

Master Degree Program in

**Data Science and Advanced Analytics**

**Business Cases with Data Science**

Case <1>: <Hotel Customer Segmentation>

<Engjulla>, <Hasani>, number: <20221404>  
<Danylo>, <Ivanchyshyn>, number: <20221692>

<Group B>

**NOVA Information Management School**

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

Universidade Nova de Lisboa

<March>, <2023>

**INDEX**

[1. EXECUTIVE SUMMARY 2](#_heading=h.gjdgxs)

[2. BUSINESS NEEDS AND REQUIRED OUTCOME 2](#_heading=h.30j0zll)

[2.1. Industry Background 2](#_heading=h.1fob9te)

[2.1.1.1. Hotel Background 3](#_heading=h.sptdzlfubto7)

[2.2. Business Objectives 3](#_heading=h.3dy6vkm)

[2.3. Business Success criteria 3](#_heading=h.1t3h5sf)

[2.4. Situation assessment 4](#_heading=h.4d34og8)

[2.5. Determine Data Mining goals 5](#_heading=h.2s8eyo1)

[3. METHODOLOGY 5](#_heading=h.3rdcrjn)

[3.1. Data understanding 6](#_heading=h.26in1rg)

[3.2. Data preparation 7](#_heading=h.lnxbz9)

[3.3. Modeling 9](#_heading=h.35nkun2)

[4. RESULTS EVALUATION 9](#_heading=h.44sinio)

[5. DEPLOYMENT AND MAINTENANCE PLANS 12](#_heading=h.2jxsxqh)

[5.1 Personnel 12](#_heading=h.fwweovsqtwa9)

[5.2 Monitoring 13](#_heading=h.rer9ozbyu6)

[6. CONCLUSIONS 13](#_heading=h.z337ya)

[6.1. Considerations for model improvement 14](#_heading=h.3j2qqm3)

[7. APPENDIX 15](#_heading=h.2xcytpi)

# EXECUTIVE SUMMARY

Hotel H needs more efficient client segmentation to improve its overall customer experience and increase revenue. Hotel H has supplied us with vital client data, including demographic information, booking trends, and preferences. As advisors, we recognize the significance of consumer segmentation in the hospitality industry. We discovered the various client categories within the hotel's customer base by employing a tailored clustering analysis which includes ….. This will help Hotel H establish more successful marketing tactics and segment-specific offerings, which we will explore in further detail below.

Clustering will give significant insights into the behavior and preferences of Hotel H's clients, allowing the hotel to make data-driven decisions that increase its market competitiveness. As consultants, we are committed to assisting Hotel H in achieving its business goals by delivering actionable insights and suggestions derived from our study.

Our analysis uncovered six client segments with distinct traits and behaviors. By gaining a knowledge of these segments, Hotel H can better adjust its sales strategy, marketing, and service offerings to its clients' requirements and preferences.

# BUSINESS NEEDS AND REQUIRED OUTCOME

During the Business Understanding phase of the CRISP-DM methodology, our team of consultants endeavored to get a thorough grasp of Hotel H's business issue. We studied the hotel's customer data to uncover patterns and trends that may aid in the development of effective client segmentation strategies. After comprehensive analysis of the data, we discovered the need for a more effective customer segmentation in order to better focus marketing activities and increase revenue. Our team advocated employing cluster analysis to categorize clients based on shared features, hence enabling more targeted and personalized marketing activities. By following this strategy, Hotel H will be in a better position to comprehend and adapt to the individual interests and behaviors of its broad client base, resulting in greater customer loyalty and profitability.

## Industry Background

The hotel business in Portugal continued to expand in 2018, as both the number of tourists and hotel occupancy rates rose. The National Institute of Statistics reports that Portugal welcomed 12,8 million international tourists in 2018, a 0.4% rise over the previous year. With 4,4 million international tourists, Lisbon, where Hotel H is located, was the most visited region in the world. The boost in tourism resulted in an increase in hotel occupancy rates, which reached an average of 71% nationwide. Yet, as hotel competition has increased and new distribution channels, such as online travel agencies and meta-search engines, have emerged, it has become increasingly vital for hotels to segment their client base in order to remain competitive and maximize income. According to the Portuguese Hospitality Atlas 2019 report (Deloitte Portugal, 2019), the tourism industry in Portugal experienced significant growth in 2018. “Cluster analysis serves to empirically answer the question of which hotel markets are complements and which hotel markets are substitutes in the context of a diversified portfolio”. Guest segmentation is the process of defining the market into identifiable segments. In other words, it is a process of grouping guests and narrowing them down.

### Hotel Background

The Lisbon, Portugal-based Hotel H, a part of the independent hotel chain C, has been applying a typical market segmentation based only on the customer's country of origin. But, the new marketing manager, A, found that this strategy is inadequate since it does not take into consideration other essential client criteria, such as their demographics, behavior, and preferences. A acknowledges that accurate customer segmentation is necessary for developing efficient marketing strategies, attracting new consumers, and maintaining existing ones. Like with the majority of hotels, Hotel H utilizes a variety of distribution channels, including travel agents, online travel agencies (OTA), brand websites, and meta-search engines. To develop effective marketing strategies, the hotel must take into account the various channels and consumer segments as marketing objectives. This project is to employ cluster analysis to identify client categories for Hotel H that will aid in the development of a more successful marketing plan. Real names and data are referenced, but anonymised for privacy considerations.

## Business Objectives

Hotel H in Lisbon, Portugal has a clear business necessity for more efficient client segmentation in order to better concentrate its marketing efforts and improve revenue. With the tourist business in Portugal enjoying tremendous expansion, competition among hotels has intensified, making it imperative for Hotel H to maintain its competitive edge and optimize its revenues. A, the hotel's new marketing manager, understands that proper customer segmentation is essential for generating effective marketing strategies, acquiring new customers, and retaining existing ones. Using cluster analysis, Hotel H will be able to identify unique client categories based on similar characteristics, so enabling more focused and customized marketing efforts. This will result in improved client loyalty, profitability, and a competitive advantage on the highly competitive Lisbon market. In addition, the hotel will be able to adapt its sales strategy, marketing, and service offerings to be more aligned with the needs and preferences of each group, resulting in a more efficient use of resources and ultimately a profit maximization.

## Business Success criteria

The major objective of the project is to raise the income of Hotel H by enhancing client segmentation and implementing more effective marketing techniques. Thus, a rise in income would be an unmistakable sign of the project's success.[[1]](#footnote-1)

Client happiness is an additional crucial indicator of success. If the project improves consumer experiences, it will likely lead to greater customer loyalty, favorable word of mouth, and repeat business.

In a competitive market like Lisbon, market share expansion is vital for long-term success. This project will help Hotel H expand its market share and achieve a competitive edge by attracting new clients and maintaining existing ones.

Return on investment (ROI)[[2]](#footnote-2): The undertaking will need a substantial commitment of time and resources. To establish if the project was worthwhile, it is necessary to compute the return on investment. The success of a project will be shown by a favorable return on investment.

Important to the success of any project is employee involvement. If staff are motivated and devoted to the project, it is more probable that they will effectively execute the required adjustments. Assessing employee engagement and satisfaction may be a valuable method for determining the project's performance.

The effectiveness of the project may also be determined by analyzing its effects on the hotel's operations. If the project results in more effective operations, cost savings and higher profitability are possible outcomes.

Ultimately, the project's performance may be determined using a mix of these factors, which represent both financial and non-financial metrics of success.

## Situation assessment

As a team of highly skilled and experienced data scientists, we have the expertise, data, computing resources, and software required to successfully complete the project. Our team is comprised of five experts in data modeling and analysis, machine learning, and statistical analysis. We have access to several datasets from reputable sources, as well as the most up-to-date analytical tools and software, enabling us to generate accurate and informative results.

While we are confident in our abilities, we are cognizant of the risks and uncertainties inherent to every project, particularly in the realm of data analysis. We have discovered potential threats include data security breaches, data quality concerns, and software or hardware malfunctions. To combat these risks, we have established stringent security standards and contingency plans to assure the project's continuation in the event of unanticipated situations.

We recognize the necessity of producing high-quality solutions within the allotted budget in terms of expenses and benefits. We have accounted for the costs related with data collecting, software licensing, and labor. By using our experience and resources, we aim to provide clients with solutions that are both cost-effective and deliver substantial benefits.

## Determine Data Mining goals

Develop a consumer segmentation model: The primary goal of the project is to improve client segmentation in order to improve the efficacy of marketing approaches. The team will employ cluster analysis to discover distinct client segments based on comparable traits. This will allow Hotel H to tailor its marketing activities, resulting in increased client loyalty, profitability, and a competitive edge.

Measure performance using financial and non-financial criteria: The project's success will be assessed by a combination of financial and non-financial indicators, such as increased income, customer happiness, market share expansion, return on investment, staff engagement, and operational efficiency.

Ensure data security and quality: The team will develop strong security standards and contingency plans to address possible risks such as data security breaches, data quality problems, and software or hardware failures.

Maximize resource use: The team attempts to develop cost-effective solutions that yield significant advantages by accounting for the expenses of data collecting, software licensing, and manpower. By creating a customer segmentation model, Hotel H will be able to better match its sales strategy, marketing, and service offerings with the requirements and preferences of each client group, resulting in a more effective use of resources and, ultimately, profit maximization.

Employ statistical approaches to examine data: The team includes expertise in data modeling and analysis, machine learning, and statistical analysis. They have access to a variety of datasets from credible sources, as well as the most up-to-date analytical tools and software, allowing them to produce accurate and useful findings.

Analyze the project's influence on operations: The efficacy of the project may also be measured by examining its impacts on the hotel's operations. If the project results in more efficient operations, cost savings and increased profitability are possible outcomes.

In summary, the project's data mining goals include developing a customer segmentation model to improve marketing techniques, measuring success with financial and non-financial metrics, ensuring data security and quality, optimizing resource use, using statistical techniques to analyze data, and evaluating the project's impact on operations.

# METHODOLOGY

The provided dataset includes information about hotel guests, such as their identification, nationality, and age, as well as their booking and stay history, revenue produced, and room preferences. The dataset contains a combination of categorical and numeric data, with certain attributes represented as binary indications (0 or 1). The initial dataset consisted of 111,733 rows.

## Data understanding

**Categorical Variables:**

Nationality: Nominal categorical variable indicating the nationality of the customer in ISO 3166-1 (Alpha 3) format.

MarketSegment: Nominal categorical variable indicating the current market segment of the customer.

Distribution Channel - Distribution Channel normally used by customer to make bookings to the hotel

SRHighFloor, SRLowFloor, SRAccessibleRoom, SRMediumFloor, SRBathtub, SRShower, SRCrib, SRKingSizeBed, SRTwinBed, SRNearElevator, SRAwayFromElevator, SRNoAlcoholInMiniBar, SRQuietRoom: Binary categorical variables indicating whether the customer usually requests a specific room feature or not.

**Numerical Variables:**

**ID:** Unique numerical identifier for each customer.

**Age:** Age of the customer.

**DaysSinceCreation:** Number of elapsed days since the customer was created.

**AverageLeadTime:** Average number of days before arrival date the customer makes bookings.

**LodgingRevenue:** Total amount of lodging revenue paid by the customer so far.

**OtherRevenue:** Total amount of other revenue (e.g., food & beverage, spa, etc.) paid by the customer so far.

**BookingsCanceled**: Number of bookings the customer made but subsequently canceled.

**BookingsNoShowed:** Number of bookings the customer made but subsequently made a "no-show".

**BookingsCheckedin:** Number of bookings the customer made, which actually ended up staying.

**PersonNights:** Total person/nights the customer has stayed at the hotel so far.

**RoomNights:** Total of room/nights the customer has stayed at the hotel so far.

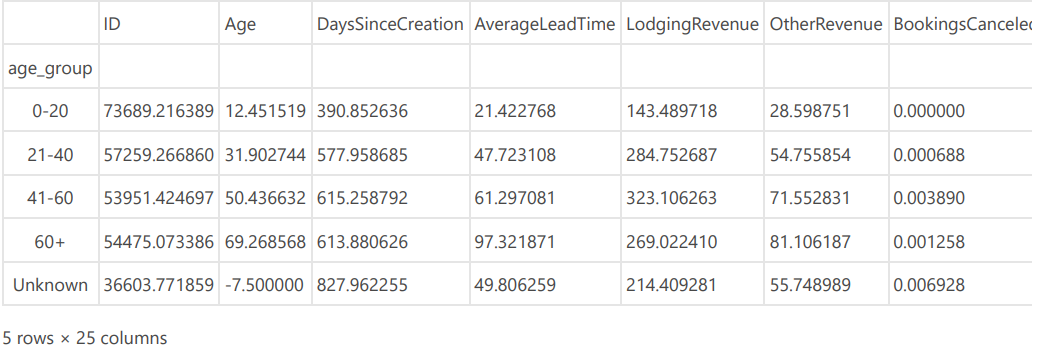
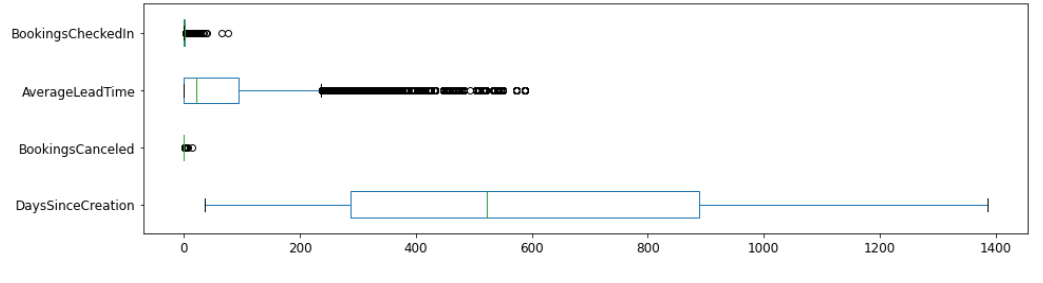
**Hash Variables:**

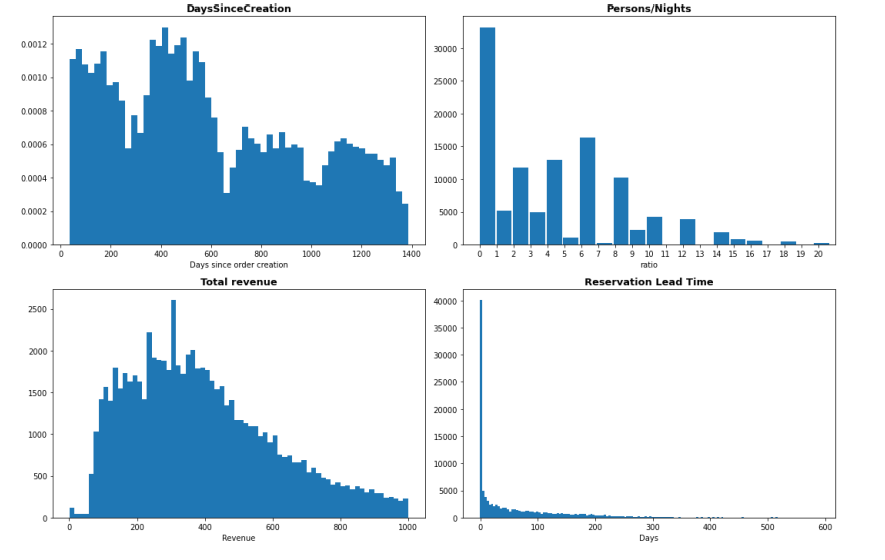
NameHash: Hash of the customer name.

DocIDHash: Hash of the customer personal document identification number (usually passport or ID card).

The data includes both categorical and numerical variables, with the majority of the variables being binary categorical indicating whether a customer typically requests a certain room feature or not. There are also several numerical variables indicating various aspects of a customer's booking history and stay at the hotel. Additionally, there are two hash variables for the customer's name and personal identification number. Please see figure 2 in the appendix to get a better understanding on the data we worked with.

## Data preparation

Data preparation is an essential aspect of any data analysis endeavor. In the provided dataset, the 'Age' column contained 3,779 missing entries. During the clustering step, we will impute missing values using appropriate strategies to address the issue. In addition, we added two new variables to our dataset to improve its description. The first variable, 'total revenue,' reflects the total amount a customer spent on lodging and other expenses, like food, spa, etc. This variable will help us determine which hotel guests spend the most on amenities and services. The second variable, 'age group,' classifies clients into one of the following four age groups: '0-20', '21-40', '41-60', and '60+'. Any missing or unsuitable data was tagged "Unknown." These new variables will provide important insight into consumer categories, allowing us to customize our marketing and sales tactics to fit their specific requirements.   
  
  
  
The graph above explains the average values for many variables for each age group, as well as a category labeled "Unknown" for variables with missing or incorrect values. It is noteworthy to observe that clients between the ages of 0 and 20 have the shortest average lead time (390.85 days) and the highest average lodging revenue (21.42), indicating that they may be more prone to book at the last minute and choose more expensive hotels. The '60+' age group has the largest average lead time (613.88 days) and average other revenue (269.02), indicating that they are more inclined to plan ahead and spend more on other hotel facilities.  
  
In order to understand our numerical variables we presented them into a box plot and histograms as seen in the figures below.   
  




Following a comprehensive study of the data, it has been determined that a number of modifications are essential. Initially, records without DocIdHash will be disregarded, but the single customer with no check-ins and revenue greater than zero will be maintained. In addition, a 'Special Requests marker' will be produced to indicate whether or not the client has previously submitted a request. Two ratios will be produced to calculate the average length of stay and room density, using IQR or manually if IQR is not applicable. RoomNights and PersonNights will be eliminated together with customers whose Lodging and Other revenue are both equal to zero. Entries with redundant DocIdHashes will be kept, even if they have the same NameHash. Efforts will be made to fill in missing Age entries using a classifier, but they will be discarded if an acceptable level of accuracy cannot be reached. Entries with an Age of 0 or below will be categorized as NaNs during the binning process. Last but not least, the BookingsCanceled and BookingsNoShown variables will be eliminated from the segmentation. These improvements will help to clean and optimize the data, allowing for a more accurate and successful segmentation process.

## Modeling

In this study, we have utilized the K-means clustering algorithm to categorize clients based on their spending patterns. K-means is a popular unsupervised machine learning technique that groups data points into a predetermined number of clusters according to the similarity of their attributes. To calculate the best number of clusters for our dataset, we employed the "Elbow Method," which identifies the value of k at which the distortion or inertia begins to decrease linearly. At the elbow, the value of k represents the ideal number of clusters. We have also evaluated the quality of the clustering results using the Silhouette score, which analyzes the distance between each data point and its allocated cluster to determine how well each data point fits into its cluster. Principal Component Analysis (PCA) was used to further visualize the clustering results in order to identify any patterns or correlations between the various features. The K-means clustering method has supplied us with significant insights into the spending patterns of different client segments, which can be applied to the development of tailored marketing and sales strategies to boost the hotel's revenue and profitability. Using the elbow method, we decided to form six clusters. In the graph, the elbow point is the point of inflection, where the rate of decrease of inertia abruptly shifts. The elbow point was about 6, indicating that six clusters would be the ideal number for our data set. In addition, we needed to construct enough clusters to adequately differentiate between the various categories of clients, but not so many that the clusters were insignificant or merged. By establishing six clusters, we were able to efficiently group clients with similar characteristics together while preserving sufficient variability within each cluster. In addition, this number of clusters matches well with prevalent marketing and sales tactics, which frequently focus on dividing customers into four to six discrete groups for focused campaigns. We normalized the data with the MinMaxScaler.

In addition to exploring the most significant variables, we employed PCA to minimize the dataset's dimensionality and accelerate model construction. Large datasets with numerous variables can impose significant computational strain and delay model development, as is well known. PCA is a frequently employed technique in machine learning for addressing this problem by selecting the most important features that contribute the most variance to the data. By utilizing PCA, we are able to reduce the number of variables in our dataset without sacrificing a substantial amount of information. This can lead to a reduction in processing time, an improvement in performance, and a more efficient use of computing resources. Using PCA, we transformed the original dataset into a new collection of main components that captured the greatest amount of variance in the data while reducing their correlation. This allowed us to deal with a smaller, more manageable dataset without sacrificing analysis accuracy.

# RESULTS EVALUATION

Using the k-means clustering technique, the data appear to have been grouped into six groups (designated clusters 0 through 5) based on the available information. However, the silhouette score of 0.3729 suggests that the clustering arrangement may not be optimal, as the majority of items do not have a high value.

Each cluster's description gives some insight into the behavior of clients belonging to that cluster. Cluster 0 clients, for instance, are often younger, have a large number of bookings, cancel or fail to appear for their reservations more frequently, but spend more on lodging and other income. Cluster 3 clients, on the other hand, are comparatively older, have a large number of bookings, but are more likely to cancel or not show up for their reservations, and have a relatively low quantity of lodging and other income. Please see appendix figure 5 to get a better comprehending on the analysis.

The research also gives some information about the booking behavior of clients in each cluster, such as the lead time for booking and the number of persons and hotel nights. Cluster 1 clients, for example, prefer to book rooms with a longer lead time and a relatively high number of person and room nights, whereas Cluster 4 customers tend to book rooms with a shorter lead time and a relatively low number of person and room nights. Below we have explained every cluster and provided a sales and marketing suggestion:   
  
**Cluster 0** depicts younger clients with a high number of bookings who cancel or fail to appear for their reservations more frequently than other customers. They spend more on accommodation and other earnings, but have fewer guests and hotel nights.

Marketing and sales techniques:

Provide lower rates for extended stays or complimentary hotel upgrades for booking a particular number of nights.

Use social media and influencer marketing to target this younger demographic and promote the hotel's popular features and activities for younger tourists.

Implementing a more flexible cancellation policy with a shorter lead time may encourage guests to book more frequently and decrease the chance of cancellations and no-shows.

**Cluster 1** consists of older clients who have been created for a longer period of time, who book rooms with a longer lead time, and who are less likely to cancel or not show up for their reservations. They have a comparatively large number of guests and room nights and spend more on accommodation and other revenue.

Marketing and sales techniques:

Provide loyalty programs or membership benefits that cater to their regular visits, such as access to unique amenities or discounted pricing for repeat bookings.

In marketing campaigns and promotional materials, highlight the hotel's history and heritage, which may appeal to elderly tourists who value tradition and elegance.

Advertise the hotel's meeting and event spaces, as older guests may be more inclined to attend conferences and business gatherings.

**Cluster 2** consists of younger clients who have been created for a shorter period of time, book rooms with a shorter lead time, and generate a relatively modest amount of lodging and other revenue. Also, they have a low quantity of guests and room nights.

Marketing and sales techniques:

Advertise the hotel's location and proximity to local attractions, as younger guests may be more interested in exploring the neighborhood.

Provide short-stay packages or bargains, such as a weekend getaway package or a last-minute booking discount.

Use social media and online reviews to highlight the hotel's fashionable and contemporary features that appeal to younger guests, such as rooftop bars and fitness centers.

**Cluster 3:** This cluster reflects elderly consumers with a significant number of bookings who cancel or fail to appear for their reservations more frequently than other customers. They have a low number of guests and room nights and a low amount of lodging and other earnings.

Marketing and sales techniques:

Provide flexible cancellation rules to accommodate senior guests with unexpected schedules or health conditions.

Advertise the hotel's wellness and relaxation amenities, as senior guests may prioritize self-care and relaxation.

Use email marketing campaigns to target this group with personalized offers or discounts that encourage them to book more frequently and prevent cancellations or no-shows.

**Cluster 4:** This cluster consists of younger clients that have a small number of bookings, reserve rooms with a shorter lead time, and generate a modest quantity of lodging and other revenue. Also, they have a low quantity of guests and room nights.

Marketing and sales techniques:

Provide unique and tailored experiences, such as food and drink sampling and outdoor activities, that cater to younger guests.

Use social media influencers or user-generated material to highlight the hotel's innovative and Instagrammable amenities that appeal to younger tourists.

Use email marketing campaigns to target this cluster with last-minute bargains and promotions that encourage them to book more frequently and boost revenue.

**Cluster 5** consists of clients that are somewhat elderly and have few bookings. They prefer to reserve rooms with a longer lead time, and their hotel and other revenue is generally low. In addition, they have a comparatively high number of guests and hotel nights.

These guests may represent a chance for the hotel to boost income by upselling them room upgrades and additional amenities during their extended stays. To entice customers in this cluster to book more regularly or extend their stays, the hotel should provide them with special packages or discounts.

**Marketing and sales techniques:**

Marketing methods could include email campaigns or targeted commercials that highlight the benefits of longer stays, such as the chance to explore the local area in depth or to relax in a pleasant hotel room. Also, the hotel could provide loyalty programs or other incentives to entice customers in this cluster to book more regularly or refer their friends.

Depending on the length of the guest's stay, sales strategies could include delivering tailored recommendations for local activities or eateries. Also, the hotel staff could offer to arrange transportation or other services to make the guest's stay more enjoyable and comfortable. The hotel can generate repeat bookings and positive word-of-mouth referrals by establishing strong relationships with customers in this cluster.

# DEPLOYMENT AND MAINTENANCE PLANS

Using the segmentation results, an AI-based deployment approach may customize the mobile app experience for each individual consumer.

The app might employ machine learning algorithms to analyze client data and behavior in order to make customized suggestions for hotel facilities, services, and promotions. For instance, if a customer is classified as a business traveler, the app could promote the hotel's meeting room features and provide discounts for business-related services. If a consumer is determined to be a leisure visitor, the app may promote the hotel's recreational activities and sightseeing packages.

To implement this plan, the hotel may collaborate with a team of AI specialists and developers to create and train machine learning models and integrate them into the app's backend. Additionally, the app might be enhanced with additional features and capabilities to provide tailored recommendations and promotions.

To monitor and maintain the performance of the model, the app's usage data could be watched and evaluated, and any bugs or errors could be swiftly remedied. Moreover, testing and upgrades could be performed on a regular basis to verify that the model continues to work efficiently and properly over time. Additionally, customer comments and ratings could be utilized to evaluate the performance of the individualized recommendations and identify areas for improvement.

## 5.1 Personnel

Data scientists, data engineers, software developers, and IT specialists might be engaged in the deployment of the model.

The data storage infrastructure, application programming interfaces (APIs), and web or mobile applications utilized by clients are examples of systems that may require modification.

The method for deploying a machine learning model into production needs collaboration across data science, engineering, and IT teams to guarantee the model is accurate, dependable, and scalable.

## 5.2 Monitoring

Once the application has been released, it is necessary to regularly monitor and maintain the system to ensure optimal performance and accuracy of forecasts. Many steps are involved, including:

Monitoring: A monitoring system should be established to track the performance of the system in real-time. This system should monitor the model's predictions, system uptime, response times, and other pertinent metrics to identify problems as they arise.

Logging: All system operations and events should be logged for later examination. This can aid in identifying areas for improvement and in diagnosing any difficulties that may arise.

Updating: The system, including the model, should be regularly updated to accommodate shifting trends, patterns, and user requirements. These updates should be rigorously evaluated to guarantee that they have no negative impact on system performance.

Retraining: The model should be frequently retrained with new data to enhance its accuracy and ensure that it remains relevant to the present scenario.

Security: There should be security measures in place to prevent unwanted access to the system and secure sensitive user data.

User feedback: Users should be encouraged to submit feedback on the system, including its precision, usability, and any faults they encounter. These comments can help direct future updates and enhancements.

In general, the upkeep of an AI-based application necessitates continual effort and vigilance to ensure optimal system performance and user satisfaction.

# CONCLUSIONS

Our examination of the hotel's customer data has yielded important insights regarding the hotel's visitors' spending and behavior habits. By analyzing six separate client clusters, we have determined Cluster 1 and Cluster 3 as the most profitable for the hotel to target. These clusters reflect the hotel's most valuable customers in terms of expenditure and loyalty.

To leverage on these clusters' potential, the hotel might use targeted marketing methods to attract and keep these important visitors. For instance, the hotel can provide customized loyalty programs and packages based on the tastes of these clusters, as well as personalized recommendations and services to enhance their stay. By doing so, the hotel will be able to increase client retention and repeat bookings, resulting in improved revenue per customer and overall profitability.

Understanding and responding to the wants and preferences of the most profitable customer segments is vital for success in the highly competitive hospitality sector. With the information gleaned from this analysis, the hotel may build targeted marketing and sales strategies to improve revenue, increase profitability, and secure its position as an industry leader.

## Considerations for model improvement

Despite the fact that our analysis has revealed useful insights into the behavior and spending patterns of hotel guests, there are a number of ways to improve the model's precision.

Initially, the use of k-means clustering and silhouette score was helpful in identifying the six separate consumer clusters; however, other clustering methods or metrics may have produced even better results. For instance, hierarchical clustering or density-based clustering could be investigated to determine whether they produce more precise and reliable grouping results. Additionally, the use of PCA (Principal Component Analysis) allowed us to reduce the dimensionality of the data and visualize the clusters in a two-dimensional space. However, it is worthwhile to investigate other dimensionality reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection) to determine if they provide even better visualizations of the clusters.

There may be more features or variables that might be added to the analysis to improve the model's precision. For instance, the purpose of the trip, the booking source, and the type of hotel rented could be included to acquire a deeper understanding of the customer's behavior and preferences.

To ensure their effectiveness and relevance, it is crucial to test the clustering results and marketing tactics with real-world data and client feedback. Incorporating new data and input into the model on a regular basis can also aid in enhancing the accuracy and applicability of the analysis over time.

In conclusion, while our current analysis has provided valuable insights into customer behavior and preferences, there are a number of ways to improve the model's precision and efficacy. By investigating various clustering methods, dimensionality reduction approaches, and incorporating more factors, we can acquire a more comprehensive understanding of customer behavior and preferences, which can result in even more effective marketing and sales tactics for the hotel.

# APPENDIX

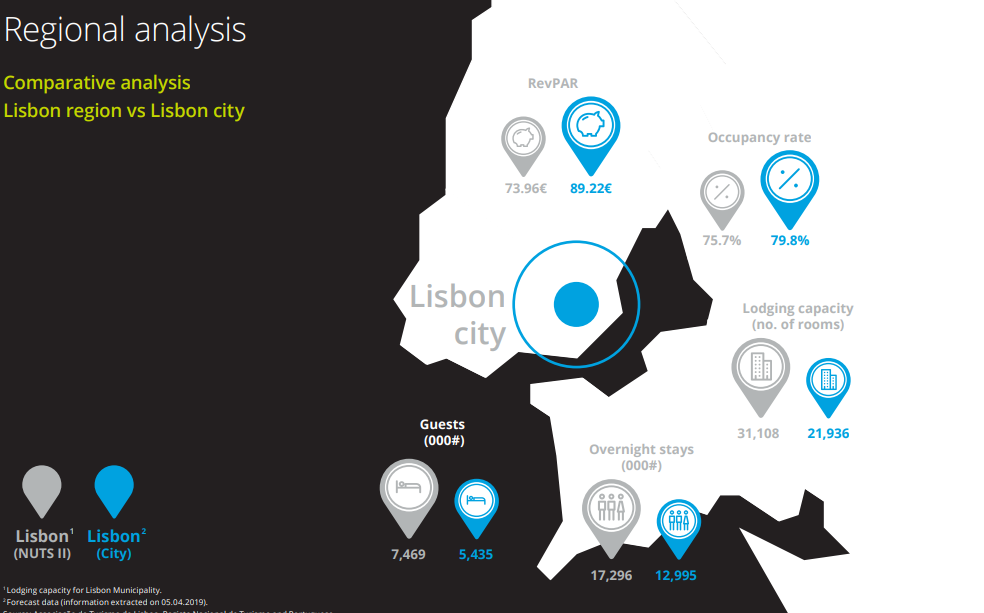
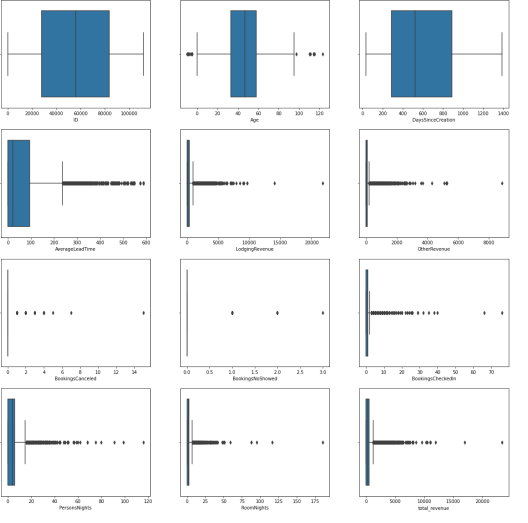


Fig 1 ( Not our finding, Deloitte graph, Retrieved at <https://www2.deloitte.com/content/dam/Deloitte/pt/Documents/transportation-infrastructures-services/Portuguese%20Hospitality%20Atlas%202019.pdf>)

  
Fig 2 ( Numerical Features)

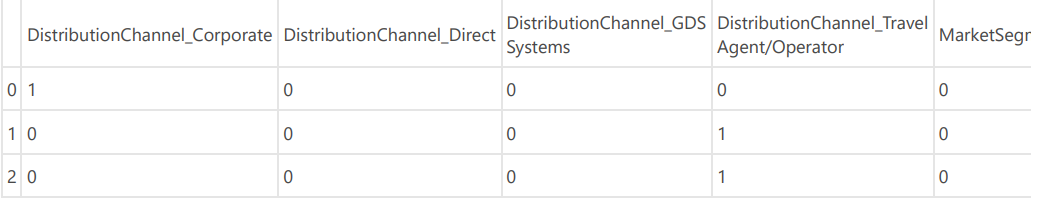
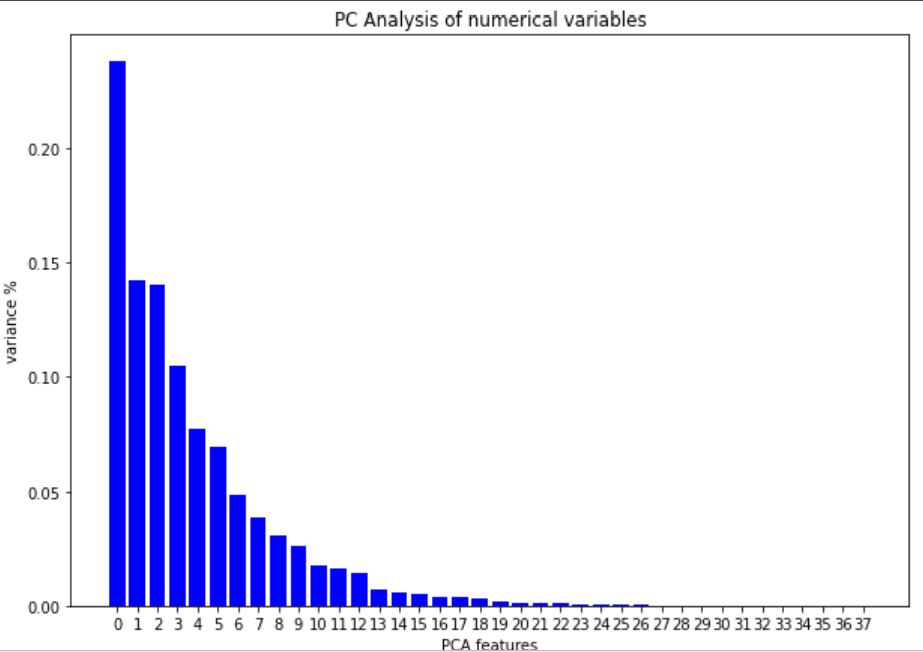
  
Fig 3 ( Dropping unnecessary components)   
  


Fig 4 ( PCA components)

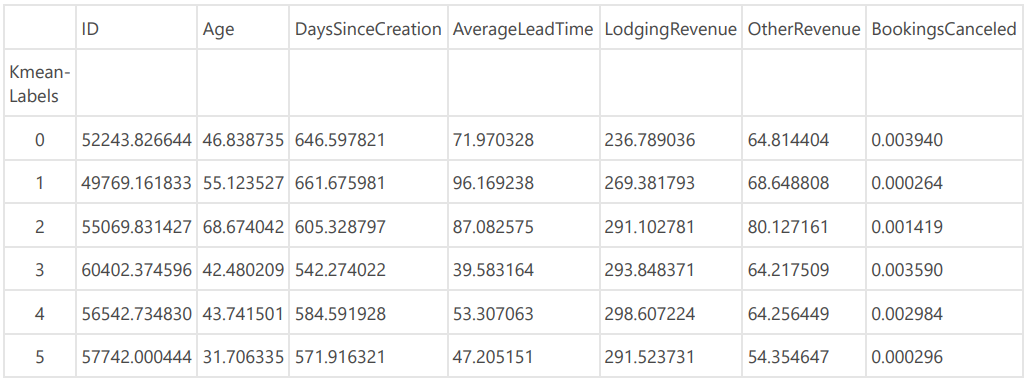


Fig 5 ( K - Means)

1. Woods, R. (2019, June 23). Clustering the top 25 hotel markets in the US. Retrieved March 7, 2023, from https://hotelinvestmentstrategies.com/clustering-the-top-25-hotel-markets-in-the-us/ [↑](#footnote-ref-1)
2. Return-on-investment approach (ROI). (2004). *Dictionary of Marketing Communications*. doi:10.4135/9781452229669.n3128 [↑](#footnote-ref-2)