NOVA IMS Information Management School

Master's Degree Program in

MDSAA

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel Customer Segmentation

Ana Caleiro

Duarte Marques

Moeko Mitani

Oumayma Ben Hfaiedh

Sarah Leuthner

Group Q

1. INTRODUCTION

The tourism and hospitality industry is a dynamic and rapidly growing sector that plays a significant role in the global economy. As consumer preferences evolve and technological advancements reshape the industry, businesses must continuously adapt to meet the changing demands of travelers. Understanding the key components, trends, and challenges of this industry is essential for sustainable growth and long-term success.

Acquiring new customers is crucial across all sectors. The journey of obtaining new customers starts with gaining as much insight and information as possible from current customers. By dividing the market into separate groups according to geographic, demographic, and behavioral factors, businesses can customize their products and services to fulfill demands, improving customer happiness and retention (Han, 2021). Subsequently, companies can make more informed decisions about market opportunities, product design, and targeting, and moreover, build marketing strategies to boost business revenue.

In this project, we adopted the CRISP-DM framework and implemented the K-Means clustering algorithm. As a result, we identified four distinct customer segments and developed targeted marketing strategies for each group to drive revenue growth of Hotel H. Moreover, we discussed development strategies, maintenance plans, and potential improvements to the model.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Hotel H is a hotel in Lisbon, Portugal, and a member of the independent hotel chain C. By 2015, they had operated four hotels. Until 2018, they used a standard market segmentation model based on customer origin for marketing decisions. Along with the acquisition of new hotels, the hotel chain's board of directors decided to increase its investment in marketing.

2.2. BUSINESS OBJECTIVES

The new marketing manager A of hotel H recognized that the current customer segmentation was not useful for the hotel marketing department because it reflected only one characteristic of the customer: the customer's place of origin.

Without a well-defined segmentation strategy, it becomes challenging to attract new customers and retain existing ones. To optimize marketing strategies and pricing, businesses must tailor their efforts to different customer segments and booking channels. In order to develop an efficient marketing strategy to grow the business, it was decided to identify and analyze new customer segmentations, taking into account geographic, demographic, and behavioral factors.

2.3. BUSINESS SUCCESS CRITERIA

The success of this project will be measured by the ability to identify and analyze new customer segments based on geographic, demographic, and behavioral factors. Additionally, the aim is to develop targeted marketing strategies for each segment to maximize engagement and revenue. The key performance indicator will be the customer retention rate, ensuring that the approach not only

attracts new customers but also strengthens loyalty among existing ones. To achieve this, various models and techniques are tested to enhance the accuracy of customer clustering.

3. METHODOLOGY

In this project, the CRISP-DM model is utilized to segment customers using different clustering techniques, enabling the development of targeted marketing strategies to boost revenue.

3.1. DATA UNDERSTANDING

The data consists of 111,733 data rows with 28 columns recording hotel customer data of 3 years and 10 months (1,400 days). The following chapters will highlight different aspects of the data.

3.1.1. Key Findings and Trends

Table 1 shows the key findings and trends of the customers in the dataset per feature.

Table 1: Key Findings and Trends

Feature	Insights and Trends
Age	Most customers are between 30-60 years old.
New Customers	Most customers joined 1.5 years ago. This means that there are more new customers than old customers.
Average Leadtime	Most customers booked the room 0-100 days in advance. The older the customer, the earlier they book.
Lodging Revenue	Most customers spent 200-650 Euros in total on lodging.
Other Revenue	25% of the customers do not spend money on other revenue. Most customers spent between 30 and 150 Euros on other expenses.
Canceled Bookings	Only 166 customers have cancelled their reservations. This is a good indication for the hotel.
No Shows	Only 57 customers did not show up for their reservations. This is a good indication for the hotel.
Checked-In Bookings	Most customers stayed only one time at the hotel. Repeat customers stayed around 5-20 times at the hotel.
PersonsNights	Most customers have a record of 4-10 people per night.
RoomNights	Most customers have a record of 2-5 rooms per night.
Old Customer Segments	When viewing the distribution of the old market segmentation, the need for a new one is inferred, as around 57% of customers belong to the segment "Other". Only smaller customer segments around 10-15% are distinct: "Direct", "Travel Agent/ Operator", and "Groups".
Customers' Nationalities	The hotel chain C has a worldwide customer base. Most customers are from Europe. The highest represented countries are France, Germany and Portugal. Regarding other continents, many customers are from the USA, Brasil and Canada. Smaller but still significant customer numbers are from Australia, China, Israel and Russia.
Special Requests	Not many customers requested special features for their rooms. If the customer has special requests, it concerns mostly the bed size or in some cases the portion of the room regarding high floor level, quietness or inclusion of cribs. Other special requests were rare.

Distribution Channel	Of all four distribution channels, “Travel Agent/Operator” was used 81% of the time. 15% of the customers booked directly at the hotel and only a fraction (<3%) were corporate customers or GDS System users.
Younger and Older Customers Trends	Regarding Lodging Revenue, underage customers result in lower revenue independent of the distribution channel as they stay only short periods at the hotel. In contrary, the older the customer, the longer they stay, or they stay in the room more.

3.1.2. Data Anomalies

Two features *Age* and DocIDHash show missing values. The percentage of missing values is smaller than 4% for both features. Also, five features are considered to require necessary outlier treatment (see Table 2).

Table 2: Five Features Requiring Outlier Treatment

Features	Outliers' description
Age	Negative values and extra-regional values are identified. The maximum value in the dataset is 123 (although the oldest person in the world is 122). Also, there are minor customers under 18 years old (following the common rules). Minors are disregarded for clustering; although they are reasonable values, they cannot be targeted with any marketing strategy.
LodgingRevenue	There are extreme outliers, for example customers who spent more than 20,000 Euros in total. These outliers might indicate companies using the hotel for their employees on business trips (e.g., frequent user, and different people staying in the hotel). Additionally, the distribution is skewed since there are many zero values in the data.
OtherRevenue	Many outliers with extreme values (8,500 Euros) have been identified. These extreme outliers could be either the amount a company spent on different people in total or an invalid value.
PersonsNights	Outliers (max. 116) indicate either false values or extreme outliers. This could be either frequent customers, families, or companies using the hotel frequently as a client. Additionally, the data is skewed as there are many zero values indicating customers who did not stay at the hotel (invalid)
RoomNights	Outliers (max. 185) indicate either false values or extreme outliers. This could be either frequent customers, families who need more rooms, or companies using the hotel frequently as a client. Additionally, the data is skewed as there are many zero values (invalid data).

Additionally, following data anomalies in the dataset were identified:

- Customers with no bookings
- Customers with same *DocIDhash*
- In *AverageLeadTime*, negative values '-1' are present.

3.2. DATA PREPARATION

3.2.1. Anomalies Treatment

Table 3 shows the treatment strategies for the anomalies found in the previous chapter.

Table 3: Anomalies Treatment Strategies

Features	Problem	Solution / Treatment
Dataset	No values, no bookings, no revenue (~33.000 rows)	Remove from the dataset, as they do not bring any value. After clustering, they will be considered again
Age	Negative values, over 100 values, under 18 values	Removed from dataset; invalid values
DocIDHash	Same Hash codes	Option 1: Duplicates if *Age* and *Nationality* is the same → aggregate rows (using max, sum, average data for other data) Option 2: Hash collision → remains in dataset
AverageLeadTime	Negative values (13 rows)	Removed from dataset; invalid values

3.2.2. Missing Values Treatment

The missing values in *DocIDHash* are not treated as they are not used in the clustering and therefore will not affect the outcome. Due to the amount of data in *Age*, they are treated by using similar records (KNNImputer). This will preserve the structure of the data and will not introduce bias due to assumptions (e.g., linear relationships). An alternative would be taking the median of the nationalities as an estimate. However, as the KNNImputer is computationally efficient, it is implemented.

3.2.3. Data Engineering

Table 4 shows the new features created with their explanation.

Table 4: New Feature Description

New Features	Description
AverageRoomPerStay	It measures how many people (adults + children) stayed per room on average. It helps to understand whether customers prefer shared rooms (higher occupancy) or individual rooms.
TotalSR	It shows the sum of special requests. It indicates if a customer tends to have many special requests or is more low key.
TotalBookings	It shows the total number of bookings no matter the booking status.
BookingCancelationRate	It shows the percentage of booking cancelation.
NoShowRate	It shows the percentage of booking No Show.
CheckinRate	It shows the percentage of checked-in bookings.
TotalRevenue	It shows total revenue.
AvgRevenuePerBooking	It shows the average revenue per booking.
OtherRevenueRate	It determines if a customer spends more on extras (food, spa...) rather than lodging.
RepeatCustomer	It shows if the customer is a repeat customer or not. If the customer has a *BookingsCheckedIn* value higher than 1, the customer qualifies a repeat customer (Boolean value).
AverageStayDuration	It indicates the average duration of stay per guest, with higher values suggesting longer stays.
RevenuePerNight	It represents the average revenue generated per night of stay, with higher values indicating more profitable stays.
RevenuePerRoomNight	It indicates if someone spends more or less in total per night including other revenues beside the lodging (profitability).
LodigingRevenuePerRoomNight	It indicates if someone spends more or less per room (profitability).

DaysBetweenStays	It roughly measures how frequent the user is, low values indicate frequent bookings or recent customers, while high values indicate old infrequent customers.

3.2.4. Outliers Treatment

Instead of relying on the Interquartile Range (IQR), thresholds were manually defined based on boxplots for outliers (see Notebook 3.2.1 Numerical Features Visualization). While this approach introduced a degree of subjectivity and limited generalization, it was necessary due to the high skewness of certain features. As a result, outliers were identified in a context-specific manner, leveraging domain knowledge for each feature.

3.2.5. Scaling

To scale the data the "*MinMaxScaler*" was used because the data is not normally distributed.

3.2.6. Feature Selection

To improve dataset quality and reduce redundancy, highly correlated numerical features (with a correlation **threshold of |0.8|**) were identified and refined. 12 features were not used, including original and newly created features, which held redundant information and no useful information.

3.3. MODELING AND MODEL SELECTION

To create better clusters, we have tried three different algorithms: K-Means, SOM, and DBSCAN. We have also considered trying Principal Component Analysis (PCA) however, this comes with the trade-off of reducing the interpretability, which is critical for our cluster analysis. Therefore, we will not proceed with this technique.

At the end, **K-Means** is chosen because of its high R^2 value. Furthermore, **4 clusters** are identified as the optimal number for the clusters with this dataset.

4. RESULTS AND EVALUATION

The model aligns with the business objectives, particularly in improving customer retention. By applying the model in a real environment, we can generate clusters that reveal insights into customer loyalty, preferences, and behavior. These clusters help identify key segments of customers, enabling targeted marketing strategies to foster loyalty and enhance retention. Through continuous evaluation and refinement of these clusters, the model can support data-driven decisions that directly contribute to meeting business success criteria focused on customer retention.

4.1. CUSTOMER PROFILING

With K-Means algorithm, we identified 4 customer clusters. Table 5 shows the profiles of customers in each cluster. With their profiles, we identified their characteristics as follows:

1. **Cluster 0: Luxury** (27,843 customers - approximately 39% of the total customers)
2. **Cluster 1: Business** (15,921 customers - approximately 22% of the total customers)
3. **Cluster 2: Family** (16,806 customers - approximately 24% of the total customers)

4. **Cluster 3: Elite** (10,923 customers - approximately 15% of the total customers)

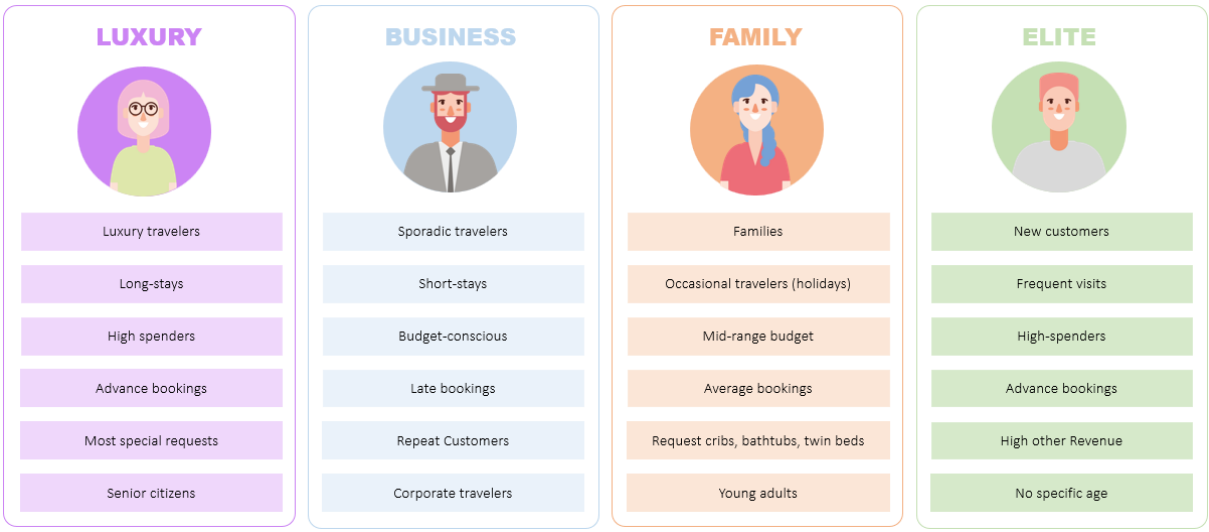

Figure 1: Customer Clusters Characteristics

Table 5: Customer Profiling

Clusters	Descriptions
Luxury	• Largest group, primarily from France and Germany • Oldest customers (45+), with a high percentage of senior citizens • Longest stays and highest lead times before booking • Spend the most on additional services (spa, restaurants, etc.) • Highest number of special requests, often for comfort and accessibility • Long term customers
Business	• Primarily Portuguese customers • Middle-aged 25-64 • Shortest stays with the lowest spending on extra services • Single rooms • Lowest lead time before booking, often last-minute reservations • Accept therefore high room prices • Most repeated customers • Tend to use travel agents and operators but also book directly or as corporate users • Frequent special request: room near an elevator, but no other special requests • Highest percentage of corporate customers
Family	• Customers aged 18-44, often families with children • Frequently request cribs, bathtubs and twin beds • Some long-standing customers in the system, but younger than cluster 0 • Occasional travel habits (large gaps between stays) • Moderate spenders, with limited spending on additional services
Elite	• New customers • Smallest customer group, mostly from Germany, France and Great Britain • Young-aged to senior customers (18+)

	• Spend the most on lodging and other revenue • Customers that visit more frequently although they book way in advance

4.2. MARKETING STRATEGIES FOR EACH CUSTOMER CLUSTER

Considering the profiles of each customer cluster, we have developed following marketing strategies for each of them (see Table 6).

Table 6: Marketing Strategies for Each Cluster

Clusters	Marketing Strategies
Luxury **(Cluster 0)**	• Promote premium packages with spa treatments, fine dining and extended-stay discounts • Offer loyalty programs tailored to long-term customers • Personalized services • Focus on high-end marketing channels and concierge services
Business **(Cluster 1)**	• Provide competitive corporate packages with flexible booking options • Focus on affordability with loyalty incentives for frequent stays • Offer loyalty programs tailored to repeat customers (10^{th} stay for free/high discount) • Optimize partnerships with travel agencies and corporate travel programs • Ensure business-friendly amenities like fast Wi-Fi, meeting rooms, an express check-in/out
Family **(Cluster 2)**	• Offer family-friendly packages, including discounts on children's services • Enhance in-room amenities for families (cribs-kid-friendly dining, family suits) • Promote seasonal campaigns for school vacations and family holidays • Create bundled offers with local attractions and kid-friendly low-cost activities
Elite **(Cluster 3)**	• More social media • Introduce exclusive welcome offers for first-time guests • Encourage long-term loyalty through high-end loyalty programs • Highlight premium accommodations and luxury experiences • Use targeted email campaigns for repeat bookings and early bird promotions

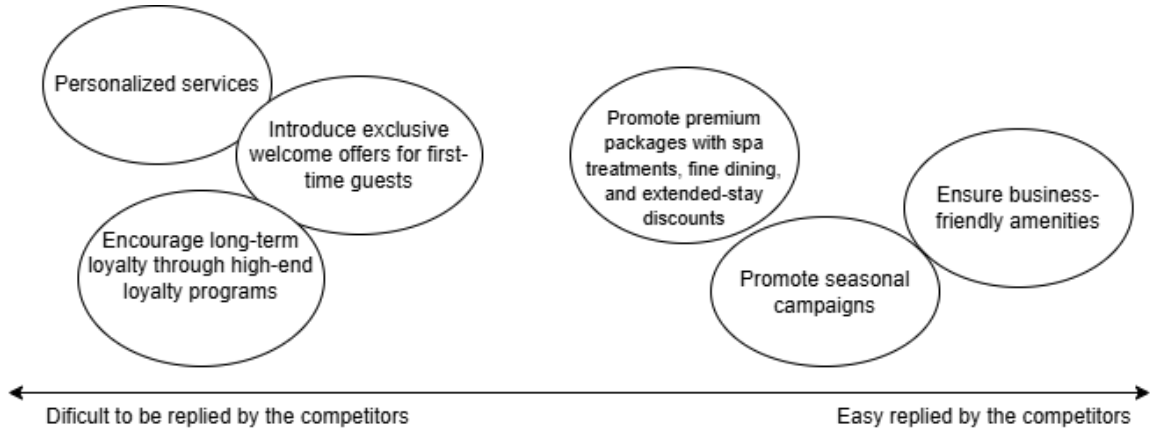

Figure 2: Recommendations for Marketing Strategy (adapted from *Customer Segmentation on Hotel Loyalty Programs, 2019*)

4.3. CASE STUDY: THE STORY OF CUSTOMERS IN EACH CLUSTER

To understand the customers in each cluster in real life, we have developed a use case for each cluster.

4.3.1. Use Case 1: Cluster Luxury (Cluster 0)

Patricia is an elegant Frenchwoman in her seventies who frequently visits Lisbon with her husband. Drawn by the city's sunny weather, she enjoys leisurely stays, and she always books well in advance to secure her favorite suite. Her days revolve around comfort and indulgence - mornings on the terrace, afternoons at the spa, and evenings at elegant restaurants. She values familiarity and excellent service, that is why she returns to the same places where staff know her preferences.

She embodies our first cluster - Luxury, which is the most populated one which is described as senior citizens, luxury long-stay travelers who usually book in advance, and high spenders who make a lot of special requests.

To ensure Patricia's continued loyalty, the strategy focuses on exclusivity, comfort, and personalization. Here are three strategies for her:

1. Premium packages featuring spa treatments, fine dining, and extended-stay discounts.
2. Introduce a loyalty program for long-term guests, with personalized services
3. Focus on high-end marketing channels to keep Patricia engaged and ensure her experiences remain exclusive and tailored.

4.3.2. Use Case 2: Cluster Business (Cluster 1)

João is a 32-year-old corporate professional from Porto, who frequently travels to Lisbon for work. He prioritizes convenience, efficiency, and proximity to his meetings, that is why he often books last-minute stays. He values reliable, well-located accommodation and is willing to pay more for comfort but avoids extra services like spas or fine dining. Typically staying in a single room, he often books through a corporate agency but occasionally makes direct reservations. A familiar guest to the staff, his main request is a room near the elevator for quick access to meetings.

João represents our second cluster - Business which is the practical and business traveler who are mostly Portuguese, who are budget-conscious and who value consistency and efficiency.

Our strategies to retain João are:

1. Provide competitive corporate packages with flexible booking options
2. Focus on affordability with loyalty incentives for frequent stays
3. Offer loyalty programs tailored to repeat customers (10th stay for free/high discount)
4. Optimise partnerships with travel agencies and corporate travel programs
5. Ensure business-friendly amenities like fast Wi-Fi, meeting rooms, an express check-in/out

4.3.3. Use Case 3: Cluster Family (Cluster 2)

Sofia is a 35-year-old mother with a toddler, traveling with her family for occasional vacations. Though her stays are not frequent, she values comfort and convenience. She usually requests cribs, bathtubs, and twin beds for her child to ensure a pleasant stay. She is a loyal guest, returning to the same hotel a few times. As a moderate spender, she focuses on the essentials and spends little on extra services. For her, family-friendly services and a comfortable, stress-free experience are key when choosing accommodation.

She embodies our third cluster - Family, who are young adults that come to the hotel along with their family with a mid-range budget and who are occasional travelers.

Our proposed marketing strategies for her are:

1. Offer family-friendly packages, including discounts on children's services
2. Enhance in-room amenities for families (cribs-kid-friendly dining, family suits)
3. Promote seasonal campaigns for school vacations and family holidays
4. Create bundled offers with local attractions and kid-friendly low-cost activities

4.3.4. Use Case 4: Cluster Elite (Cluster 3)

Alex is a 28-year-old professional from Germany who frequently travels for both work and leisure. He belongs to a smaller, but valuable customer group, often booking stays well in advance. Despite being a relatively new customer, he enjoys frequent visits to his favorite hotel, where he tends to spend the most on lodging and additional services like dining and spa treatments. Whether for business or pleasure, he values comfort and quality, often opting for premium accommodations that match his preferences. He doesn't mind planning ahead to secure the best deals, and his advanced bookings reflect his desire for a stress-free, well-organized experience. Although relatively young, Alex's travel habits align more with those of long-term, loyal customers - he enjoys consistency and the assurance that his accommodation will meet his high standards.

Alex represents our final cluster - Elite, who are luxury travelers and new customers who book in advance and belong to the high spender category.

Our marketing strategies to retain him are:

1. Increased Social Media Presence
2. Exclusive Welcome Offers: Providing special deals for first-time guests to make Alex's initial stay feel rewarding.

3. High-End Loyalty Programs: Encouraging repeat visits by offering a tailored loyalty program that rewards him for his frequent stays.
4. Highlight Premium Accommodations: Showcasing luxury rooms and experiences that align with Alex's preferences for comfort and quality.
5. Targeted Email Campaigns: Sending personalized emails with early bird promotions and repeat booking offers to incentivize future stays.

5. DEVELOPMEMT AND MAINTENANCE PLANS

To optimize the insights provided by the clustering model, Hotel H could use a cloud-based infrastructure to deploy it, following the success case of the Pestana Hotel Group and others. It is well known that the hospitality industry is going through a digital transformation, and cloud computing has been a cornerstone. It is estimated that the amount spent on hotel software will increase by 7-8% in the next 3 years. Starting to use this technology means that the hotel can process new data in real time and update customer segments dynamically in a structured way. In summary, cloud integration will combine all essential operations into a single interface, guaranteeing that all departments are in sync, minimizing errors, and allowing more insights to support decision-making through advanced analytics.

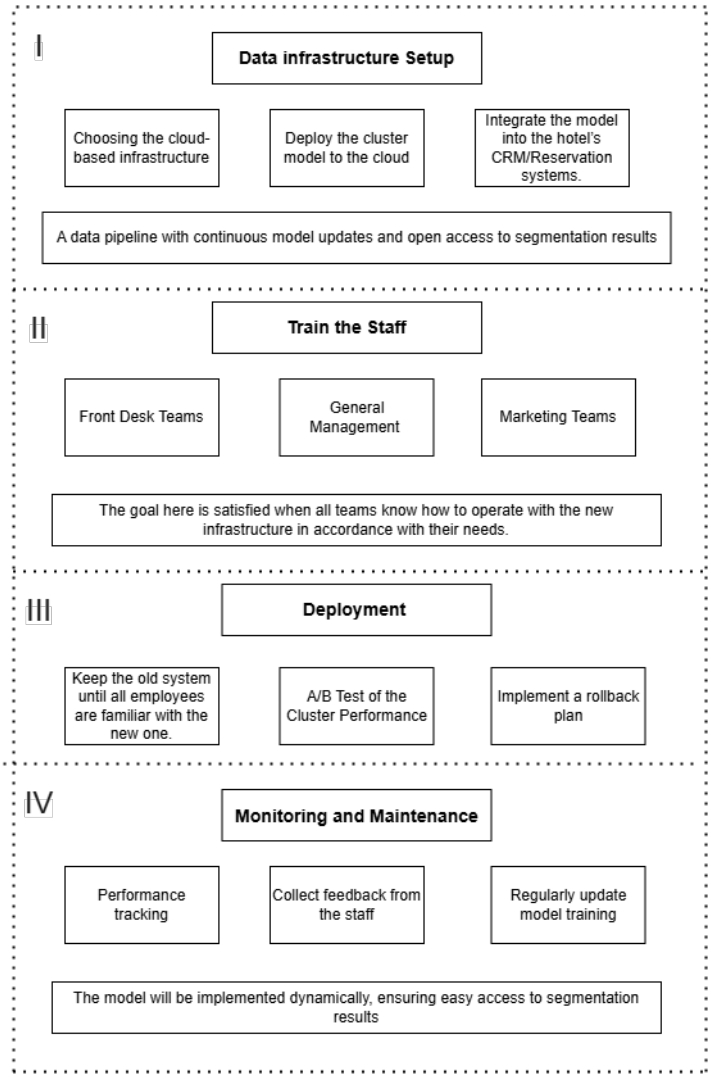

Figure 3: Model Deployment

To implement this strategy, we recommend that the hotel follows four essential steps. The first step is selecting and setting up a cloud-based infrastructure. Some of the most widely used cloud platforms in the industry include Oracle Cloud, Microsoft Azure, and RoomRaccoon, with the latter being specifically designed for the hospitality sector. Cost-effective options like Google Colab Pro or Paperspace can also be used for testing. Next, deploy the clustering model in the cloud and integrate it with the existing system, ensuring that the data pipeline runs smoothly and it's able to continuously be updated.

The second step will focus on the training stage for the staff. All areas in hospitality that add new entries to the data or take insights should grasp the basics of the cloud. A 2-week intensive workshop should be enough for the front desk, general management, and marketing teams to become familiar with the new technology.

In the actual deployment, we recommend that Hotel H keep the old system for at least one month or until the majority of staff is comfortable using the cloud. A/B tests should be run on the clustering model to assess the impact of changes in customer segmentation. Additionally, always have a rollback plan in place to ensure that you can return to a stable system without losing critical data if something goes wrong.

Finally, we reach the continuous stage of monitoring and maintenance of the cloud and the clustering model. The hotel should perform regular performance checks and collect feedback from the staff to understand what can be improved. Since the environment uses real data, regular updates are necessary. However, it is more financially viable to update the cloud monthly instead of in real-time or weekly.

This deployment is only a recommendation and should be discussed and further analyzed with Hotel H.

6. CONCLUSION

To enhance Hotel H's marketing strategy, we began with data exploration and preprocessing, addressing all inconsistencies to ensure data quality. Next, we engineered key features, extracting deeper insights by combining existing variables. For clustering, we tested multiple models—including k-means, self-organizing maps (SOMs), and DBSCAN—evaluating their performance based on R^2. The best model was selected for each segment, leading to the identification of four distinct customer clusters.

Each cluster reveals valuable insights into customer behavior and preferences, providing actionable data to refine marketing strategies and enhance customer engagement.

6.1. BUSINESS IMPLICATIONS

With the implementation of clustering segments and their respective automated pipeline, the hotel can expect an improvement in customer retention. This outcome is achieved due to the ability to develop tailored marketing strategies that will help acquire and satisfy customers, from the very start.

6.2. CONSIDERATIONS FOR MODEL IMPROVEMENT

In the future, while the model is in place, we can improve it with new data entries. This is why the monitoring and maintenance stage is crucial for integrating the model into the business decision-making process. To make the clusters even more accurate, we could also experiment with different types of clustering algorithms.

7. REFERENCES

BI4ALL. (2020). Caso de Sucesso: Pestana Hotel Group. *Directions*. Retrieved March 9, 2025, from https://directions.pt/inteligencia/caso-de-sucesso-pestana-hotel-group

Caetano, J. (2019). Customer Segmentation on Hotel Loyalty Programs: Leveraging Loyalty with Data Mining [Master's thesis, NOVA Information Management School, Universidade Nova de Lisboa]. The Universidade Nova de Lisboa's Repository. https://run.unl.pt/handle/10362/113175?locale=en

Han, H. (2021). Consumer behavior and environmental sustainability in tourism and hospitality: a review of theories, concepts, and latest research. Journal of Sustainable Tourism, 29(7), 1021–1042. https://doi.org/10.1080/09669582.2021.1903019

Tkachova, N., Pylypiv, V., Vinnikova, V., Pylypiv, V., & Zaritska, N. (2021). The relevance of the strategic management of the hotel cluster based on a balanced scorecard. *Research Article, 20*(3). https://www.abacademies.org/articles/the-relevance-of-the-strategic-management-of-the-hotel-cluster-based-on-a-balanced-scorecard-10858.html