

**MDSAA**

Master Degree Program in

**Data Science and Advanced Analytics**

**Business Cases with Data Science**

Case 1: Lisbon Hotel Customer Segmentation

Beatriz, Monteiro, number: 20240591  
Catarina, Gonçalves, number: 20230083   
Margarida, Raposo, number: 20241020   
Teresa, Menezes, number: 20240333

Group P

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

March, 2025

INDEX

[*1.* EXECUTIVE SUMMARY 2](#_Toc192542548)

[2. BUSINESS NEEDS AND REQUIRED OUTCOME 3](#_Toc192542549)

[2.1. Business Objectives 3](#_Toc192542550)

[2.2. Business Success criteria 3](#_Toc192542551)

[2.3. Situation assessment 3](#_Toc192542552)

[2.4. Determine Data Mining goals 4](#_Toc192542553)

[3. METHODOLOGY 4](#_Toc192542554)

[3.1. Data understanding 5](#_Toc192542555)

[3.2. Data preparation 7](#_Toc192542556)

[3.2.1. Errors: 7](#_Toc192542557)

[3.2.2. New Features (KPIs) 8](#_Toc192542558)

[3.3. Modelling 10](#_Toc192542559)

[3.4. Evaluation 10](#_Toc192542560)

[3.4.1. Demographic and Financial Clustering 11](#_Toc192542561)

[3.4.2. PCA (Principal Component Analysis) Results 11](#_Toc192542562)

[3.4.3. UMAP (Uniform Manifold Approximation and Projection) Results 11](#_Toc192542563)

[3.4.4. Final Decision and Analysis 11](#_Toc192542564)

[3.4.5. Outlier Reclassification: 12](#_Toc192542565)

[4. RESULTS EVALUATION 12](#_Toc192542566)

[4.1. Cluster Analysis 12](#_Toc192542567)

[4.2. Marketing Strategies 14](#_Toc192542568)

[5. DEPLOYMENT AND MAINTENANCE PLANS 15](#_Toc192542569)

[5.1. Proposed Next Phases 15](#_Toc192542570)

[5.2. Internal Data Analysis Solution: A Cost-Effective Alternative 16](#_Toc192542571)

[5.2.1. Key Features of Our Solution 17](#_Toc192542572)

[5.2.2. Cost Comparison 17](#_Toc192542573)

[6. CONCLUSIONS 18](#_Toc192542574)

[6.1. Considerations for model improvement 18](#_Toc192542575)

[7. REFERENCES (IF APPLICABLE) 19](#_Toc192542576)

[8. APPENDIX 20](#_Toc192542577)

[8.1. Glossary: 20](#_Toc192542578)

[8.2. Data Preparation Process (detailed) 20](#_Toc192542579)

# EXECUTIVE SUMMARY

In this report, we analyse **Hotel H**, located in **Lisbon, Portugal**, with the goal of understanding the key characteristics and behaviours of its existing customers. Our objective is to segment customers **geographically, demographically, and behaviourally** to develop a comprehensive customer segmentation strategy. This analysis will enable us to provide targeted marketing recommendations to attract new customers while also enhancing engagement with existing guests.

In order to achieve this, we followed the **CRISP-DM methodology**, which is widely used for planning and executing data mining projects. This approach allows us to gain a deeper understanding of customer needs, identify, collect, and analyse relevant data, perform data preparation, and conclude with the modelling, evaluation, and deployment phases, which we will explain further.

From our **clustering analysis**, we identified interesting **patterns and behaviours** that have helped us to tailor targeted marketing initiatives, such as interactive gamification, discounts, free extra night and cultural immersive experience. The main objective of these strategies is to enhance customer satisfaction, improve the hotel reputation by delivering a differentiated experience to the customer and ultimately increase profits. Through all these, we aim to give frequent guests an engaging reason to keep visiting the hotel while also attracting new customers. Additionally, these marketing tactics will help distinguish our hotel from competitors by offering a **unique and personalized experience** that sets us apart in the market.

To move forward, collaboration among several key teams will be crucial. The data analysis team will continue refining the segmentation model and integrating new analytical capabilities, focusing on customer behaviour insights and market trends. The marketing team will provide their expertise in understanding market dynamics and ensuring the customer segments align with real-world behaviours and business objectives. The commercial team will collaborate in assessing how these segments can be applied to optimize B2B strategies and identify long-term opportunities.

To sum up, we will use data-driven insights to improve consumer engagement, speed up the booking process, and refine our marketing strategy. This will guarantee long-term success and establish Hotel H as a **unique option in the market**.

# BUSINESS NEEDS AND REQUIRED OUTCOME

A clear understanding of the business needs and requirements is essential for data-driven solutions to align with real world demands. As such, this chapter lays the foundation for suitable results, by assessing the business’ current situation, defining primary objectives, and outlining how data mining can address key business challenges throughout the completion of the project.

## Business Objectives

Hotel chain C is an independent hotel chain that is not affiliated with any major international industry players, such as Marriott or Hilton. As of 2015, C managed four hotels and, following acquisitions, decided to increase investments in marketing, leading to the creation of a marketing department in 2018. The new marketing manager found the existing customers' segmentation, solely based on sales origin, inadequate for strategic decision-making.

Hotel H is a four-star hotel, a member of chain C, and it operates in Lisbon, Portugal. In the present project, we will be analysing customer data originating from H’s activity.

As such, the primary business objective is to improve customer segmentation to enhance marketing effectiveness, by the end of the analysis. By making use of the demographic and behavioural data, the main goal is to further personalize offers and campaigns, leading to customer satisfaction and revenue maximization.

Additionally, the present project aims to answer related business questions, such as “How can the marketing department target the customers in the different clusters?” and “How should the integration of the new insights occur, in a deployment stage?”.

## Business Success criteria

The new segmentation impact will be essentially measured in two different stages. In a first stage, the goal is to assess the quality of the clustering, by evaluating how well the attained clusters' profiling represents the customers in the respective cluster, and how different customers from distinct clusters are from each other. The more similar customers in the same cluster are from each other, and the more distinct they are from clients in different clusters, the better the segmentation.

The second assessment stage will follow the deployment of the strategic marketing decisions, based on the new segmentation. This evaluation will be done using marketing KPIs, which will reflect the impact of the strategies in the customer behaviour. By making good used of the new customer information and creating a strategy based on it, we hope to see an increase in **Average Daily Rate** (average revenue per occupied room), **Occupancy Rate** (percentage of rooms occupied at a given time), and **Customer return Rate** (percentage of repeated customers) – reflecting an increase in income, due to increasing customer satisfaction. Other KPIs that could be used to measure the marketing efforts’ success would be the **Return on Investment** (total return relative to the cost of a campaign), **Hotel review volume** and **rating**.

## Situation assessment

This project, executed by a team of four business analytics students in collaboration with the hotel marketing department, aims to analyze customer data from Hotel H to generate actionable marketing recommendations. The dataset consists of fixed extracts of customer and potential customer interactions, including check-ins, bookings, cancellations, no-shows, and loyalty program sign-ups via the hotel's website. The project will be developed using Jupyter Notebook for analysis and Streamlit for app deployment, with an estimated duration of two weeks. Confidentiality is ensured by anonymizing individual and company names, and legal compliance regarding data usage will be regarded. The results must be comprehensible, high-quality, and actionable while adhering to security and legal standards.

Key assumptions include the uniqueness of the *DocIdHash* identifier and the general correctness of the data, particularly for unverified nationality entries. Additionally, Hotel H is assumed to operate within a competitive market with similar offerings to its peers, allowing for meaningful marketing insights. Constraints include limited access to deeper operational data and potential challenges in fully understanding business processes due to data abstraction. The project faces risks such as data inconsistencies, time limitations, or a misalignment between analytical insights and real-world business applications. Contingency plans involve iterative data validation and adaptive modelling approaches. While initial resource allocation in marketing campaigns can be costly, especially in a testing phase, the potential benefits include increased customer acquisition, revenue growth, and expanded market share, ultimately strengthening Hotel H’s competitive positioning.

Terminology such as *CRISP-DM (Cross Industry Standard Process for Data Mining)*, *K-means Clustering, Hierarchical Clustering, Self-Organizing Maps (SOM), UMAP (Uniform Manifold Approximation and Projection), Silhouette Score, Market Basket Analysis, R2* will be used*.* Refer to the glossary in the appendix for better understanding of these concepts.

## Determine Data Mining goals

The primary objective, from a data mining point of view, is to significantly improve customer segmentation utilizing all stored customer details. The success of this task will be evaluated based on cluster **cohesion** - how similar customers within a cluster are from one another – and **distinctiveness** – how distinct customers in different clusters are from each other. Beyond well-defined clusters, the insights taken from the segmentation should be useful for marketing experts, meaning they should be interpretable and actionable, ensuring the practical value of the achieved outcomes.

Our **project plan** will begin with analysing the market, using tools like Google Trends to understand market fluctuations and gain knowledge that will help in the next stages. Then, we will look into our current customer base to understand who they are, where they come from, and how they behave. We will also assess why the current segmentation is not meeting its goals. This phase will focus on identifying any data quality issues that need to be resolved before developing the clustering algorithm, which will help create personalized strategies for each customer segment.

# METHODOLOGY

Following the **CRISP-DM methodology**, the process began with **Business Understanding** to define segmentation goals. **Data Understanding & Preparation** involved transforming variables and creating new ones. In **Modelling**, **K-means, Hierarchical Clustering** and **Self-Organizing Maps (SOM)** were applied, evaluated using the **Silhouette Score and R2**. Finally, insights were translated into business recommendations. For implementation details, refer to the notebook *BCwDS\_Case1\_groupP*.

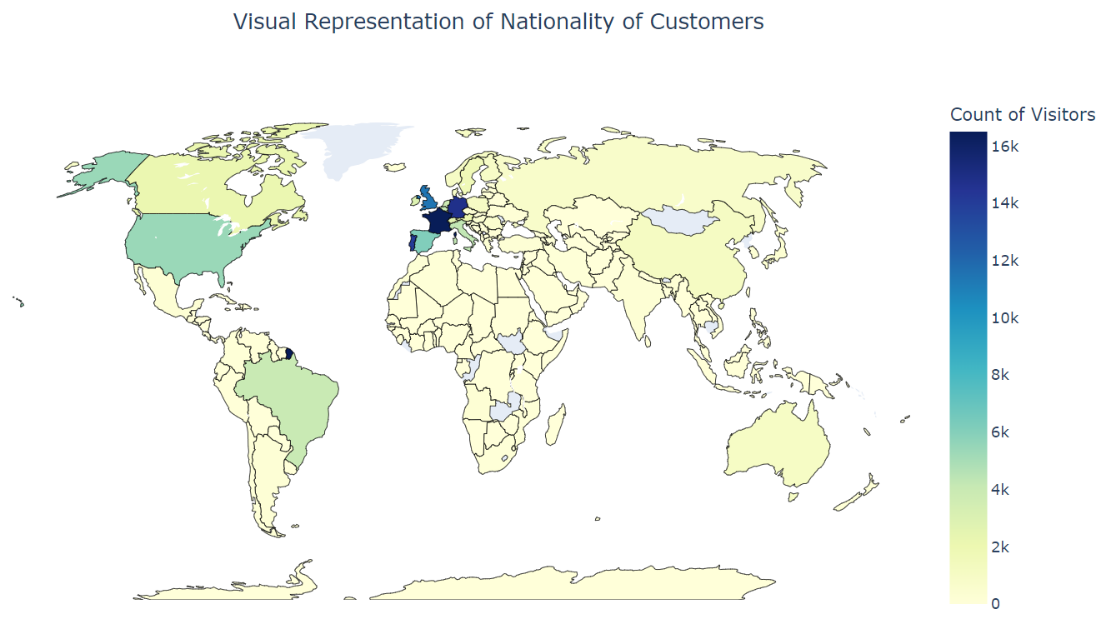
## Data understanding

The dataset provided contains information on 111,733 records, representing either customers or potential customers of a hotel in Lisbon. It comprises 29 variables, including 10 numerical features, 13 boolean indicators, 5 categorical attributes, and a unique identifier for each entry.

The numerical features include attributes such as age, days since customer registration, and revenue metrics. Boolean variables indicate customer behaviours and preferences, while categorical features represent nationality, customer segment classifications, and booking-related attributes. The dataset provides a comprehensive view of customer profiles and booking behaviours, which aligns with the study' segmentation objectives.

The initial exploratory data analysis (EDA) revealed inefficiencies in data types, missing values in the *Age* and *DocIDHash* fields, and a duplicate *DocIDHash* record appearing 3,032 times. Additionally, inconsistencies, duplicate entries, and outliers were identified, emphasizing the need for thorough data cleaning and preprocessing. Furthermore, new variables were created in the feature engineering section to enhance data exploration and extract deeper insights.

As shown in *Figure 1*, our top 4 customer nationalities are **France (15%)**, **Germany (13%)**, **Portugal** (National customers – 13%), and **Great Britain (10%)**.



**Figure 1.** Geographic Map – representation of Customers' Nationality

The Market Basket analysis reveals a notable association between cribs and king-size beds. Approximately **7.98%** of all considered guests request both a crib and a king-size bed, indicating a moderate pairing frequency. Furthermore, the confidence level of **85.41%** suggests that when a guest requests a crib, there is a high likelihood that they will also request a king-size bed. The lift value of **1.14** confirms a positive association, meaning that the presence of a crib request makes it **1.14 times more likely** that a king-size bed will also be requested. Based on these insights, the hotel could enhance its offerings by introducing **family-friendly room packages** that automatically include these features, simplifying the booking process for guests with children. Additionally, **pre-booking suggestions** could be implemented, where guests selecting a crib receive a prompt recommending a king-size bed, further streamlining the reservation experience and improving customer satisfaction.

The **previous segmentation approach** was not optimized, as it categorized customers based mainly on the variable *DistributionChannel*, without considering key influencing factors. The treemap visualization (Figure 2) shows that a significant portion of customers were grouped under an undifferentiated category labeled 'Other' (57%), indicating a lack of granularity in segmentation. Additionally, inconsistencies in customer characteristics across segments suggest that the model did not effectively differentiate between distinct customer behaviours. As a result, the previous approach lacked precision, leading to suboptimal categorization and limited strategic insights. A more data-driven and feature-optimized segmentation strategy is needed to enhance accuracy and improve decision-making. In today's increasingly competitive market, segments should provide us with more than just information about sales origins; they should also give us insights into who our clients are.

A screenshot of a computer screen

AI-generated content may be incorrect.

***Figure 2.*** Treemap of Market Segments and Distribution Channels

During our analysis, we noticed that some users have registered on our loyalty program but have zero bookings cancelled, zero no-show bookings, and zero bookings checked-in, which represent 26,853 clients in total. They represent potential customers who have shown interest in making a reservation with us, presenting a **significant market opportunity** to capture these clients. We will implement a targeted marketing strategy to encourage them to book with us.

Even though this data may not be entirely accurate, since the hotel collects more information during check-in, which never happened in these cases, and some details may have been filled in by default, we still find it important to understand the main characteristics of these clients, as they represent around 30% of our total data.

These potential clients are typically middle-aged adults, around 46 years, who are new to our business, with one year or less of enrolment in our loyalty program. They tend to book primarily through travel agents or operators as their distribution channel, indicating a preference for third-party booking services rather than direct hotel reservations. Their market segment is mainly classified as ‘Other’ suggesting a diverse customer base that does not fall under the conventional corporate, leisure, or business categories.

To convert these potential clients into active guests, we will focus on the main reasons that may lead to them not booking. **Website issues** such as a complex booking process and slow or unresponsive program, **lack of trust** due to bad reviews or unclear security features, **competitors offering** better deals or experiences, indicating the customer is price-sensitive, and lastly, **payment issues** with limited payment methods or payment failures.

We can characterize this group as **price-conscious**, which might also indicate that they are **indecisive** due to the variety of available options, and they find themselves in a loop of comparing the different hotels and getting the best value for their money. They also tend to be **cautious** about security and about user experience because they can discourage users from finishing the booking by unclear trust signals or technical difficulties. Additionally, this group is **sensitive** to friction in the reservation process, thus any inconvenience with payment processing, navigation, or pricing transparency could

To **prevent** this and **retain these potential clients**, we will simplify the reservation process by cutting down on processes and offering clear instructions. We will provide clear pricing that includes all fees and transaction expenses up front. To promote trust, we will also show security certificates and client testimonials. To encourage reservation completion, we are additionally using targeted marketing techniques including email reminders for booked reservations that have been abandoned and special incentives like time-sensitive discounts, complimentary breakfast, room upgrades, or loyalty points.

This strategy would be put into action in **January-March**, as that is when most tourists start to look for/book flights to Lisbon and, of course, make hotel reservations. By launching our campaign during this peak booking period, we aim to capture their attention at the right moment and position our hotel as the preferred choice.

[A graph showing the price of a stock market

AI-generated content may be incorrect.](https://trends.google.com/trends/explore?date=2022-01-01%202025-03-04&q=flights%20to%20portugal,lisbon%20hotels&hl=pt-PT)

***Figure 3.*** Trends in Searches for Flights to Portugal vs. Lisbon Hotels (2022-2025) by GoogleTrends

## Data preparation

### Errors:

* **Challenges with Identifying Unique Guests:** Normally, a *DocIDHash* (hashed version of an ID document) should uniquely identify a single individual. However, if an operator enters placeholders like ".", "Na", or similar generic values instead of a real ID, the hash function will always produce the same *DocIDHash* for those placeholder values. Since the hashing process is **deterministic**, the same input always results in the same hash. If many guests are assigned the same placeholder ID, their *DocIDHash* will also be identical, making it impossible to tell them apart. Without a valid *DocIDHash*, we can't uniquely identify a person, as names alone are unreliable—multiple guests from the same country can have identical names.

Handling Different Scenarios:

* **Case 1: Same Name Appearing Multiple Times Under the Same *DocIDHash*** → This suggests the same person is checking in multiple times. We assume they are the same individual and aggregate their data.
* **Case 2: A Single Name Appearing Once Under a Shared DocIDHash** → This could indicate a shared identifier, like a *company standard used for multiple employees*. We separated these entries to avoid confusion.
* **Case 3: Multiple Different Names Under the Same DocIDHash** → This is the most complex case. Since different people might be assigned the same placeholder ID (e.g. companies), further investigation is needed to determine if they are the same person or truly different individuals.

This was the major quality issue identified. Additionally, we observed that many cases where **Age** was missing were linked to either a missing ID or an ID that was shared across more than 3,000 records. This suggests that these entries are likely to belong to individuals who did not provide their ID. To **improve data quality and integrity in the future**, we recommend **making ID submission mandatory both during online reservations and at check-in**. In cases where this requirement is not met, it would be beneficial to discuss a solution with the **technical** **team**. One possible approach is to ensure that if an ID field is left blank or filled with a predefined placeholder, the **system assigns a randomized unique value** before hashing. This would prevent duplicate *DocIDHashes* caused by placeholders and improve accuracy in identifying unique individuals.

For a detailed overview of the **data preparation process**, please refer to [**Table 1 in the appendix**.](#data_preparation)

### New Features (KPIs)

These newly engineered key performance indicators (KPIs) not only enhance the understanding of customer behaviour and spending patterns—such as customer loyalty, booking tendencies, and revenue contribution—but also serve as tools for our clustering model and targeted marketing strategies.

**Booking Frequency** (measures how often a customer has interacted with the hotel, regardless of the outcome check-in, cancellation, or no-show):



* **High** → Frequent bookers (potential loyal customers). **Low** → One-time or new customers.

**Booking Success Rate** (measures the proportion of bookings that actually resulted in a stay. It helps assess customer reliability):

A black text with black letters

AI-generated content may be incorrect.

* **High** → Reliable customers who mostly check in. **Low** → Many cancellations or no-shows.

**Total Special Requests** (count of all requests indicators a customer has):



* **High** → Customers with specific preferences. **Low** → Customers with fewer demands.

**Total Revenue** (is the total amount spent by the customer, including room charges and additional services. It helps evaluate customer value):



* **High** → High-value customers. **Low** → Budget-conscious or short-stay customers.

**Spending per Booking:**

A black text with black letters

AI-generated content may be incorrect.

* **High** → High-spending customers per booking. **Low** → Budget travellers or guests making frequent low-cost bookings.

**Revenue Per Person Per Night** (measures the average revenue generated per person per night. It helps assess spending behaviour on an individual basis):

A black text on a white background

AI-generated content may be incorrect.

* **High** → Luxury travellers or high-value bookings. **Low** → Budget-conscious guests.

**Average Occupancy** (Both operators involve nights as a unit (i.e., both include the number of nights stayed). Therefore, the night part is common in both the numerator and denominator. As a result, the nights in both cancel out. This leaves us with the ratio of persons to rooms, i.e., how many people are staying per room):

A close-up of black text

AI-generated content may be incorrect.

* **High Average Occupancy** → This indicates more people per room, which could be the case for family/group bookings, or the rooms are being used at higher capacity.
* **Low Average Occupancy →** Fewer people per room, which could be indicative of business travelers, single occupants in rooms, or underutilized space in the hotel.

**ADR (Average Daily Rate)** measures the **average revenue per room night sold**.[[1]](#footnote-2)

A close-up of a sign

AI-generated content may be incorrect.

* **High** → Hotel is charging premium rates. **Low** → Discounts, low season, or budget accommodations.

**Binning *DaysSinceCreation*** (number of days since a customer’s record was first created in the system. It helps understand how long a customer has been associated with the hotel, but it does not necessarily indicate engagement or loyalty) into four categories: **Newly Registered** (Customers who have recently entered the system); **Developing** (Customers with some history but not long-term); **Established** (Customers with a significant duration in the system); **Longstanding** → Customers who have been in the system for an extended period.

**Binning *AverageLeadTime*** (number of days between the booking date and the check-in date) into four categories: **Last-minute** (Bookings made close to check-in); **Moderate Planners** (Guests who book a few weeks in advance); **Advance Planners** (Guests who plan their stays well in advance); **Long-term Planners** (Guests who book months ahead).

## Modelling

Our modelling process aimed to identify the optimal combination of feature selection, data preparation methods, and clustering techniques. We tested different approaches by applying:

* We experimented with multiple normalization and scaling techniques, including Min-Max, Standard, Log, and Power transformation.
* Additionally, we assessed whether Winsorization should be applied to handle outliers.

**Dimensionality Reduction & Clustering (3 Approaches):**

* No Dimensionality Reduction, instead using separate **perspectives for client dimensions**:
  + Financial Perspective: Variables related to spending behaviour (e.g., revenue per person per night, ADR).
  + Demographic Perspective: Characteristics such as age, country of origin, and stay details.
* **Principal Component Analysis (PCA)** to reduce dimensions.
* **Uniform Manifold Approximation and Projection (UMAP)** for nonlinear dimensionality reduction.

**Clustering Methodology:** For each approach, we applied **K-Means Clustering** and **Hierarchical Clustering Variants** (Ward’s method, complete linkage, average linkage, single linkage). The best clustering model was selected based on the **R² score**, which measures how well the clustering explains data variance.

**Selection of Optimal Number of Clusters (K):**

* For the **Financial and Demographic perspectives**, we determined K by analysing the **R² evolution curve** (stabilization point). To merge the two perspectives into one final cluster solution, we used **hierarchical clustering**.
* For **PCA & UMAP**, we tested a range of K values and evaluated cluster quality using: **Elbow Method** and **Silhouette Score; Cluster Sizes & Differentiation** (ensuring clusters are distinct and useful for marketing strategies).

## Evaluation

Our best results were obtained with the combination of **Winsorization and Min-Max scaling** across all three approaches. The steps followed were:

### Demographic and Financial Clustering

* For the **Demographic** perspective, the optimal number of clusters was determined to be **4**.
* For the **Financial** perspective, the optimal number of clusters was **5**. These values were chosen based on the R² evolution curve.
* **Hierarchical clustering was applied to combine both perspectives** (Demographic and Financial), resulting in **4 clusters**.
* However, this approach produced suboptimal results. The categorical encoded columns (which were based on the origin of the clusters) dominated the analysis, leading to a segmentation primarily based on the origin, without distinct differentiation in other numerical variables. This resulted in the creation of clusters that were overly influenced by the origin and lacked variety based on other factors.

### PCA (Principal Component Analysis) Results

The best model using PCA involved **7 components[[2]](#footnote-3)**, which explained **96% of the cumulative variance** in the data. The **optimal number of clusters** identified using **K-Means clustering** was **5**, based on a balance between the R² score and the Silhouette Score:

* **R² score:** 0.50 (explains 50% of the variability in the data).
* **Silhouette Score:** Approximately 0.26, indicating moderate cluster separation.

### UMAP (Uniform Manifold Approximation and Projection) Results

Using **3 components[[3]](#footnote-4)** from UMAP, the optimal number of clusters was also **5**, consistent with the PCA results. This was determined through:

* **R² score:** 0.793466
* **Silhouette Score:** Approximately **0.53**, suggesting a good separation between clusters.

The **Silhouette Score** indicated well-separated clusters, confirming that UMAP is a viable alternative to PCA. It demonstrated strong performance in preserving the data’s structure and separation.

### Final Decision and Analysis

Both **PCA** and **UMAP** identified **5 clusters** as the optimal number for customer segmentation. **UMAP** was ultimately chosen due to its superior ability to preserve the data structure and provide clear separation between clusters. The resulting clusters were well-balanced, with the following distribution: **Cluster 0:** 10,182 samples; **Cluster 1:** 19,946 samples; **Cluster 2:** 22,037 samples; **Cluster 3:** 6,036 samples; **Cluster 4:** 16,573 samples.

The clusters were relatively evenly distributed, ensuring that no single cluster was overrepresented.

### Outlier Reclassification:

When reclassifying the outliers using a decision tree model, the model achieved an **accuracy of 93%**, further validating the robustness of the clustering approach.

# RESULTS EVALUATION

## Cluster Analysis

**Cluster 0** – *Generations on the go*

Cluster 0 stands out for having customers who visited once, frequently travel in family (biggest average occupancy) and for less than a week. This is the youngest cluster, but likely due to their trip’s nature, they enjoy planning in advance (average of 3 months).  They are the second greatest spenders in total revenue, but in fact the ones generating the highest average daily rate (wich means that the they tend to stay all in the same room or stay less nights having in mind the nature of our average daily rate feature). Moreover, they are moderate demanders, standing out in crib request, which aligns with the familiar character of their visit.

They display an above average preference for direct bookings, but as all clusters, their majority books through an agent or operator. This is an extremely diverse cluster, origin wise, with most guests coming from everywhere in Europe, and some coming from other parts of the world.

**Cluster 1** - *Holiday now, bills later*

Customers in cluster 1 are characterized by their high-risk, high-reward nature. They are one of the most likely to cancel or even not show to their booking, while if they do show, they are often returning customers and the largest spenders.

Their preference is to travel in pairs, solo or small families (average of 2 persons per room), and their length of stay is usually close to a week. In comparison to other clusters, their age is average, peaking at older adults. As most clusters, they are mainly spontaneous guests, but a peak in a longer lead time (mainly due to the limitations imposed to extreme values) makes them, on average, some of the biggest planners.

As mentioned before, this is the group of clients bringing in the most total revenue. They usually spend more in rooms and are valuable users in other hotel services. The overall individual revenue per night these customers bring to the business is the second highest, as is the average daily rate.

However, all earnings come with a price, as these are the highest demanders, often requesting quietness and special bed options.

Preferentially, they book through agents or operators but also represent a big portion of customers that book directly with the hotel. Even though nearly 95% of customers in cluster 1 are new, this is still the cluster with the second highest rate of returns, which aligned with their spending habits make them very valuable to the business. Once again, most of these visitors are foreigners, coming from European countries, with especially high concentration of French and German guests.

**Cluster 2** – *Long-weekend retreat*

Customers in cluster 2 are mostly one-time guests in the hotel. They mostly enjoy travelling in pairs or couples, but some solo traveling behaviour was captured as well, and for usually less than a week, likely some extended weekend trips. These customers are the oldest out of all five clusters, and even though their majority are last-minute visitors, nearly 2/5 of the big long-term planners in the dataset can also be found in this group, making them the group with the highest overall lead time.

Regarding their spending habits, these customers do not enjoy spending a lot. In fact, they are the second to last group in terms of total revenue brought into the business and the last in terms of revenue per person, per night. Besides the few spending they do during their stay, in comparison to other groups, they often make requests for special bed options and quietness. A detail that stands out is the fact that they are the group that most requests twin beds, which might be an indication of a strong presence of non-couple pairs.

Most customers in this cluster show a preference for booking trough an agent or operator, but due to its large size, it does represent as well the highest percentage, out of every group, of customers booking directly with the hotel. Close to 100% of guests are new customers.

All in all, origin wise the cluster 2 is very diverse, with approximately 80% of customers coming from various European countries. Despite this overpowering European presence, this is the segment with the largest concentration of North American and customers from other non european origins.

**Cluster 3** – *The budget friends*

The cluster 3 contains customers that are at least one-time guests, with few to none repeated visits. They are mostly pairs or couples, characterized by their short stay.

These are most economical travellers, generating the least amount of total revenue out of all customer segments. Their low spending is likely connected to the brief duration of their trip.

Another important factor is that this is one of the youngest clusters, peaking in middle-aged adults. They are mostly spontaneous travellers, with the overall lowest average lead time.

Visitors in cluster 3 are amongst the most demanding. Like most, they prefer to book through agents or operators and are mainly new customers. There is a clear concentration of German clients, as well as other European countries.

**Cluster 4** – *Frequent visits, solo style*

Customers in cluster 4 are characterized by their repeated bookings, as these are the guests with the highest returning rate, aligned with the highest risk of cancelation and no-show behaviour. This characteristic is likely due to the uncertain nature of their trips.

In its large majority, they enjoy travelling solo, with some exceptions to the rule which might be pairs or couples. Their length of stay may vary a lot, from only a few days to 1 week and, according to the lead time, these are mainly spontaneous trips, as last-minute guests represent close to half of this group.

Furthermore, they are average spenders in terms of total revenue, but account for the highest overall revenue per person, per night. Their ages vary, peaking at middle-aged adults, but there is a high concentration of young tourists. These customers are the least demanding, showing a slight preference for special bed options.

Nearly 3/4 of overall corporate travellers are concentrated in this group, aligning with the solo trip assumption, regardless of having a majority of clients booking through agents and operators. This indicates that, even though most corporate customers are concentrated in this cluster, we should not forget about clients visiting with different objectives, thus the strategies produced should appeal to both.

Lastly, 1/3 of Portuguese clients are represented in the present segment, while most customers come from all around Europe.

## Marketing Strategies

**Cluster 0** – *Generations on the go*

For these clients, we propose a set of unique initiatives that enhance family bonding, create surprise elements, and offer parents much-needed rest while making the hotel stay unforgettable for everyone.

To create an immersive cultural experience, families can choose to schedule a surprise mystery itinerary upon booking. The agenda, which is only disclosed upon check-in, would include unique local experiences like a private cooking session with a Lisbon chef, a hands-on Portuguese tile-painting workshop, or a city treasure hunt. This unpredictability makes the stay exciting while ensuring families leave with a one-of-a-kind story to tell back home.

**Cluster 1** - *Holiday now, bills later*

Cluster 1 guests are big spenders with long stays but have high no-show rates. To maximize their value, we introduce a free extra night offer which is unlocked whenever clients spend a certain amount on upscale experiences like private tours, fine meals, spa treatments, or upgraded rooms. This tactic turns a cost into a high-return investment for the hotel by encouraging extravagance while prolonging their stay and making sure that by the time they utilise their complimentary night, they have already spent more than its value.

**Cluster 2** – *Lovebirds’ long-weekend retreat*

To captivate cluster 2 customers, the hotel would launch Private Pop-Up Experiences. These would include exclusive, private moments such as candlelit dinners, private Fado concerts, sunset boat rides, wine tastings, art workshops, and cooking sessions with local chefs. These experiences emphasize exclusivity, personalization, and cultural immersion, tailored to create unforgettable memories. We believe that these would be a great strategic addition making the hotel the go-to destination for couples.

**Cluster 3** – *The budget friends*

For cluster 3 guests, we would provide discounts for each extra person in the reservation for these guests, who are economical travelers who usually travel in small groups or in pairs. This strategy not only lowers the cost of their stay but also increases general occupancy and boosts income from group reservations. Additionally, as they are early planners, we would implement a reward system that offers free breakfasts or exclusive discounts on early reservations. This encourages long-term planning while securing bookings well in advance, helping to lock in early revenue. To ensure the profitability and sustainability of our strategic plan, there should be a limited number of rooms allocated to this strategy.

**Cluster 4** – *Frequent visits, solo style*

To captivate solo travellers and frequent guests we propose unconventional strategies designed to enhance loyalty and engagement while creating a sense of exclusivity.

Inspired by the hidden menus of trendy restaurants, the first initiative introduces a secret list of perks accessible only to (returning) solo guests. This menu, available upon request, could include complimentary surprise cocktails, an invitation to a solo-travel networking dinner, or even a whole plan to visit the city designed for these kinds of travellers, who may not have much time to plan. By making these benefits feel like an insider’s privilege, we differentiate our experience with that of competitors, and increase their connection with the hotel, transforming each stay into a discovery-driven experience.

Solo travellers often seek spontaneity, so we introduce an interactive gamification element to their stays. Guests can opt into a set of challenges, such as trying five different dishes from the hotel restaurant or visiting three partner locations in the city. These challenges would then unlock exclusive prizes like free spa visits, ideal for couples, personalized souvenirs, or room upgrades, if available.

These tactics, which combine exclusivity and gamification, transform typical lone stays into exciting and fulfilling experiences, ensuring continued loyalty and differentiation in the hotel industry.

# DEPLOYMENT AND MAINTENANCE PLANS

*‘If You Fail to Plan, You Are Planning to Fail’* - Benjamin Franklin.

This is especially true in IT and data-driven projects, where poor planning and communication breakdowns can significantly hinder success. The *Bull Survey (Spikes Cavell, 1998)* found that **57% of IT project failures** stem from communication failures, followed by **39% due to inadequate planning** and **35% due to poor quality control**. More recently, *Gartner (2019)* reported that **only 20% of analytic insights** effectively translate into business outcomes.

Recognizing these challenges, we have established a structured deployment plan to ensure clarity, alignment, and continuous improvement. By integrating a phased approach—covering research, validation, and long-term model updates—we aim to mitigate risks, enhance segmentation accuracy, and drive meaningful business impact.

## Proposed Next Phases

* **Next Milestone: Validation & Feature Enhancement**

Once approval is secured, the next phase will focus on refining the segmentation approach and enhancing its practical application for corporate clients. In collaboration with the **marketing and commercial teams**—**whose expertise in market trends and B2B dynamics is crucial**—the newly developed customer clusters will be analysed to ensure they accurately reflect real behaviours and market patterns.

To strengthen the segmentation model, additional analytical capabilities will be introduced if available. We will examine **seasonal trends**, specifically the *month during which clients stay*, to understand fluctuations in customer demand and analyse how different times of the year affect purchasing behaviour and price sensitivity. Furthermore, we will assess the **length of customer stays** as a key factor in determining engagement levels and long-term value. We will also incorporate **ratings analysis** to evaluate customer satisfaction and identify underlying trends in feedback, ensuring that marketing efforts are aligned with customer expectations.

In addition to these analytical improvements, the user interface will be further refined based on real-world feedback. Enhancements will target the optimization of design, usability, and information presentation, ensuring that the tool is both intuitive and effective for end users. By incorporating these new insights and refinements, the segmentation model will evolve into a more precise, data-driven framework that supports strategic decision-making.

* **Continuous Improvement & Model Updates**

To maintain relevance, the clustering models undergo regular updates, incorporating new data to reflect evolving market dynamics. Key variables are systematically stored to facilitate long-term enhancements. Additionally, the user interface is periodically refined, integrating industry best practices and assessing the feasibility of external business intelligence (BI) tool integration to maximize analytical efficiency.

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 4.*** Deployment Plan

## Internal Data Analysis Solution: A Cost-Effective Alternative

Our solution provides an interactive, user-friendly, and cost-efficient alternative to traditional BI tools like Power BI and Tableau, while also being more flexible than outsourced app development.

### Key Features of Our Solution

***Table 1.*** Internal data analysis solution

|  |  |
| --- | --- |
| **Feature** | **Description** |
| *Interactive Dashboards* | Ready-to-use visualizations tailored for the hotel industry. Users can interact with visualizations, create custom views, and share notes and insights. |
| *Predictive Clustering* | Users can input data and receive cluster predictions instantly while also seeing data quality warnings. |
| *Real-Time Notes & Feedback* | The marketing team can annotate insights directly within the app. The information is saved and can be used for brainstorming in future meetings. |
| *Customizable & Scalable* | Easily adaptable to new data requirements without additional licensing fees. |
| *No Technical Expertise Required* | Unlike tools like Power BI or Tableau, which often require specialized training, our solution is designed to be much more user-friendly. This allows non-technical teams to easily answer their own questions. They can also set up data alerts and generate recommendations without relying on technical expertise. Our goal is to empower everyone to fully utilize analytical capabilities and leverage important industry knowledge without the need for advanced training or technical skills. |
| *Download Options* | Users can download visualizations as images (.png) for presentations and export summary data per cluster in tabular format. |

### Cost Comparison[[4]](#footnote-5)

We analysed the costs of leading BI tools and the development of a custom internal app through outsourcing. Below is the cost breakdown **for a team of 10 users**:

***Table 2.*** BI Tools Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tool** | **Our Streamlit App** | **Power BI Pro** | **Tableau Viewer** | **Outsourced App Development** |
| *Interactive Dashboards* | **Yes** | Yes | Yes | Yes |
| *Predictive Clustering* | **Yes** | Yes | No | Depends on development |
| *Real-Time Notes & Feedback* | **Yes** | No | No | Depends on development |
| *Customizable & Scalable* | **Yes** | Limited | Limited | Yes |
| *Technical Expertise Required* | **No** | Yes | No | Depends on development |
| *Download Options* | **Yes** | Limited | Limited | Depends on development |
| *Pricing* | **One-time development + maintenance** | €9,40 + IVA user/  month | €35 + IVA  user/  month | High initial cost; varies based on complexity |
| *Annual Cost*  *(10 Users)* | **Lower long-term cost** | €1,128€ | €4,200€ | Simple app development price tag: €4,600 – €46,000[[5]](#footnote-6) |

**Why Choose Our Solution?**

For hotels looking for a powerful yet cost-efficient data analysis tool, our solution provides the best balance between functionality, usability, and affordability. Our platform is an adaptable and immediate solution for marketing teams to start working.

* **Cost Savings:** No recurring per-user fees, making it a budget-friendly option.
* **Faster Implementation:** Ready to use without the need for extensive setup or training.
* **Greater Flexibility**: Easily modifiable based on marketing team feedback.
* **Predictive Capabilities:** Built-in clustering analysis without additional software costs.
* **Collaboration:** Teams can leave notes and refine their strategies in real time.

You can check our Internal Data Analysis Proposal here: [Hotel H Interface](https://case-1-business-cases-with-data-science-25.streamlit.app/).

# CONCLUSIONS

## Considerations for model improvement

A key objective of this project was to address the inefficiencies in the previous segmentation approach, which was heavily influenced by the ***Distribution Channel*** variable. This bias led to customer clusters that were primarily defined by the distribution channels, overshadowing other important factors and undermining the granularity of the segmentation. During the development phase, we focused on mitigating this issue by excluding categorical variables, such as the distribution channel, from the algorithm. By doing so, we ensured that the resulting clusters were not dominated by any single category. This allowed us to uncover more accurate customer segments that reflect underlying behavioural patterns rather than simple categorizations based on distribution channels. As a result, the new segmentation approach is unbiased and provides a more insightful view of customer behaviour.

Additionally, suggestions for future improvements were identified during the deployment phase. One such suggestion is to incorporate seasonal trends, particularly by analysing the month of customer stays, length of customer stays, and ratings analysis to assess customer satisfaction and identify trends in feedback.

Finally, we recognize the importance of improving data quality and integrity, particularly when identifying individual customers. To address this, we recommend making ID submission mandatory during both online reservations and check-ins. In cases where this requirement is not met, we propose discussing potential solutions with the technical team. One option is to ensure that any missing or placeholder ID fields are automatically filled with a unique, randomized value before hashing. This will help prevent duplicate DocIDHashes caused by placeholders and improve the accuracy of identifying individual customers in the system.

These improvements will not only enhance the accuracy and reliability of the segmentation model but also help create more effective strategies for customer engagement and decision-making moving forward.

# REFERENCES

* **Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R.** (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* Retrieved from <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>
* **Hotelchamp.** (n.d.). *Top 5 reasons for booking engine abandonment and how to fix them.* Retrieved from <https://www.hotelchamp.com/blog/5-reasons-booking-engine-abandonment>
* ***Google.*** *(2024). Service unavailable. Retrieved from* [*https://trends.google.com/trends/explore?date=2022-01-01%202025-03-04&q=flights%20to%20portugal*](https://trends.google.com/trends/explore?date=2022-01-01%202025-03-04&q=flights%20to%20portugal)
* **Ruiz, P.** (2017). *Project success and new ventures' outcomes: How often do partners' potential benefits and losses really converge?* ResearchGate. Retrieved from <https://www.researchgate.net/publication/312040736_Project_success_and_new_ventures'_outcomes_How_often_do_partners'_potential_benefits_and_losses_really_converge>‌
* **Health Data Management.** (2019, January 18). *Gartner's top data and analytics predictions for 2019*. Retrieved from <https://www.healthdatamanagement.com/articles/gartners-top-data-and-analytics-predictions-for-2019>
* **Data Science PM.** (2024, November 16). *Why big data science & data analytics projects fail*. Retrieved from <https://www.datascience-pm.com/project-failures/>
* **Innowise.** (n.d.). *Custo de desenvolvimento de aplicações: Quanto custa desenvolver uma aplicação móvel em 2025?* Retrieved from <https://innowise.com/pt/blog/mobile-development-cost/>
* **Tableau.** (n.d.). *Buy Tableau – Pricing and plans.* Retrieved from <https://buy.tableau.com/en-de/>
* **Microsoft.** (n.d.). *Power BI pricing.* Retrieved from <https://www.microsoft.com/pt-pt/power-platform/products/power-bi/pricing?market=pt>
* **Dogtiev, A.** (2025). *App development cost (2025).* Business of Apps. Retrieved from <https://www.businessofapps.com/app-developers/research/app-development-cost/>
* **BinaryFolks.** (n.d.). *What is the cost of creating an app?* Retrieved from <https://www.binaryfolks.com/blog/how-much-does-an-app-cost>

# APPENDIX

## Glossary:

* **CRISP-DM (Cross Industry Standard Process for Data Mining)** – A structured methodology for data mining projects, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.
* **Clustering** – A machine learning technique used to group similar data points together based on shared characteristics, commonly used for customer segmentation.
* **K-means Clustering** – A partitioning clustering algorithm that assigns each data point to the nearest of *k* centroids based on Euclidean distance.
* **Self-Organizing Maps (SOM)** – A neural network-based clustering method that maps high-dimensional data into a lower-dimensional space while preserving topological relationships.
* **UMAP (Uniform Manifold Approximation and Projection)** – A dimensionality reduction technique used to visualize high-dimensional data while preserving its structure.
* **Silhouette Score** – A metric used to evaluate clustering quality by measuring how similar a data point is to its assigned cluster compared to other clusters.
* **Market Basket Analysis** – A data mining technique used to identify relationships between items in transactions, commonly applied in retail and customer behavior analysis.
* **R2 (Coefficient of Determination)** – A statistical measure that indicates how well a model explains the variance in the dependent variable. It ranges from **0 to 1**, where **1** means the model perfectly fits the data, and **0** means it does not explain any variance.

## Data Preparation Process (detailed)

**Table A 1.** Data Preparation Process

|  |  |
| --- | --- |
| **Step** | **Action** |
| ***Duplicates***  ***Treatment*** | Duplicate detection was performed using different approaches, each addressing a specific scenario |
| ***Handling Data Inconsistencies*** | Resolving inconsistencies found during the Exploratory Data Analysis (EDA) that could lead to misleading results in analytics and machine learning models |
| ***Missing Values*** | Missing values treatment using the K Nearest Neighbour (KNN) Imputation method. |
| ***Outliers*** | The chosen method ensured a balance between data retention and quality, maintaining dataset integrity while reducing noise and improving feature relevance. Extreme values were manually removed using predefined thresholds, resulting in only 0.04% data loss. Winsorization was applied to further reduce the impact of remaining extreme values without losing additional data. |
| ***Features almost no variability*** | These features needed further analysis since no variability does not contribute to distinguishes between clients. |
| ***Encoding*** | Transform categorical strings into numerical ones. |
| ***Scaling & Normalization*** | Algorithms like K-Means rely on Euclidean distances, making feature scaling essential. |

1. Unlike the industry-standard ADR, which divides total room revenue by rooms sold, our method is constrained by data availability and calculates ADR using total revenue per client divided by (rooms × nights), as we cannot separate the number of nights per booking. [↑](#footnote-ref-2)
2. From the variables: 'Age','OtherRevenue', 'PersonsNights', 'Total\_Revenue', 'RevenuePerPersonNight', 'AvgOccupancy', 'ADR', 'DistributionChannel\_Corporate','Origin\_Portugal','AverageLeadTime'. [↑](#footnote-ref-3)
3. From the variables 'Age', 'OtherRevenue', 'PersonsNights', 'Total\_Revenue', 'RevenuePerPersonNight', 'AvgOccupancy', 'ADR', 'AverageLeadTime'. [↑](#footnote-ref-4)
4. Note: These estimates are approximate and can vary based on specific project requirements and regional development costs. [↑](#footnote-ref-5)
5. Prices for app development can vary based on location and complexity. In Spain and Portugal, app development companies or freelancers typically charge between €25 and €80 per hour. The development time for simple apps ranges from 2 to 4 months, mid-complexity apps take about 4 to 6 months, while complex apps may require 9 months to over a year to complete. [↑](#footnote-ref-6)