure 2 shows the two-dimensional representation of the clustered customers using the t-SNE dimensionality reduction technique, which preserves neighborhood relationships between observations in high-dimensional spaces. This visualization reveals a distinct structure among the five identified clusters, with well-defined dense regions that support the internal consistency of each group. Although some clusters exhibit partial overlap, as is common in domains with fuzzy boundaries such as customer behavior, clearly segregated areas are still observed, indicating distinct patterns in the original data. Overall, this visualization acts as a qualitative validation mechanism that complements the numerical analysis and reinforces the strategic interpretation of the identified profile groups.

Discussion

The research reviewed shows that machine learning has revolutionized hotel management by enabling a deep understanding of customer behavior, optimizing revenue, and consequently improving the customer experience (Parvez, 2021; Cherenkov et al., 2024). In this regard, the k-means algorithm has been used to segment user profiles, identify consumption patterns, and generate personalized strategies (Eibl et al., 2024; Yildirim et al., 2022).

The applications of k-means align with the findings of this research, as this algorithm enabled the identification of five distinct customer groups, or clusters, characterized by variables such as booking

lead time, length of stay, group composition, and the level of personalization required. These groups or clusters are consistent with those found in the work of Alghamdi(2023), and Shu et al. (2023) demonstrate that grouping data based on certain similarities is effective for defining consumer profiles.

Specific studies, such as those by Kaya et al. (2022) and Bondzi-Simpson and Ayeh (2019), support k-means as an effective algorithm for classifying customers or users based on quantitative criteria. It also identifies organizational or cultural aspects, such as the level of human capital preparedness to offer highly specific services. This approach is reflected in the profiles identified in this analysis, in which the groups called “high-demand couples” or “families” present differentiated requirements, indicating the need to anticipate operational needs during hotel stays. Likewise, the “spontaneous travelers” group showed greater loyalty and use of direct booking channels, validating the study by Puh and Bagić (2023) on the strategic value of real-time customer behavior data.

In this regard, the literature also shows that other algorithms can be more effective than k-means in specific contexts, such as high-risk reservations, as noted by Kumar et al. (2024). The results of this research confirmed its high demand and usefulness as an initial segmentation method, coinciding with the findings of Zhang and Yu (2017) and Ramzan et al. (2019), who used k-means to classify emotions in customer reviews. Furthermore, the t-SNE test was used to reinforce the structural validity of the five groups or clusters, providing visual evidence consistent with the qualitative differentiation observed in the analyzed data.

From a business management and strategic administration perspective, the results of this research allow for informed decision-making from a quantitative perspective to generate loyalty strategies, not just transactional ones, and to design personalized services and generate grouped or segmented marketing content. For example, the low retention rates for groups or clusters 0, 1, and 4 present opportunities for organizational management to create highly specific loyalty programs through communication channels focused on hotel service attributes, such as social media, digital banners, and email. Consequently, the use of the k-means algorithm not only provides analytical value but also serves as a bridge to strategic actions that can strengthen competitiveness and generate competitive advantages within the hotel sector.

The profiles obtained through the k-means algorithm represent a tool for designing differentiated marketing, revenue marketing, and customer relationship management strategies.

For example, the couples segment, with high demand, shows a high level of spatial requests, suggesting the need to implement value-added offers and premium experiences. In contrast, the spontaneous traveler segment, characterized by less advanced booking, is well-suited to dynamic pricing strategies and digital campaigns implemented during low-demand seasons in hotels.

As for the families segment, they require specific offers and benefits that contribute to repeat purchases and loyalty. These potential applications are consistent with research by Chen et al. (2023) and Vives and Ostrovskaya (2023), which highlights how customer decision-making can optimize pricing and strengthen customer loyalty in the hotel industry.

A key finding is the low percentage of repeat guests observed in most clusters. This suggests that hotels have significant room to strengthen their loyalty strategies tailored to each segment's characteristics. Clustering models help identify customers with higher potential value, thereby enabling the design of more effective loyalty programs.

In this regard, the resulting profiles can be used to develop more personalized campaigns via booking platforms, email, and social media, thereby increasing the likelihood of repeat business. Thus, cluster analysis serves as a tool to support strategic decision-making aimed at strengthening organizational competitiveness, rather than merely a descriptive tool, according to de Souza et al. (2024).

Regarding the k-means algorithm, it is sensitive to the centroids and the scale of the variables used, meaning that different configurations can yield partially distinct clusters. In this study, the two-dimensional representation known as t-SNE was used, which showed a separation between the five identified groups. This technique constitutes a qualitative validation and does not replace procedures such as cross-validation.

Therefore, future research could utilize cluster analysis and consistency, and compare the performance of k-means with other algorithms such as DBSCAN or hierarchical clustering. These comparisons would allow us to determine the sensitivity of the algorithm used to different combinations of data and thus strengthen the results obtained.

The results demonstrate the usefulness of the k-means algorithm for identifying behavioral patterns in hotel reservations. It is important to acknowledge the methodological limitations. First, the number of clusters was selected using the elbow method, which is widely used but can yield different solutions depending on the dataset. For future research, quantitative internal validation coefficients such as the Silhouette coefficient, the Davies-Bouldin index, or the Calinski-Harabasz criterion can be used to strengthen the segmentation proposed in this study.

Secondly, the clusters were not evaluated using cross-validation techniques. Therefore, data partitioning is necessary in future research to determine if the identified profiles are consistent despite small variations in the dataset. Thirdly, the analysis reflects customer behavior at a specific point in time so that preferences may change due to factors such as holiday seasons, non-holiday seasons, financial conditions, and social influences.

Finally, regarding the generalizability of the results, they must be interpreted in light of the analyzed dataset. Therefore, profiles obtained for different hotel categories should be applied with caution when making generalizations, taking into account the target market, the type of establishment, and other indicators.

Conclusion

The k-means algorithm identified five hotel customer profiles based on operational and behavioral variables, including booking lead time, length of stay, travel group, special requests, hotel type, and booking channels. The results support the idea that the hotel market comprises segments with distinct needs, consumption patterns, and levels of interaction, reinforcing the use of machine learning techniques to enable more objective segmentation processes based on large volumes of data.

Additionally, the structure of the identified clusters was visually validated using t-SNE, which showed consistent separation between the groups in two-dimensional space, strengthening the qualitative interpretation of the resulting profiles. Regardless of how the five clusters are characterized, this study contributes to the understanding that segmentation strategically supports hotel service management, as each identified profile represents distinct opportunities for strategic application.

An example of these segmentation opportunities is the high-demand couples profile, as previously mentioned. These couples represent high commercial value, enabling the development of personalized service strategies and exclusive experiences that would increase both guest satisfaction and average revenue per guest. In contrast, spontaneous travelers, identified by booking with less advance notice, represent an opportunity for implementing revenue management strategies. This involves making price adjustments and developing strategies, such as digital campaigns, to ensure the hotel maintains occupancy even during off-peak seasons.

Regarding the family segment, the opportunities lie in designing packages with complementary services and specific benefits that meet their booking requirements. Similarly, groups or profiles that book well in advance provide the hotel with information about their operational planning, which can reveal opportunities for improvement, enable better evaluation of marketing strategies, and help estimate future demand through these reservations.

From a strategic marketing perspective, the research findings enable the development of more refined communication campaigns that deliver value to each group or profile, taking into account factors such as behavioral patterns, requested services, and so on. Furthermore, the low proportion of repeat guests observed across all clusters highlights the need to strengthen hotel loyalty programs. This could foster long-term customer relationships if these strategies are properly implemented, potentially enabling better allocation of financial resources to run advertising campaigns for each customer profile identified through data mining.

Finally, this research provides evidence of the potential use of data mining and machine learning as tools to support organizational decision-making. Clustering models enable a more precise understanding of guest behavior, which is crucial for developing evidence-based business strategies, such as loyalty programs and revenue management. The study thus offers an analytical framework that can contribute to sustainable competitiveness within the hotel industry.

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

The data used in this study are publicly available from the repository associated with DOI: [<https://doi.org/10.1016/j.dib.2018.11.126>]. No new datasets were generated during this study.

Funding

This research received no external funding.

Declaration of generative AI and assisted technologies in the writing process

Google Translate was used to assist with the translation of the article while the authors were reviewing it. In addition, Grammarly was used to check the style of the entire text.

The research team declares that the use of generative artificial intelligence tools was limited exclusively to style correction and minor adjustments in the manuscript's wording. These AI-assisted technologies did not intervene in the conceptual development, data analysis, interpretation of results, or the generation of substantive academic content.

It is also specified that the participation of AI represented a minimal percentage within the overall writing process, serving solely as linguistic support to improve the clarity and fluency of the text through Perplexity. Full responsibility for the ideas, methods, results, and conclusions presented rests entirely with the authors of the study.

References

- Alghamdi, A. (2023). A Hybrid Method for Big Data Analysis Using Fuzzy Clustering, Feature Selection and Adaptive Neuro-Fuzzy Inferences System Techniques: Case of Mecca and Medina Hotels in Saudi Arabia. *Arabian Journal for Science & Engineering*, 48(2), 1693. <https://doi.org/10.1007/s13369-022-06978-0>
- Ampountolas, A. (2025). Predicting hotel booking cancellations: a comprehensive machine learning approach. *Journal of Revenue and Pricing Management*, 24, 539-550. <https://doi.org/10.1057/s41272-025-00532-x>
- Ampountolas, A., & Legg, M. (2025). A comprehensive approach to enhancing short-term hotel cancellation forecasts through dynamic machine learning models. *Tourism Economics*, 32(2). <https://doi.org/10.1177/13548166251318768>
- Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*, 22, 41-49. <https://doi.org/10.1016/j.dib.2018.11.126>
- Bagirov, A. M., Taheri, S., & Ugon, J. (2016). Nonsmooth DC programming approach to the minimum sum-of-squares clustering problems. *Pattern Recognition*, 53, 12-24. <https://doi.org/10.1016/j.patcog.2015.11.011>
- Bandyapadhyay, S., Fomin, F. V., Golovach, P. A., Lochet, W., Purohit, N., & Simonov, K. (2023). How to find a good explanation for clustering? *Artificial Intelligence*, 322, 103948. <https://doi.org/10.1016/j.artint.2023.103948>
- Bondzi-Simpson, A., & Ayeh, J. K. (2019). Assessing hotel readiness to offer local cuisines: a clustering approach. *International Journal of Contemporary Hospitality Management*, 31(2), 998-1020. <https://doi.org/10.1108/IJCHM-12-2017-0820>
- Bulchand-Gidumal, J. (2022). Impact of Artificial Intelligence in Travel, Tourism, and Hospitality. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken, *Handbook of e-Tourism*. Springer, Cham. https://doi.org/10.1007/978-3-030-48652-5_110
- Chen, J., Xu, Y., Yu, P., & Zhang, J. (2023). A reinforcement learning approach for hotel revenue management with evidence from field experiments. *Journal of Operation Management*, 69(7), 1176-1201. <https://doi.org/10.1002/joom.1246>
- Cherenkov, E., Benga, V., Lee, M., Nandwani, N., Raguin, K., Sueur, M. C., & Sun, G. (2024). From Machine Learning Algorithms to Superior Customer Experience: Business Implications of Machine Learning-Driven Data Analytics in the Hospitality Industry. *Journal of Smart Tourism*, 4(2), 5-14. <https://doi.org/10.52255/smarttourism.2024.4.2.2>
- de Souza, A., Moro, S., & Pereira, R. (2024). Cluster-Based Approaches toward Developing a Customer Loyalty Program in a Private Security Company. *Applied Sciences*, 14(1), 78. <https://doi.org/10.3390/app14010078>

- Eibl, S., Fina, R., & Auinger, A. (2024). Exploring Segmentation in eTourism: Clustering User Characteristics in Hotel Booking Situations Using k-Means. In F. S. Nah, *HCI en empresas, gobiernos y organizaciones*. (Vol. 14720, pp. 157-175). https://doi.org/10.1007/978-3-031-61315-9_11
- Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93, 95-112. <https://doi.org/10.1016/j.patcog.2019.04.014>
- Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, 104, 104388. <https://doi.org/10.1016/j.engappai.2021.104388>
- Herrera, A., Arroyo, Á., Jiménez, A., & Herrero, Á. (2024). Forecasting hotel cancellations through machine learning. *Expert Systems*, 41(9), e13608. <https://doi.org/10.1111/exsy.13608>
- Kaya, K., Yilmaz, Y., Yaslan, Y., Ögüdücü, Ş. G., & Çıngı, F. (2022). Demand forecasting model using hotel clustering findings for hospitality industry. *Information Processing & Management*, 59(1), 102816. <https://doi.org/10.1016/j.ipm.2021.102816>
- Kumar, A., Mahato, B., Kujur, A., Pandey, M., Kumar, B., & Alam, S. (2024). Predicting Cancellations of Hotel Reservations with Unsupervised Machine Learning Methods. *2024 International BIT Conference (BITCON)*, (pp. 1-6). Dhanbad, India,. <https://doi.org/10.1109/BITCON63716.2024.10985602>.
- Maaten, L. V., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579-2605.
- Oliveira, F. B., Silva-Filho, M. W., Barbosa, G. A., Freitas, J. P., Penna, C., & Miranda, P. B. (2025). Machine Learning and Time Series Analysis to Forecast Hotel Room Prices. In A. V. Paes, *Intelligent Systems* (Vol. 15414). Springer, Cham. https://doi.org/10.1007/978-3-031-79035-5_25
- Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, 56, 6439-6475. <https://doi.org/10.1007/s10462-022-10325-y>
- Parvez, M. O. (2021). Use of machine learning technology for tourist and organizational services: high-tech innovation in the hospitality industry. *Journal of Tourism Futures*, 7(2), 240-244. <https://doi.org/10.1108/JTF-09-2019-0083>
- Puh, K., & Bagic', B. (2023). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*, 6(3), 1188-1204. <https://doi.org/10.1108/JHTI-02-2022-0078>
- Ramzán, B., Bajwa, I. S., Jamil, N., Amin, R. U., Ramzan, S., Mirza, F., & Sarwar, N. (2019). An Intelligent Data Analysis for Recommendation Systems Using Machine Learning. *Scientific Programming*, 1-20. <https://doi.org/10.1155/2019/5941096>

- 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. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, 89, 102546. <https://doi.org/10.1016/j.ijhm.2020.102546>
- Sánchez-Rivero, M., Rodríguez-Rangel, M. C., & Ricci-Risquete, A. (2023). K-Means segmentation of tourism accommodation based on the active use of websites: Its application to an emerging destination (extremadura, Spain). *Journal of Vacation Marketing*, 29(4), 654-669. <https://doi.org/10.1177/13567667221117303>
- Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Exploration Newsletter*, 25(1), 36-42. <https://doi.org/10.1145/3606274.3606278>
- Shu, Z., Carrasco, G. R., García-Miguel, J. P., & Sánchez-Montañés, M. (2023). Clustering using ordered weighted averaging operator and 2-tuple linguistic model for hotel segmentation: The case of TripAdvisor. *Expert Systems with Applications*, 213, 118922. <https://doi.org/10.1016/j.eswa.2022.118922>
- van Leeuwen, R., & Koole, G. (2022). Data-Driven Market Segmentation in Hospitality Using Unsupervised Machine Learning. *Machine Learning with Applications*, 10, 100414. <https://doi.org/10.1016/j.mlwa.2022.100414>
- Vargas-Calderón, V., Moros, O. A., Castro, N. G., & Camargo, J. E. (2021). Machine learning for assessing quality of service in the hospitality sector based on customer reviews. *Information Technology & Tourism*, 23(3), 351-379. <https://doi.org/10.1007/s40558-021-00207-4>
- Viverit, L., Yoonjoung, H. C., Nobre, P. L., & Tiana, G. (2023). Application of machine learning to cluster hotel booking curves for hotel demand forecasting. *International Journal of Hospitality Management*, 111, 103455. <https://doi.org/10.1016/j.ijhm.2023.103455>
- Vives, A., & Ostrovskaya, L. (2023). Exploring the impact of hotel attributes and services on price and revenue: A hedonic pricing model approach with a focus on internal and external segmentation and repeat customers. *Tourism Economics*, 30(4), 813-843. <https://doi.org/10.1177/1354816623120037>
- Yildirim, M. E., Murat, K., & Ibrahim, F. (2022). A Case Study: Unsupervised Approach for Tourist Profile Analysis by K-means Clustering in Turkey. *Journal of Internet Computing and Services*, 23(1), 11-17. <https://doi.org/10.7472/JKSII.2022.23.1.11>
- Zhang, X., & Yu, Q. (2017). Hotel reviews sentiment analysis based on word vector clustering. 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI). Beijing, China. <https://doi.org/10.1109/CIAPP.2017.8167219>