NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel H

André, Lourenço, number: 20240743
Emir, Kamiloglu, number: 20240945
Manuel, Andrade, number: 20240571
Rute, Teixeira, number: 20240667
Victor, Silva, number: 20240663

Group T

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 20

1. EXECUTIVE SUMMARY

Given the simplicity of the current data processing and segmentation by the hotel, this analysis has been conducted to optimize customer segmentation and enhance the marketing campaign, as Meta advertisements have become a pillar of modern marketing strategies in the hospitality sector.

By developing segmentation analysis through proper data processing and algorithm implementation, we will strengthen the hotel's marketing campaign, as Meta's ad algorithm thrives on precise customer data. The main reasons for this are twofold: first, by processing data on various demographics, interests, and preferences, we can achieve accurate ad targeting. Secondly, we can further refine our funneling to better adapt to these clients.

An efficient marketing campaign is key to better capital allocation. More precise targeting will increase conversion rates and lower customer acquisition costs. Over time this may free up additional resources for research and development in other areas of the hotel, increasing shareholder value (ceteris paribus). Beyond improving campaign efficiency, the use of advanced machine learning techniques will also give the hotel a competitive edge over rivals still relying on outdated marketing strategies.

Through developments in data-driven insights, this approach will not only improve ad performance but also position the hotel for long-term success. The ability to continuously adapt and optimize targeting will ensure more effective spending and higher marketing ROI, keeping the hotel ahead in a rapidly evolving industry.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

This chapter is related to the Business Understanding phase of CRISP-DM.

2.1. BUSINESS UNDERSTANDING AND INDUSTRY CONTEXT

The business case in analysis belongs to Hotel H, a hotel that is a member of the independent chain C, located in Lisbon, Portugal. For the period this data was gathered, the total revenue is 38 million euros, of which 18% comes from in-stay guest expenses.

Hotel accommodation in Lisbon has been steadily growing for the past 5 years, with its lowest points only describing the seasonality characteristic to this industry: colder seasons show lower demand. The increasing popularity of Lisbon as a vacation destination validates the potential there is within this tourism sector for growth, translating into higher revenues for hotel business such as Hotel H.

No time-stamped data was disclosed, so we can't conduct a comparison analysis of Hotel H's performance against other hotels in the same sector. This limits our understanding of how the hotel measures up against industry benchmarks or historical trends, which could be crucial for identifying specific areas for improvement or growth opportunities.

Business Need and Required Outcome

Our client, Hotel H, identified a clear business need: a new client segmentation strategy that goes beyond sales origin by distribution channel to incorporate geographic, demographic or behavioral characteristics, such as age, country of origin, and number of stays. This requires a reviewed clustering solution, with more detailed profiling.

2.2. BUSINESS OBJECTIVES

By the end of this project, we aim to outline a clear and actionable segmentation strategy, based on deep analysis and insights collected from the data provided. Our segmentations strategy should be used by the marketing department to provide:

1) A tailored guest experience, leading to higher satisfaction, positive feedback, and a stronger reputation—key to attracting new customers, and

2) In-depth market research, providing clear insights into core clientele versus potential customers, their revenue contribution, and the associated acquisition costs.

2.3. BUSINESS SUCCESS CRITERIA

The success of this project's implementation will be measured by the quality and effectiveness of our cluster solution and profiling. Specifically, this can be translated into:

a. Meaningfully sized clusters - to ensure that we are segmenting clusters represent a significant portion of our clientele, either in count or revenue - with a **minimum threshold of 8%**
b. Distinctive cluster characteristics – Each cluster should have **at least one clearly identifiable strength or weakness**, ensuring a differentiated approach for targeted marketing strategies.
c. Clear and actionable customer profiles, to outline an effective marketing strategy that addresses each cluster's specificities - **at least five different profiles**, highlighting key differences between clusters
d. Practical number of clusters, for a feasible and cost-effective market implementation, the total number of clusters should be **limited to a maximum of six**.

2.4. SITUATION ASSESSMENT

Available Resources and Algorithms

This project is being conducted within a two-week timeframe by a team of five, leveraging computing resources with a maximum of 32GB RAM. The primary algorithms used for segmentation are DBSCAN and K-Means, along with aggregation operations that require additional computational time.

Internal Data

The dataset provided to us collects over 110,000 entries of clientele information, from 29 different features. We identified a total of 98,000 different clients, spanning 192 different nationalities, aged from 0 to 100 years.

External Data

To enhance our analysis, we incorporated industry performance data from **Portugal's National Statistical Institution (INE)**, specifically:

- **Number of stays and hotel revenue** for the **Lisbon region over the past five years**

Constraints and Limitations

During our analysis, we identified several limitations that impacted the depth of our findings:

- **Missing key business context – The dataset lacks crucial hotel-specific information such as total capacity and hotel star rating, making it difficult to assess luxury levels and customer purchasing power.**
- **Absence of time-oriented data – Without time stamped data, we cannot account for seasonality effects, a critical factor in the hotel industry.**
- **Lack of essential Key/Performance Indicators (KPIs) for segmentation, including:**
    - **Whether special requests were fulfilled**
    - **Guest satisfaction level (positive, negative, neutral, or no feedback)**
    - **Number of available rooms and average occupancy rate**
    - **Promotional offers used**

Additionally, no financial budget or time frame was provided to guide long-term marketing projections.

Assumptions

To bridge data gaps, mitigate uncertainty and proceed with a structured analysis, we made the following key assumptions:

- **A database entry is created when a customer interacts with the hotel platform, filling at least one personal info field, or when a customer stays at the hotel.**
- **The age group 0-15 and 16% of the "Disengaged" segment were excluded from the analysis, assuming these users were simulating reservations without purchasing power or intention of actual booking a reservation**
- **Checked-in guests under 16 were assumed to be staying with their caretakers.**
- **Guests who canceled or did not show up were assumed not to incur any fees.**
- **All customer revenue is captured under "LodgingRevenue" or "OtherRevenue" features**

2.5. DETERMINE DATA MINING GOALS

The established data mining goals include:

1. Thorough data exploration to assess the quality and structure of existing data
2. Comprehensive visual analysis and support that helps understand the different features and how they interact with each other
3. Feature engineering that extracts maximum information between features, without leading to redundant information
4. Organize data meaningfully to extract actionable insights that support the marketing department's operations.
5. Identify missing data and establish strategies for improved data collection in future initiatives.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

The data in question is derived from a Portuguese Hotel, we estimate that the data is 8 - 10 years old, given 81.5% of bookings were conducted via Travel Agent/Operator. The raw dataset dimensions are 111,733 x 29.

The purpose of this analysis is to improve the hotel's customer segmentation strategy, addressing the shallow assessment conducted by the previous marketing department. The previous analysis lacked detail in three key areas: geographic, demographic, and behavioral characteristics of the clientele.

By incorporating this three-fold approach, we believe the hotel can significantly optimize its marketing campaigns. A more in-depth analysis of customer attributes will help identify an audience with high check-in rates and greater revenue potential, ultimately improving the hotel's capital allocation in the marketing department.

3.2. DATA PREPARATION

In this section we will elaborate on the structure used for data preparation, a crucial point in the analysis as this section must be tackled in line with the business fundamentals and needs. Preparing such data on customers of the hotel, will allow for a robust analysis, which will be crucial for the marketing campaign. Given the relevance of Meta advertising in digital marketing, by having the data normalized, we will obtain a more robust analysis, which is crucial for Meta Ads as their algorithms thrive on precise targeting and patterns. The steps involved in this section involved handling missing values, duplicates, outliers and correlations.

3.2.1. Handling Duplicates

We started the data preparation by analysing duplicates. We initiated the analysis by checking standard duplicates, these are rows that have duplicates whilst excluding the columns ID. We verified 1169 duplicates that appear sequentially, which we hypothesize being entry errors**.** The third checkmade was understanding rows with repeated client, using the features Age, Nationality, and DocIDHash. There are 9366 entries with repeated DOCIDHash info, of which 2797 is the count of unique clients. We aggregated the rows with the repeated DocIDHash values into unique row, to aggregate customers. We then had 548 rows where we found duplicate DocIDHash, these values we believe represent no-show customers, as it is quite possible that the hotel used a specific DocIDHash for missing values. This information may be imported in the clustering section as might find customer characteristic patterns for those who end up not completing their bookings.

3.2.2. Incoherencies

Having concluded the section on duplicates we checked for data incoherencies, where we found several columns that might need further attention. We started by the feature Age, where we found inconsistent min and max values. We found a negative age and an excessively high max value. For the maximum age we set 100 using IQR fundamentals and logic.

The following feature was AvergaeLeadTime, here we removed all the negative values, as it is not possible to have a negative average number of days between customer booking completion and arrival. We deleted 12 clients where this was the case.

The columns mentioned above were the sole features with issues regarding incoherencies, the remaining were aligned with typical hotel booking data.

3.2.3. Feature Engineering

In this section we elaborate on the features created which we believe may enhance customer understanding and possibly pick up relevant trends for the hotel marketing campaign.

3.2.3.1. First Engineering – TotalRevenue

This simple feature was created by using the sum of LodgingRevenue & OtherRevenue. This column reveals all the revenue attributed from a customer.

$$Total\ Revenue = Lodging\ Revenue + Other\ Revenue$$

3.2.3.2. Second Engineering – AveragerevenueBooking

Here we created the feature which captures the average revenue a customer brings per booking.

$$AverageRevenueperBooking = \frac{(Bookings\ CheckedIn\ +\ Bookings\ no\ showed)}{Total\ Revenue}$$

3.2.3.3. Third Engineeering – TotalRequest

This column captures the sum of requests a customer has made, this column will reveal customer tendencies.

$$TotalRequest = \sum(All\ Special\ Requests)$$

3.2.3.4. Fourth Engineering – CheckinRate

To measure customer booking reliability, we introduce the Check-in Rate feature. This metric evaluates the proportion of bookings that successfully result in a check-in, helping to analyze customer behavior and predict booking fulfillment trends.

$$CheckinRate = \frac{(BookinksCheckedIn)}{(BookingsNoShowed + BookingsCanceled + Bookings\ ChecckedIn)}$$

3.2.3.5. Fifth Engineering – FallBackRate

To assess customer booking uncertainty, we introduce the FallBack Rate feature. This metric represents the proportion of bookings that did not result in a check-in. indicates potential revenue loss due to cancellations and no-shows.

$$FallBackRate = 1\ - CheckinRate$$

3.2.3.6. Sixth Engineering – Frequency(Days)

To analyze customer retention and booking behavior, we introduce the Visit Frequency (Days) feature. This metric estimates how frequently a customer makes a booking based on the time elapsed since their first recorded booking.

$$Frequency(Days) = \frac{BookingsCheckedIn}{DaysSinceCreation}$$

3.2.3.7. Seventh Engineering – Frequency(Days)

The Average Person Per Room metric calculates the average number of guests per room during a customer's stay. This helps in understanding room occupancy trends, which can optimize hotel pricing, resource allocation, and marketing strategies.

$$AveragePersonPerRoom = \frac{RoomNights}{PersonNights}$$

3.2.3.8. Eight Engineering – Geographic Segements

This feature groups customers into specific country-based or continent-based segments:

- If the customer is from one of the top 5 most frequent countries (Portugal, Spain, Germany, France, United Kingdom), they are assigned their respective country.
- All other customers are categorized by continent.

$$\text{Tourism Segment 1} = \begin{cases} \text{CountryName}, & \text{if CountryName is in Top 5} \\ \text{Continent}, & \text{otherwise} \end{cases}$$

3.2.3.9. Ninth Engineering – Booking Window

We created BookingWindow to understand the planning behaviour of our customers.

Lead Time (Days)	Category	Interpretation
0-7	Last-Minute (0-7)	Urgent bookings, likely spontaneous travellers.
8-30	Business Travels (8-30)	Guests booking short-term stays, often for work.
31-90	Gateways (31-90)	Leisure travellers planning weekend or short trips.
91-180	Planners (91-180)	Customers planning for vocations.
180+	Specific Travels (+180)	Long-term planners, usually for events or high-season trips.

3.2.4. Missing Values

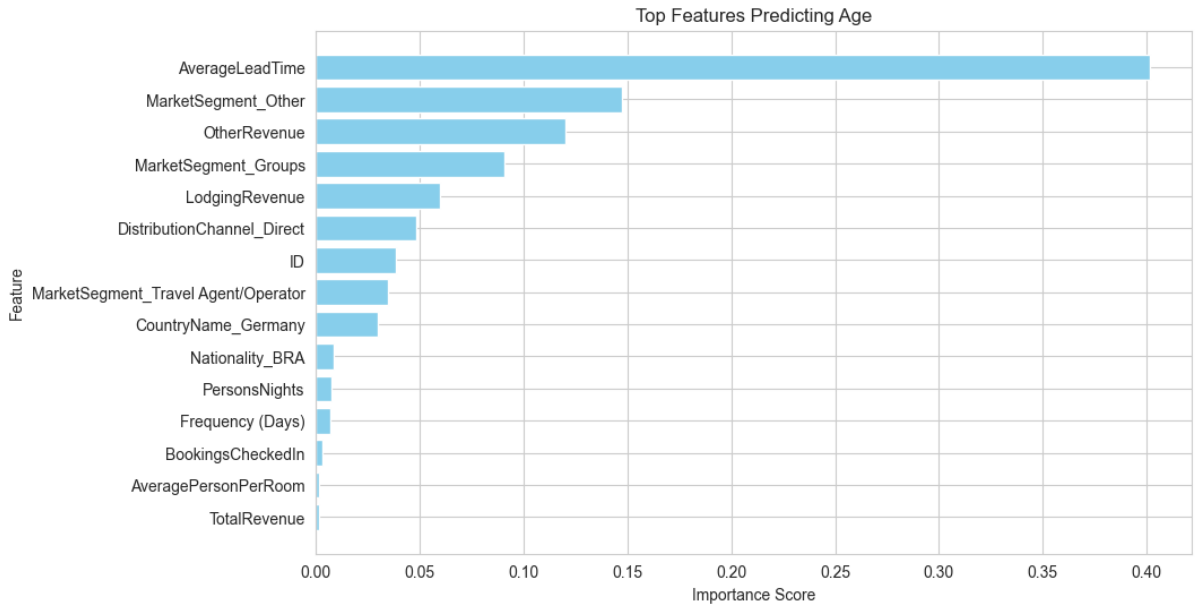
The dataset contained missing values in the "Age" column. Initially, we suspected missing values in "DocIDHash," but after verification, we confirmed that this column contains no missing values and does not require further processing, as it was already addressed in the Duplicates section.

For the "Age" column, which represents the age of a given customer, we identified 227 missing values. To avoid dropping these rows, we decided to use Decision Tree regression to predict the missing values.

To determine the most relevant features for prediction, we analyzed the relationship between Age and both numerical and categorical features:

- Among numerical features, the highest correlations were observed with AverageLeadTime and DaysSinceCreation.
- Among categorical features, we conducted an ANOVA test and found that DistributionChannel was a strong predictor of Age, with a statistically significant p-value at $\alpha = 0.01$.

We then trained a Decision Tree model using these key features and successfully predicted the missing Age values, fully eliminating missing values from this section. As a result, the dataset is now completely free of missing values. The most influential predictor for Age was AverageLeadTime.



(Diagram 3.2.4. – Top features for Age imputation)

3.2.5. Outliers

Outliers in the hotel data set are another issue we must face to have robust results. Identifying and handling these values is essential to ensure the accuracy and reliability of our analysis, as outliers can distort statistical measures and impact model performance.

The methodology used in this section is the Interquartile Range(IQR). We apply this method to statistically detect outliers, entailing a scientific based solution. The results achieved revealed that revenue related variables presented the highest count of outliers. We found 4803 and 4681 outliers for LodgingRevenue and OtherRevenue, respectively. This reveals how certain customers have considerably higher hotel spending habits relatively to the common customer. Additionally, booking related variables, such as BookingCheckedIn also reveal high counts of outliers, having 3318 outliers, reinforcing the value of our variable BookingWindow which reveals customer booking behaviour.

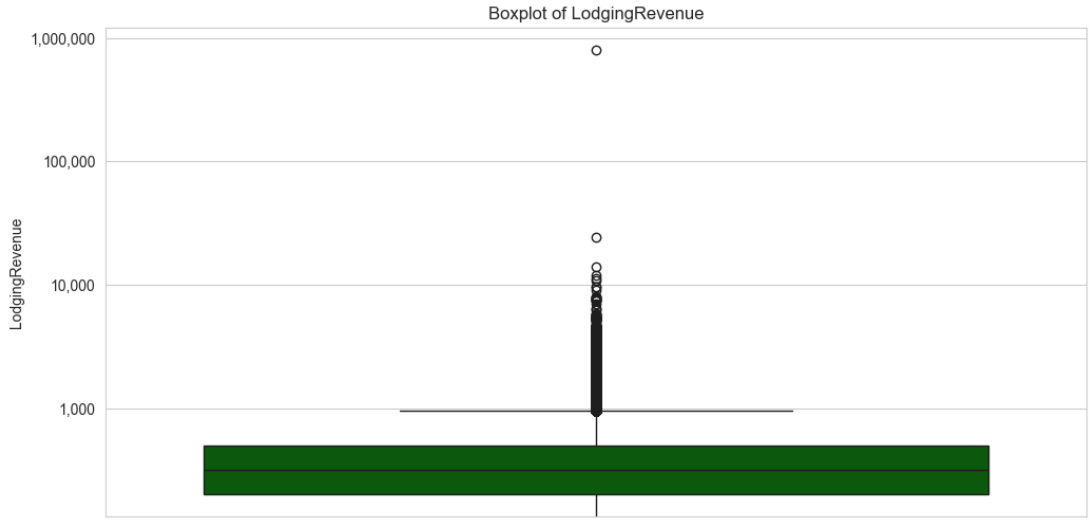
Finally, the Cancellation and no-shows revealed the lowest amont of outliers. Revealing the cancelation and no-show behaviour values don´t tend to have major deviations of the median.

3.2.5.1. LodgingRevenue and OtherRevenue

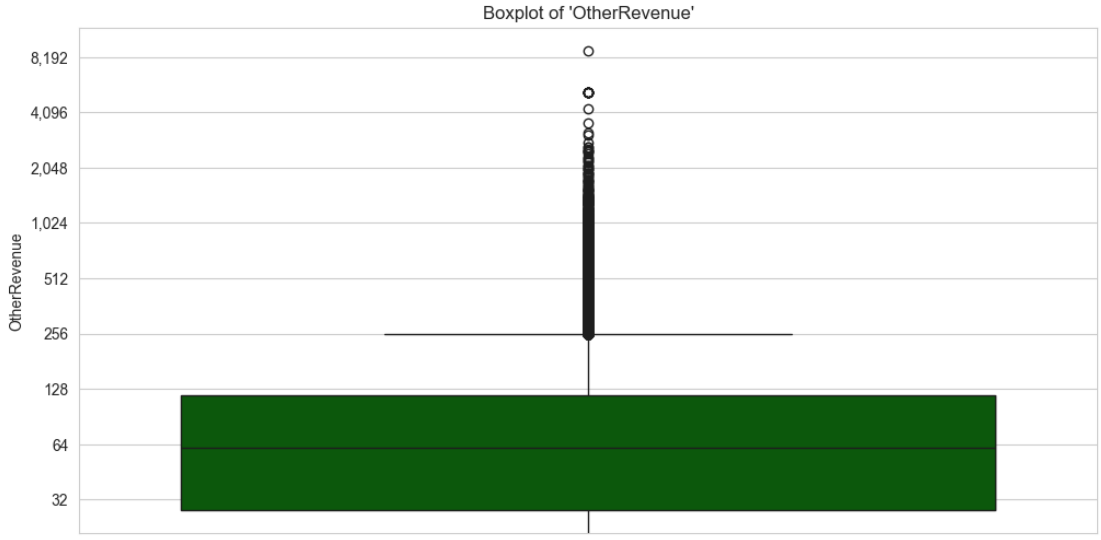
Conventional outlier detection methods like IQR flag a large number of points as outliers (4,399 in this case), but they can be overly conservative since they do not consider data distribution and density. Many of these flagged points may still hold meaningful information rather than being true anomalies. Instead, we propose defining outliers based on where data density significantly declines, ensuring that only truly extreme deviations are removed while preserving relevant observations.

Since an alternative method is crucial for our analysis, we chose to integrate the DBSCAN Outlier Detection method, given its capacity for overcoming outliers based on density in a multivariate space. DBSCAN evaluates data points near a local neighborhood, which it uses to identify whether they belong to a dense region. Consequently, this allows for the handling of nonlinear data patterns, which traditional outlier handling methodologies are lacking. This makes the outlier detection analysis robust, as we are using DBSCAN and IQR, which have their own strengths and shortcomings.

To apply this method, feature scaling was needed as it is a crucial preprocessing step when applying density-based clustering algorithms such as DBSCAN. Since it relies on Euclidean distances to identify clusters and outliers, differences in the scale of numerical features can significantly impact the results. If features have vastly different ranges, the clustering process may become biased, as high-magnitude features can dominate the distance calculations. Hence, we scaled the data using MinMaxScaler, which ensures that all numerical variables contribute equally to distance measurements by transforming them into a fixed range between 0 and 1.



(Diagram 3.2.5.1.1. – LodingRevenue Boxplot)



(Diagram 3.2.5.1.2. – OtherRevenue Boxplot)

3.2.6. Featurue Selection

In the following section, we performed feature selection using correlation analysis to remove low-value features or those that share redundant information. The goal of this step was to reduce dimensionality while preserving the most informative variables for further analysis, particularly for clustering and outlier detection. By eliminating highly correlated variables, we ensure that our model is not biased by duplicate information, leading to more efficient and interpretable results.

To achieve this, we used the Spearman correlation matrix, which is particularly useful because it can detect both linear and non-monotonic relationships. Unlike Pearson correlation, which is limited to linear dependencies, Spearman's rank correlation provides a more robust measure when dealing with

data that may not follow a strict linear pattern. This makes it ideal for datasets containing skewed distributions, ordinal relationships, or complex dependencies, as seen in our dataset.

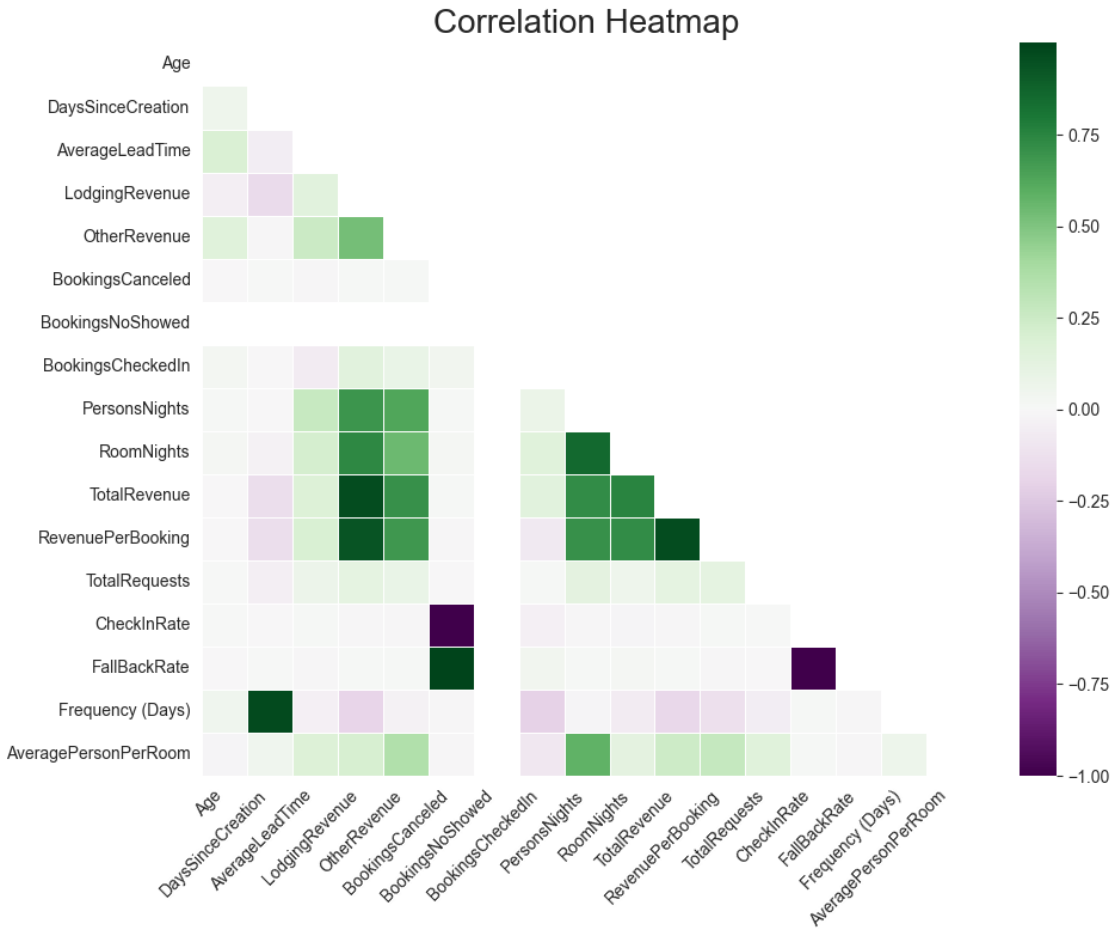
Using this method, we set a correlation threshold of 0.70, meaning that if two variables had a correlation coefficient greater than 0.70, we considered them too similar and removed one of them. This threshold helps strike a balance between removing redundant features while retaining sufficient information for accurate clustering. As a result of this analysis, we identified and removed the following features:

- Days Since Creation – Highly correlated with Frequency (Days), making it redundant.
- Persons Nights & Room Nights – Both were strongly correlated with Total Revenue, meaning they contributed overlapping information.
- Check-in Rate & Fallback Rate – These variables were perfectly negatively correlated (-1.00), meaning one could be derived from the other.

Days Since Creation – Highly correlated with Frequency (Days), making it redundant.

Persons Nights & Room Nights – Both were strongly correlated with Total Revenue, meaning they contributed overlapping information.

Check-in Rate & Fallback Rate – These variables were perfectly negatively correlated (-1.00), meaning one could be derived from the other.



(Diagram 3.2.6 – Correlation Heatmap)

3.3. MODELING

The objective of our clustering analysis is to understand how the customer base interacts with the hotel. We analyzed customers based on their segments, which derived from their booking and spending behaviors. By having a solid overview of these characteristics, we can optimize our targeting on Meta ads for better conversion rates. We tested various clustering algorithms and considered carefully the strengths and shortcomings of the Hierarchical, DBSCAN and K-means, clustering algorithms. In this section of the report, we start considering the various algorithms, analyzing the strengths and shortcomings of each, and elaborating why we chose the K-Means clustering. Once this is elaborated, we specify the various clusters used to match the analysis with business needs of the hotel, where features are also considered based on the business needs and their significance.

Additionally, we created a 5th cluster, Cluster 4, which represents the disengaged customers, these were customers that had never made a booking or contributed to any of the hotel revenue thus far. These customers were set to a separate dataset and treated in isolation. We will use the characteristics to understand the demographics for the no revenue generating customers for further marketing optimization, and for potential service improvements for customer acquisition.

The next section of the modelling represents the choice for the optimal number of clusters, here we used two forms in tandem, the silhouette score and the elbow method. We concluded the section by mentioning the limitations of our modeling.

3.3.1. Clustering Algorithm Selection

When choosing the clustering algorithm for this analysis, we had to consider the strengths and shortcomings of each method. By analyzing and testing DBSCAN, Hierarchical, and K-Means clustering algorithms, we carefully selected K-Means, as we believe it is the most appropriate for the project in question. Below, we will analyze each of these three.

3.3.1.1. K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm widely used for customer segmentation. The goal of this algorithm is to group similar customers, meaning customers that are statistically similar based on the values present in their feature set. This will be essential for optimizing the hotel's advertising campaign.

The main strength of this clustering method lies in its interpretability, which aligns with the hotel's needs, as understanding the customer base is crucial for the practicality of the project. Using

clustering techniques that reduce dimensionality (e.g., PCA) would make it difficult for the hotel to apply the findings of the analysis effectively. Additionally, since the hotel's marketing department does not have technical machine learning expertise, an easy-to-interpret and presentable analysis is optimal, ensuring minimal jargon.

The limitations of K-Means are related to the need to predefine the number of clusters and their sensitivity to outliers. However, by using the Silhouette Score and Elbow Method, we were able to identify the optimal number of clusters, effectively overcoming this limitation. Additionally, we handled outliers in the data processing stage using the Interquartile Range (IQR) and DBSCAN density-based filtering.

3.3.1.2. Hierarchical Clustering

Hierarchical clustering was the first algorithm tested in the analysis, but it led to memory issues from the start. This occurred due to the computational expense of hierarchical clustering on larger datasets, whereas K-Means can efficiently handle large volumes of data.

3.3.1.3. – DBSCAN

The final method we tested was DBSCAN, but this algorithm struggled to create clearly separated clusters. Additionally, more parameters need to be fine-tuned compared to K-Means, making DBSCAN less practical and harder to implement for this particular analysis.

3.3.2. Choosing Optimal Number of Cluster (k)

As mentioned above, one of the requirements for K-Means is the need to predefine the number of clusters (k) beforehand. To find the optimal k, we used two methods, the Silhouette Score and the Elbow Method. Below we elaborate on these techniques.

3.3.2.1. Silhouette Score

This method measures how well clusters are formed based on two main pillars: Cohesion and Separation.

Cohesion refers to how well the points within a cluster are grouped together. A high Cohesion value indicates that points within a cluster are close to each other, forming a well-defined group.

Separation refers to how far apart clusters are from each other. Higher separation means clusters are more distinct and independent.

The optimal value of k is the one that maximizes these two metrics, ensuring that points within each cluster are close together while clusters themselves remain well-separated. This allows for the most interpretable clustering results.

3.3.2.2. Elbow Method

The foundation of this method is Inertia, which measures how compact clusters are.

The Elbow Method finds the k value where the marginal improvement in compactness becomes negligible, this is the "elbow point", where adding more clusters leaves to a marginal change of 0, since it no longer significantly improves the structure.

Using both the Elbow Method and the Silhouette Score provides a more reliable way to determine the optimal number of clusters in K-Means clustering. The diversity of these two methods allows us to consider two different perspectives when making the decision. This combined approach essentially mitigates the risk of choosing too many or too few clusters, ensuring the segmentation is both accurate and actionable.

3.3 Limitation of the Modelling

The main drawback we faced in this analysis was that we applied K-Means related to the sensitivity of the algorithm to both outliers and random initialization, which impacted the cluster outputs.

Given that K-Means relies on centroids that are randomly initialized at the initial stage of implementation and recalculated through iterations, small changes in outliers led to significant differences in the final cluster outputs. As mentioned in the preprocessing section, we used IQR and DBSCAN, and subtle changes to the aggressiveness of the outlier implementation significantly affected the clusters obtained.

Additionally, we found that the algorithm was sensitive to different random states. This sensitivity in K-Means was the main drawback of the modeling process, however, the benefits in interpretability and computational efficiency outweighed this limitation.

3.4. EVALUATION

The evaluation of our clustering model is based on cluster quality and business relevance. We rely on both quantitative metrics, Silhouette Score and Elbow Method and qualitatively through aligning the results with the business needs of the hotel.

The analysis initially suggested that k=5 could be the optimal number of clusters, we chose k=4, considering a marketing and strategic perspective. Using four clusters allows for a clearer segmentation that aligns better with our target audience, making it easier to design executable strategies. Additionally higher n could lead to over-segmentation, making customer groups harder to define and market too effectively. By keeping the segmentation broad, the hotel can optimize its campaigns without unnecessary complexity.

To validate the clustering quality, we analyzed the Silhouette Score and the Elbow Method. The Silhouette Score peaked at k=2 and k=3, indicating that the clusters were well-separated at lower values of k. However, since our goal was to achieve a balance between differentiation and business interpretability, we selected k=4. This point represents the ideal point as additional clusters would not significantly improve compactness, as the marginal benefit flattens. This decision ensures that the clusters are executable and practical for the marketing department.

After validating the statistical performance, we evaluated whether the clusters provided meaningful customer segmentation. The analysis identified four distinct customer groups: frequent high-spending guests, long-term planners, last-minute budget travelers, and occasional mid-spending customers. Each of these groups represents a unique consumer behavior, allowing the hotel to implement tailored marketing strategies. As an example, high-spending guests can be targeted with loyalty programs, while last-minute budget travelers may respond better to flash sales and last-minute discounts. These insights make the segmentation not only statistically sound but also actionable for business decisions.

In conclusion, our evaluation confirms that k=4 provided the most effective balance between statistical validity and business usability. The chosen segmentation allows for clear customer profiling, strategic marketing execution, and data-driven decision-making. While some sensitivity to initialization and outlier handling remains, the methodology applied ensures consistent and interpretable clusters that are actionable for optimizing hotel operations and marketing strategies.

4. RESULTS EVALUATION

In this section, we shall develop the results of our cluster and profiling analysis to understand the effectiveness of our model in identifying customer patterns and improving business understanding. The results evaluation shall reflect whether we have detected executable actions that can optimize the hotel's marketing campaign.

The evaluation criteria used will be based on several key performance indicators, ranging from cluster differentiation, customer understanding, and business execution/marketing actionability. Below we will study the Clusters including profiling in the analysis

The first cluster study was Behavior-Based Segmentation, as the features studied were AverageLeadTime, Frequency, OtherRevenue, LodgingRevenue, and RevenuePerBooking. Given the optimal K analyzed in the evaluation section of modeling, we obtained four clusters: Cluster 0, 1, 2, and 3, each representing a different customer type.

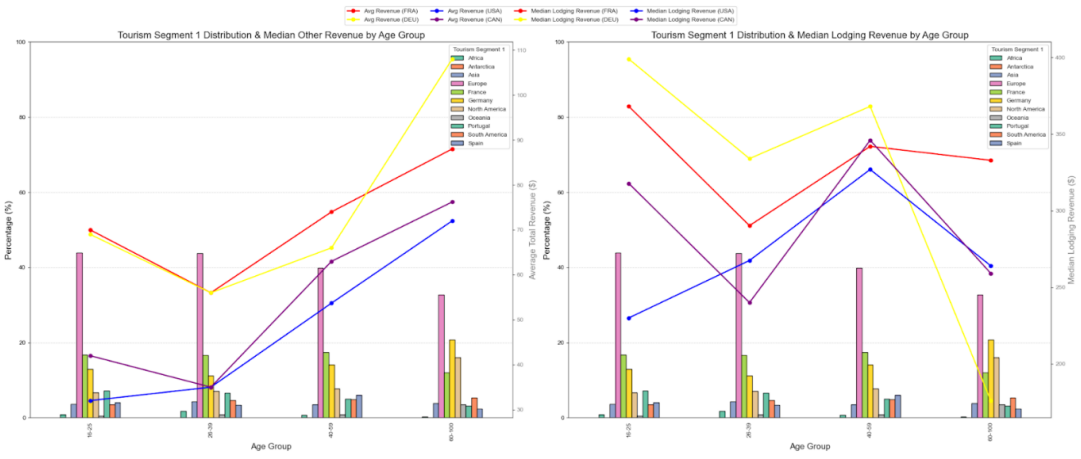
We named Cluster 0 as customers with considerably high spending and frequency. The highest revenue per booking in this segment was €597, with a lodging revenue of €510. They booked with moderate time in advance, given an AverageLeadTime of 191 days, and were frequent visitors, with a Frequency of 286 days. Naturally, they also presented high spending in OtherRevenue, with the highest value of €116. The demographics revealed here tended to be mid-range customers, between ages 30 and 50, who planned their journey.

The second cluster analyzed was Cluster 1, which characterized infrequent visitors who were budget-conscious and made last-minute bookings. This was evident from their large Frequency value (1041 days) and lowest revenue per booking (€440) and lodging revenue (€358). Their OtherRevenue was the lowest of the four clusters, reinforcing their budget-conscious nature. Their last-minute booking pattern was also revealed by their shortest lead time of 46 days. Dominated by younger customers (ranging between 20 and 35 years), mostly business travelers.

Next, we analyzed Cluster 2, which revealed moderate spending, frequent visitors, and last-minute bookings. These customers visited more frequently (Frequency: 381 days) and had mid-range revenue per booking (€481) and LodgingRevenue (€440). Their OtherRevenue was low (€87), indicating that they spent less beyond lodging. The demographics here also leaned towards business travelers, also between the ages of 20 and 35.

The final cluster of the main dataset, Cluster 3, represented high-planning, low-frequency visitors who showed high expense in other costs. They visited infrequently (Frequency: 927 days), and their AverageLeadTime of 214 days indicated that they booked well in advance. Their revenue per booking (€498) and lodging revenue (€389) were moderate. However, their OtherRevenue (€109) revealed a tendency to spend significantly on non-lodging services. The demographics here revealed middle aged individuals again, which showed preferences for quality services, and were specific travelers. By specificity here is revealed by the large AverageLeadTime, reflecting a booking for a specific event in advance.

The final cluster of the analysis was cluster 4, where we analyzed the disengaged customers, given these customers only revealed values consistent ages and present in the other clusters. The notable difference in this section related to a stronger percentage of U.S. and Canadian individuals when looking into ages above 60 (as seen by the light brown columns in the diagram below)



(Diagram 4.1 - Distribution & Median of OtherRevenue & LodgingRevenue by Age Group)

(Diagram 4.2 – Heatmap of all clusters except for the disengaged cluster)

5. DEPLOYMENT AND MAINTENANCE PLANS

To ensure the model's applicability and reusability, efficiency and scalability must be prioritized. The code should be Pythonic and modular, allowing seamless updates as the hotel's operations expand.

We recommend updating the model either monthly or every 10,000 new bookings. While a time-based approach ensures regular updates, tracking new check-ins offers a more dynamic adjustment based on actual booking trends.

As the hotel's data collection evolves, new features may be introduced. The code should be structured with functions to accommodate these changes effortlessly, ensuring adaptability and long-term usability.

6. CONCLUSIONS

This segmentation breakdown makes it obvious—understanding customer behavior isn't just about categorizing people, it's about using that data to make better business decisions that drive revenue. Each cluster represents a unique opportunity, and the hotel needs to stop treating all customers the same. High-spending frequent guests (Cluster 0) should be locked in with VIP loyalty perks, last-minute budget travelers (Clusters 1 & 2) need dynamic pricing that pushes them to convert and disengaged but high-potential customers (Cluster 4) shouldn't be chased with room discounts when what they spend on is everything else, spa, dining, experiences.

The real play here isn't just filling rooms it's maximizing the lifetime value of every guest. Instead of focusing purely on occupancy rates, the hotel needs to start thinking about the bigger picture: how much total revenue can be extracted from each customer type? Clusters 3 & 4 show that lodging isn't always the main driver, and the real money is in upselling premium experiences. This entails shifting the scope from simplistic occupancy strategies to an exhaustive optimized model that merges targeting, pricing intelligence, and personalized engagement.

Automate this, refine it with real-time data, and start executing smarter strategies that maximize revenue per guest not just per room while delivering a boutique-like service that reflects each guest's unique personality. The more dialed-in this segmentation gets, the stronger the long-term growth potential. This isn't just about knowing who the customers are, but about making sure they boost revenue, stay loyal, and engage with what's being offered.