NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel customer segmentation

Chloe, Deschanel, number: 20240693

Diogo, Carvalho, number: 20240694

Ingrid, Lopez, number: 20240692

Rúben, Marques, number: 20240352

Group D

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March 10, 2025

INDEX

1. EXECUTIVE SUMMARY

Hotel H is looking to improve its customer segmentation strategy to enhance guest experience, marketing efficiency, and revenue generation. The hotel currently relies on a basic segmentation model based on customer origin, which does not fully capture differences in customer behavior, booking trends, or spending patterns. To address this, we performed a data-driven clustering analysis using advanced segmentation techniques.

Using hospitality customer records, key characteristics such as booking behavior, lead time, revenue contributions, distribution channels, and special requests were analyzed. Through a structured data preparation process, including handling missing values, removing outliers, and feature engineering, a segmentation model was developed using K-Means clustering. This allowed for four distinct customer groups, each with unique characteristics that can inform targeted marketing strategies, pricing models, and improvements to service offerings.

By leveraging these segmentation insights, Hotel H is able to minimize churn rates, increase direct bookings, and optimize promotional efforts. The clusters highlight opportunities to offer personalized services and tailored marketing campaigns, helping the hotel create differentiation in Lisbon's ever-growing and highly competitive hotel market. The resulting clustering provides valuable insights into the behavior and preferences of Hotel H's customers, enabling data-driven decisions that translate into better market positioning and increased profitability.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

According to the World Tourism Organization (2025), Europe is one of the top world destinations, with an estimated 747 million international arrivals in 2024. The industry has seen a remarkable recovery, reaching 99% of pre-pandemic levels and continuing its growth. Particularly, Portugal welcomed 30 million tourists in 2023, contributing to nearly 70 million overnight stays in 2022 (Global Asset Solutions, 2024). This surge has significantly boosted both the economy and the tourism sector's contribution to GDP. Subsequently, the strength of the tourism industry has contributed to Portugal's hotel market growth, with the Lisbon region alone seeing a 14.8% increase in supply. For example, over 4,600 new hotel rooms are currently under construction or in the planning phase (Global Asset Solutions, 2024), reflecting the dynamic evolution of the industry.

Market segmentation is a strategic targeted approach that can enable businesses to efficiently allocate resources and proactively adapt to market trends, thus gaining competitive advantage and customer satisfaction (Singgalen, 2024). However, the hospitality sector has become increasingly competitive, highlighting that traditional customer segmentation, based solely on customer origin, is no longer a valuable option. This can result in less effective marketing efforts and potential revenue loss. In order to move beyond these traditional segmentation methods, it is essential to adopt a data-driven approach, such as data mining, to gain deeper customer insights for better tailored services and promotional strategies, ultimately sustaining business growth (Singgalen, 2024).

Hotel H, situated in Lisbon, Portugal, is part of an independent hotel chain C. This hotel uses a standard market segmentation based on the origin of the customer. However, as previously mentioned and identified by the new marketing manager of Hotel H, this approach is not sufficient in a competitive landscape. Indeed, Hotel H must adopt a more advanced segmentation and data-driven approach, focusing on key customer characteristics such as country of origin, demographics, behaviours and spending habits, as well as consider the multiple distribution channels (e.g. travel agencies, travel operations, online travel agencies, brand websites, etc.). This project will leverage data mining algorithms to better identify and target different groups of customers. In turn, this will allow personalized offers, including pricing, promotions and service offerings, as well as improve satisfaction and effectively allocate resources.

2.2. BUSINESS OBJECTIVES

To proceed with the project, it is essential to clearly determine business objectives for Hotel H. Firstly, one of the main objectives is to increase revenue through advanced customer segmentation that enhances marketing effectiveness. Key metrics to consider for success include:

- Identify a new market segmentation method that allows management to outline specific marketing strategies that effectively capitalize on each segment's strengths.
- Increase the average total revenue per customer, focusing on high value segments identified through cluster analysis.
- Higher engagement and personalization, with success indicators such as increased repeat bookings, longer stays, and customer satisfaction.

- Reduction in booking cancellations and no-shows, leading to higher occupancy and revenue stability.

Secondly, another objective is to increase direct bookings and reduce reliance on online travel agencies (OTAs), hence minimising commission costs. The key success factors include:

- Increase in direct booking share, reducing third-party bookings.
- Improved return on investment (ROI) by assessing the revenue generated against the costs of implementing the new marketing strategy, as well as improvements in operational efficiency, including cost reductions and optimisation of pricing, promotions and service offerings.

By achieving these objectives, Hotel H will enhance its brand reputation and differentiate itself in Lisbon's highly competitive hospitality market through unique guest experiences and stronger loyalty programmes. This data-driven approach will support continuous improvement, with regular reviews of key performance indicators (KPIs) and a long-term strategy that fosters collaboration between sales and marketing.

2.3. Data Mining Objectives

The primary goal is to create a customer segmentation to support more effective marketing, by using clustering techniques like K-Means to group common attributes like demographics, booking patterns, and expenditure patterns. To identify if the marketing strategy is successful, financial (e.g. revenue growth, market share expansion, and ROI) and non-monetary (e.g. employee engagement, operational performance) success metrics can be used. Another goal is to ensure the success of the deployment and maintenance plan.

3. SEGMENTATION PROCESS

3.1. Exploratory Data Analysis

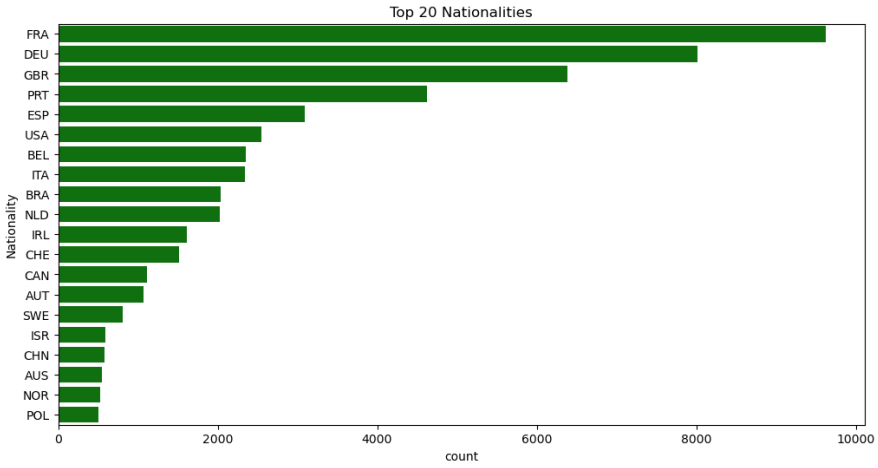
The dataset has 111,733 entry rows and 26 columns including information on customers, their preference and revenue. There are 5 categorical columns, which are ID, nationality, distribution channel, market segment, as well as the hash of customer names and document IDs. The rest of the columns are numerical or binary ones, and pertain to age, average lead time, revenue, bookings, total person per night, total room per night and types of requests.

To facilitate the analysis, the dataset is indexed by customer ID to uniquely identify each entry. However, there are 111 duplicate IDs, which are subsequently removed. When checking missing values, age and document ID contain 4,092 and 932 missing values respectively. Additionally, 8,141 duplicate entries in the document ID field are dropped, and non-essential columns such as market segment (as new segments will be created), document ID and Name (which serve only as identifiers) are removed. To ensure consistency in the data, a coherence checked is performed, identifying the following points:

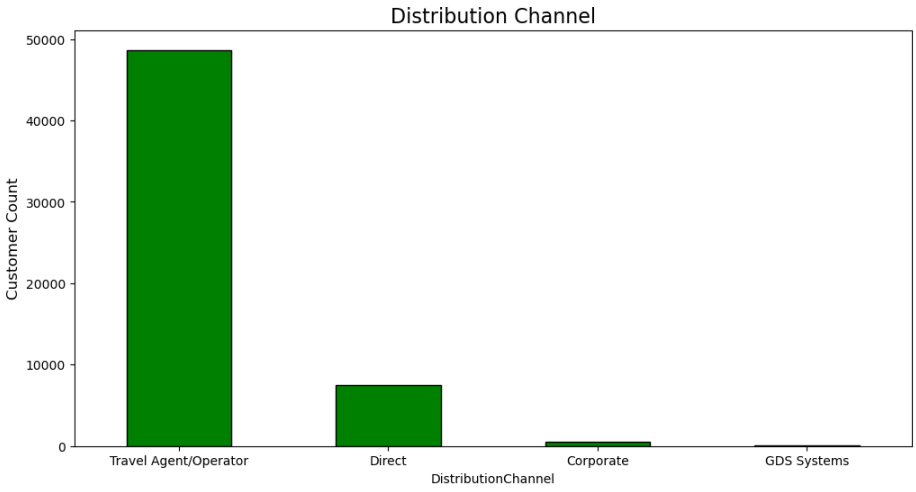
- Approximately 30% of the dataset contain no recorded revenue (i.e. lodging and other revenue).
- 4 bookings exceed the total number of person-nights recorded.

- 10 cases occur where room-nights are greater than person-nights.

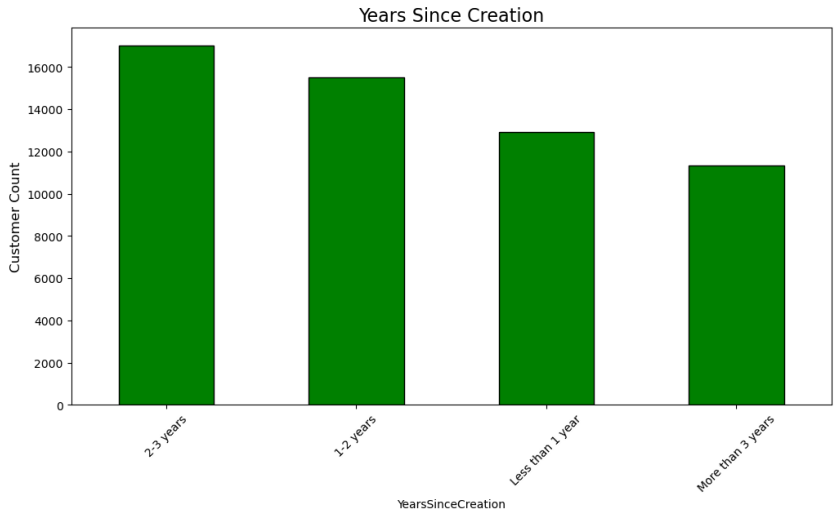
These cases are filtered out of the analysis to avoid inconsistencies and provide accurate information for later segmentation. From the cleaner dataset, the exploratory data analysis was conducted. Nationality revealed to be a categorical variable with high cardinality, with 173 unique nationalities. However, approximately 91% of total customers are represented by only 20 nationalities, with French being the most recurring one, representing around 17%. Over half of the nationalities present have fewer than 10 occurrences.



Regarding distribution channels, the majority of bookings are made through a travel agent or operator, representing around 84% of cases, while only 13% of bookings are made directly with the hotel. This shows that there is a strong dependency on intermediaries.



Hotel H's customers appear to be mainly middle aged, with a mean and median age of 48 years old. Customers also appear to be recent ones, having joined within the last two years, but the data shows a large spread, suggesting variability in customer tenure.

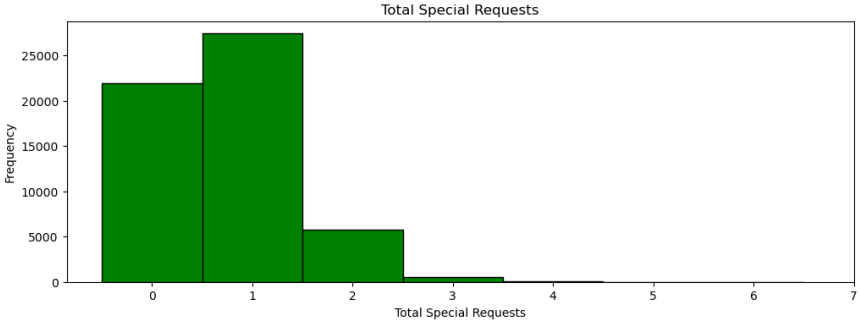


Additionally, customers seem book the three months in advance on average, with an average lead time of 97 days. Generally, there are very little occurrences where a booking is cancelled or there is a no-show, and two thirds of customers have only checked in the hotel once. Moreover, regarding the type of requests, in other words binary variables, customers do not seem to make many special requests, although having a king-sized bed is the most frequent one.

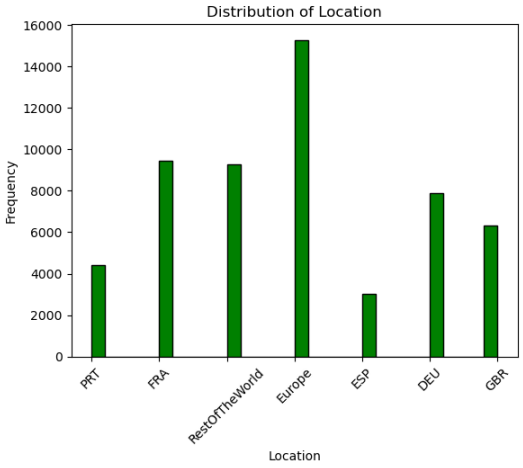
3.2. DATA PREPARATION

The preparation phase ensures that the data is clean, consistent, and suitable for clustering and customer profiling. From the exploration phase, the age variable contained negative variables, which were converted to missing value. These, along with 145 other missing values, were estimated using the KNN Imputer with 5 nearest neighbors to ensure accurate imputations. This variable also contained outliers which were dropped, including 0.1% of customers above the age of 90 years old. Customers below the age of 18 years old were also filtered out, as it is illegal to make bookings while underaged in Portugal. Consequently, these customers are not considered as active ones. Additionally, customers with lodging revenues above 8k, and other revenues above 4k, were filtered out as they are considered as outliers to the normal analysis. The next step involved feature engineering, where the following features were created:

- **Total SR** was introduced to quantify the total number of service requests made by a customer. This was calculated by summing all service request (SR) features. This feature serves as a valuable metric for understanding customer preferences for specific accommodations.



- The **Cancellation Rate** metric was created to assess customer reliability when it comes to booking consistency, and determine whether a customer frequently cancels reservations, which is crucial for revenue forecasting and optimizing booking policies. A high cancellation rate may indicate an unreliable customer, whereas a low rate suggests consistent booking behavior.
- The **Revenue Per Night** feature measures the average revenue generated per room night, providing insights into which customers contribute the most revenue relative to their stays. This metric is essential for distinguishing high-value customers from budget-conscious travelers.
- **Location** was added to classify customers based on nationality while simplifying the categorization of less frequent nationalities. The five most common nationalities in the dataset—France (FRA), Germany (DEU), Portugal (PRT), United Kingdom (GBR), and Spain (ESP)—were retained as distinct categories. All other nationalities were grouped into either Europe or Rest of the World, creating a more intuitive approach to geographic segmentation. This classification enables targeted marketing efforts and service customization based on customer origin.



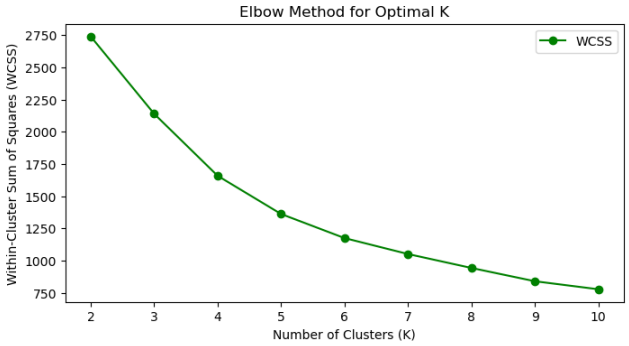
- The **Occupancy Rate** feature was created to measure the efficiency of room usage by customers. This feature is the ratio between Persons per Night and Rooms per Nights, showing how fully a room is occupied during a guest's stay and understanding customer booking patterns.
- The **Average Stay Length** feature was created to measure the duration of a customer's stay. It is the ration of Rooms per Nights and Bookings Checked In. This feature can help determine customers who prefer short or long stays, and thus, optimizing marketing strategies.

Finally, the numerical features used for clustering were scaled using the min-max scaler, which scales features by reducing the range to a minimum and maximum (i.e. between 0 and 1).

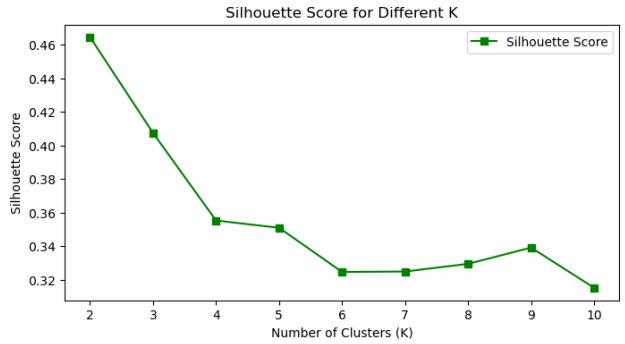
3.3. Modelling and Evaluation

This section explains the model process that was used to segment the customers. We used KMeans Clustering, an unsupervised machine learning model, to segment customers into groups based on the nature of e dataset. The purpose is to segment customers with similar characteristics to

facilitate targeted promotions and customer retention. To identify the best number of clusters, we utilized the Elbow Method that checks the Within Cluster Sum of Squares (WCSS). The method finds the point where the increase in the number of clusters no longer reduces inertia significantly, showing the most effective segmentation.



We also checked the quality of the clustering using the Silhouette Score, which measures the quality of fit of a data point to its assigned cluster. High silhouette scores show strongly defined, relevant groups.



When we compared different cluster sizes ranging from K = 2 to K = 10, we determined that K = 4 provided the best trade-off between interpretability and segment distinction. To assess the effectiveness of our clustering approach, we calculated the R^2, which measures the proportion of variance in the dataset explained by the clusters. Our analysis showed that 43.89% of the variance is explained by the clusters, capturing significant patterns of the dataset.

Additionally, it is important to note that we also tested two other algorithms and compared results. We used density-based clustering, specifically Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which identifies clusters of points with high density while treating low-density areas as noise or boundary regions. During our analysis, this model produced a lower R^2 score than KMeans, so we decided to discard it. We also tested the Mean Shift clustering algorithm, which, like DBSCAN, is a density-based approach that identifies clusters by shifting points toward the densest regions. Unlike DBSCAN, which relies on a fixed density threshold, Mean Shift iteratively moves data points toward local density peaks. In our analysis, this model achieved a higher R^2 score than KMeans, however we decided to discard this approach because the cluster sizes were highly unbalanced with one cluster having almost all customers, while the remaining clusters had only a few. Since KMeans provided the best R^2 result and a good number of clusters, we decided to keep this algorithm and proceed with the analysis and profiling of clusters.

4. CUSTOMER SEGMENTATION

4.1. Cluster Analysis and Profiling

To gain a deeper understanding of cluster behavior, we calculated the mean values for features used in clustering. Additionally, we analyzed the minimum and maximum values within each cluster to ensure distinct customer profiles. The table below provides key data on each cluster.

Cluster 0	Established, Budget-Conscious Customers	• Number of customers: 18,593 • Days Since Creation: 1140 days (Longest customer history) • Average Lead Time: 61 days (Moderate planning time before booking) • Average Stay Length: 3.2 days (Short stays) • Total Revenue: €453 (Lowest revenue contribution) • Age: 48.2 years (Older customer base) • Total Special Requests (SR): 0.68 (Few requests) • Location: Europe • Distribution Channel: Travel Agent/ Operator
Cluster 1	High-Spending, Advanced Planners	• Number of customers: 15,848 • Days Since Creation: 262 days (Newest customers) • Average Lead Time: 102 days (Longest planning period) • Average Stay Length: 3.4 days • Total Revenue: €593 (Highest revenue contribution) • Age: 47.7 years • Total Special Requests (SR): 0.77 (Moderate requests) • Location: Europe • Distribution Channel: Travel Agent/ Operator
Cluster 2	Standard, Consistent Customers	• Number of customers: 15,002 • Days Since Creation: 708 days • Average Lead Time: 60.8 days • Average Stay Length: 3.1 days • Total Revenue: €530 • Age: 47.7 years • Total Special Requests (SR): 0.73 • Location: Europe • Distribution Channel: Travel Agent/ Operator
Cluster 3	Older, High-Commitment Customers	• Number of customers: 6,191 • Days Since Creation: 846 days • Average Lead Time: 263 days (Extreme advance planning) • Average Stay Length: 3.3 days • Total Revenue: €494 • Age: 55.7 years (Oldest customer group) • Total Special Requests (SR): 0.75

		• Location: Germany • Distribution Channel: Travel Agent/ Operator

Cluster 0: Established, Budget-Conscious Customers

These are long-time customers who book their stays moderately in advance and stay for a short period. Their revenue contribution is the lowest, possibly indicating budget-conscious behavior. They make relatively few special requests, suggesting they have simple lodging preferences.

Cluster 1: High-Spending, Advanced Planners

These are relatively new customers who plan well in advance and have the highest revenue contribution. They likely book premium packages or make additional purchases at the hotel. They also make more special requests than other segments, possibly indicating a preference for tailored experiences.

Cluster 2: Standard, Consistent Customers

This cluster represents regular, mid-tier customers. Their booking behavior is predictable with moderate lead time and a relatively stable revenue contribution. They are slightly older and make a fair number of special requests, suggesting a balance between standard and customized experiences.

Cluster 3: Older, High-Commitment Customers

This cluster consists of older, highly organized customers who book significantly in advance. Their revenue is moderate, but they might be loyal, repeat guests. They also have slightly more special requests than other groups, possibly indicating a demand for comfort and specific preferences.

4.2. BUSINESS IMPLICATIONS

Depending on the segmentation of the hotel clientele, different behavioral patterns are observed that facilitate marketing and operational strategies.

Group 0, consisting of long-term, budget-conscious hotel guests, tend to book their stays moderately in advance and are the lowest revenue earners. To retain their loyalty, the hotel must offer targeted promotions, personalized discounts and upsell opportunities for additional services such as dining or spa packages. By maintaining competitive pricing and intensifying customer loyalty programmes, repeat bookings and increased consumption would be encouraged.

In contrast, Group 1 consists of high-spending customers who plan their stays well in advance. Their considerable revenue contribution suggests that they would be willing to invest in premium experiences. To boost their engagement, the hotel should offer exclusive packages, room upgrades, and personalized recommendations based on their past preferences. As these customers book in advance, email marketing campaigns with personalized offers could further consolidate their engagement and encourage direct bookings.

Group 2 represents standard and consistent customers with predictable booking behavior and moderate revenue contribution. As they fall between the economy and premium segment, the hotel

should focus on maintaining consistent engagement through occasional incentives and personalized discounts. By offering special offers for extended stays or personalized experiences based on average length of stay, the hotel could increase the loyalty of these customers.

Lastly, Group 3 is the smallest group and is characterized by older customers who book well in advance and have demonstrated a strong commitment to this hotel. Their preference leans towards comfort and reliability, making them ideal candidates for personalized services such as senior discounts, extended stay offers and personalized customer services. A personalized approach, where their preferences are taken into account and catered to, can reinforce their loyalty and increase the likelihood of future bookings.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. PERSONNEL

The success of the implementation of the clustering model in Hotel H will depend on the participation of a multidisciplinary team that guarantees its correct implementation and maintenance. Given that customer segmentation is a key element in the hotel's commercial strategy, each role within the team has a specific function to translate the results of the model into specific strategic actions. The personnel needed to deploy and maintain include:

- Data scientists - Responsible for the tuning and maintaining the model. They must continuously evaluate segments to ensure that the clusters reflect actual customer behavior patterns. This can be achieved by closely collaborating with the marketing team to interpret and gain expert knowledge of segments.
- Data Engineers – Responsible for the preparation and data flow. This model requires a constant and up-to-date flow of data to stay relevant. For example, they can integrate external data sources, like market data or global trends. This team would ensure that the data is clean, structured and ready for use in cluster analysis.
- Software developers - Responsible for the implementation and accessibility of the model. For customer segmentation to have a real impact on hotel operations, the model must be integrated with the tools used by the sales, reservations and customer service teams.
- IT Specialists - Responsible for security and maintenance. As the hotel deals with sensitive customer data, it is crucial to have adequate security protocols in place. This team must ensure that data security and privacy, complying with regulations such as GDPR (General Data Protection Regulation). They must also oversee the system infrastructure, ensuring that the model works efficiently without affecting other hotel operations.

5.2. MONITORING

Following the implementation of the segmentation model in Hotel H, it is necessary to establish a continuous monitoring process to ensure its accuracy and usefulness in commercial and operational decision making. The goal is to track customer trends, evaluate the effectiveness of segmentation-based strategies, and make data-driven adjustments when necessary. Assessing the permanence of customers in their assigned segments over time is crucial, especially if unexpected changes occur, as these may indicate evolving customer preferences or external market influences. Changes in customers' booking behavior and stay preferences include:

- Changes in average wait time
- Changes in length of stay
- Changes in average length of stay
- Changes in revenue contribution

Such information will aid Hotel H to adapt its marketing campaigns and pricing strategies accordingly. For example, if a specific group starts booking more last-minute stays, the hotel will need to adjust its discount policies or promotional strategies to accommodate these behavioral changes. Beyond individual customer trends, monitoring must also evaluate the overall impact of segmentation on business strategy to determine the success of promotions aimed at attracting bookings from specific groups, whether loyalty incentives encourage repeat stays, and whether pricing adjustments based on group behavior optimize revenue. Finally, if these strategies do not produce the expected results, further adjustments may be necessary. Implementing a proactive monitoring framework ensures that customer segmentation remains relevant, actionable and aligned with business objectives, while continuously improving marketing and revenue management strategies.

6. CONCLUSION

Our analysis of Hotel H's customer data has provided valuable insights into guest spending habits, booking behaviors, and customer loyalty. By identifying four distinct customer clusters, we have selected Cluster 1 (High-Spending, Advanced Planners) and Cluster 3 (Older, High-Commitment Customers) as the most strategically important for the hotel. These clusters represent the most profitable customer groups due to their high revenue contribution and strong engagement with the hotel, despite Cluster 3's smaller size. Loyalty programs, premium service offerings, and personalized promotions should be employed in order to attract and retain these kinds of guests. Group 1, for example, is made up of guests who plan well in advance and generate the most revenue, so offering customized premium packages and early booking incentives can encourage direct bookings and increase their contribution to overall profits. Meanwhile, Group 3, made up of older, more organized travelers, can benefit from personalized services, discounts for longer stays and premium comfort options to enhance their experience and ensure repeat visits.

Due to the growth of the highly competitive hotel market, as demonstrated above, it is necessary to be able to respond to the preferences of high-value customer segments. By using data-driven insights to optimize pricing, marketing and service personalization, we seek to drive direct bookings by avoiding intermediation costs and maximizing revenue per guest. We believe that by pursuing such a strategic approach, the hotel's position in the market will be strengthened and its profitability will improve in the long term.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Although the segmentation model has provided key information about Hotel H's customers, there are opportunities for improvement to optimize its accuracy and applicability to the hotel's business strategy.

Although K-Means was the best choice in terms of trade-off between R^2 and number of clusters, other techniques can be explored to detect more complex patterns. For example, Hierarchical

Clustering could help to visualize relationships between customers and group similar profiles in a more flexible way. This could reveal sub-clusters within existing segments. For example, within high-spending customers, more clusters could emerge based on loyalty duration or booking preferences. The use of Gaussian Mixture Models allows customers to belong to several segments simultaneously, which would help distinguish hybrid travelers, customers who might alternate between long and short stays, for example. A parameter tuning in K-Means could have also been used to test different centroid initializations and selection techniques to improve the reliability of the clusters.

Adding refined demographic and behavioral indicators could improve differentiation between customer types. A potential new variable could be related to travel motives (business vs. leisure travelers). While business travelers tend to prefer short stays, mid-week bookings and premium services, leisure travelers may book in advance, travel in groups and prefer package deals. This distinction would allow Hotel H to optimize promotions and service offerings by segment.

Additionally, while location was considered as part of the profiling phase, the approach could have been more detailed in capturing regional spending trends. The current method grouped nationalities into broad categories (e.g., Europe vs. rest of the world), which provided an overview of customer origins, but could have missed variations in booking behaviors and spending patterns. A more detailed breakdown, for example distinguishing between Western and Eastern European travelers, or between domestic and international customers, would have provided more accurate data on the degree of customer loyalty and its contribution to revenue.

Lastly, although the dataset did not include date and time variables, it was not possible to assess how customer behaviors fluctuate across travel periods (e.g., peak vs. off-peak, vacation vs. off-peak travel). However, an alternative method could have been to indirectly infer seasonal trends by analyzing how booking times and revenues vary by customer location or distribution channel. This could have helped to approximate seasonal influences even without direct date-time data.

7. REFERENCES

Singgalen, Y. A., (2024), "Hotel Customer Segmentation for Marketing Strategy Optimization Using CRAF Framework", *Journal of Business and Economics Research*, Vol 5, No 2, pp. 188-200.

World Tourism Organisation, (2025), *International Tourism Recovers Pre-Pandemic Levels in 2024*, Available at: https://www.unwto.org/news/international-tourism-recovers-pre-pandemic-levels-in-2024 (Accessed on the 7th March 2024).

Global Asset Solutions, (2024), *Portugal Hotel Market Outlook 2024*, Available at: https://globalassetsolutions.com/portugal-hotel-market-outlook-2024/ (Accessed on the 7th March 2024).