NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel Customer Segmentation

Carolina, Caldeira, number: 20230440

Goncalo, Caldeirinha, number: 20230469

Madalena, Figueirinhas, number: 20230436

Martyna, Kmiecik, number: 20230452

Alina, Metzger, number: 20230998

Group C

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

LIST OF TABLES AND FIGURES

1. EXECUTIVE SUMMARY

The report, prepared for the Hotel H in Lisbon, Portugal, presents an analysis and grouping of the hotel's clientele into individual market segments, with the aim to learn as much as possible from the existing customers to eventually also find new customers. The project was initiated in response to a need for the hotel, which, thanks to a new marketing manager, decided to move to a more sophisticated customer segmentation approach that takes into account multi-faceted customer preferences, demographics and behaviors. This urge arose through expansion and increased marketing investment following the chain's acquisition of new hotels, and to increase the hotel's profitability.

The methodology used, called CRISP-DM, included careful data understanding, preparation and modeling using principal component analysis (PCA) and the K-Means ++ clustering algorithm to segment the hotel's customer base into five distinct clusters. These groups were characterized based on demographic, geographic, psychographic, and behavioral attributes, providing a more detailed picture of the hotel's customers than the previous segmentation method.

The success of the project has been assessed based on various quantitative and qualitative criteria, including clarity in identifying distinct customer segments allowing for effective targeted marketing strategies, increased efficiency and improved customer loyalty and satisfaction. The implementation and maintenance plans for the segmentation model emphasize the need for regular updates, monitoring, and collaboration between not only marketing but also IT departments to ensure that the model is still effective and relevant in view of the new circumstances.

Business conclusions from the analysis include marketing strategies tailored to each identified segment, including loyalty programs, personalized benefits and family-friendly amenities for European travelers; corporate packages and facilities for business travelers as well as offerings for personalized experiences for mature travelers and international guests. These strategies are designed to improve customer satisfaction and loyalty, and the overall company performance.

The report closes with recommendations for further improving the segmentation model and recognizes potential challenges related to data quality and privacy issues. Finally, the analysis provides Hotel H with a strategic framework to improve business performance and engage customers through data-driven insights.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Hotel H, located in Lisbon, Portugal, is a member of the independent hotel chain C. Hotel chain C operated 4 hotels until 2015. The acquisition of new hotels made the hotel chain board invest more in marketing. In 2018, a marketing department was created, and a new marketing manager hired, 'A'. Historically relying on customer segmentation based primarily on one customer characteristic, its sales origin, Hotel H, under the direction of the new marketing manager 'A', recognizes the need for a distinct approach. This need comes from the realization that the current segmentation does not adequately capture the diverse preferences and behaviors of its customers, particularly considering the varied distribution channels through which customers now book their stays.

2.2. BUSINESS OBJECTIVES

In order to find new customers and continue to captivate the current ones, learning as much as possible about the existing customers is crucial. By properly identifying customers' groups, Hotel H' marketing manager 'A' can make better strategic choices.

Firstly, one objective of the project is to develop a multidimensional customer segmentation model that uses geographic, demographic, psychographic and behavioral characteristics.

Secondly, use the enhanced segmentation to come up with new targeted marketing strategies, product development and service offerings that meet the distinct needs of different customer groups.

Lastly, the goal of this project is to keep existing customers, attract new customers and enhance overall business performance.

2.3. BUSINESS SUCCESS CRITERIA

The business success criteria will be evaluated based on both quantitative and qualitative measures. The analysis' ability to clearly identify distinct customer segments that allows the creation of effective targeted marketing strategies is particularly crucial. The new segments should include the characteristics of the customer that were previously omitted, such as age, number of stays or nationality. It is also expected to lead to improved marketing efficiency, through increased engagement rates and a greater return on marketing investment. Additionally, another success criteria are the strengthening of the hotel H's competitive advantage in the hospitality market, due to revenue growth. Lastly, the increase in customer satisfaction and loyalty can be measured through the rise in repeated bookings rate and positive customer feedback.

2.4. SITUATION ASSESSMENT

Regarding the resources, the data used for the evaluation is Hotel H's existing data on customer details, transaction histories and booking patterns. The marketing department, led by manager 'A', will work closely with the IT department to ensure data quality and accessibility. The data in this project is managed by the team of five analytics students from Nova IMS, using Jupyter Notebook, a web-based interactive computing platform, capable of handling complex data mining tasks, along with customer relationship management systems and data visualization tools.

Some risks include data privacy concerns and the potential for inaccurate data leading to improper segmentations. However, to mitigate these risks, some contingencies can be made. Hashing customer names and document ids, as in the case of Hotel H's dataset, is a contingency action regarding data

privacy. The collection of supplementary data is another contingency plan that can be done to mitigate the risk of inaccurate data.

Regarding terminology, it is important that stakeholders understand the project key terms such as 'K-means clustering', 'Principal Component Analysis (PCA)', 'Elbow Method' and 'Silhouette Score'. In the appendix a glossary of these terms will be provided for clarity and alignment.

The costs associated with this project include the team's work hours and potential data acquisition costs. In return, the expected benefits are considerable, boosting customer satisfaction and loyalty, offering the potential for revenue growth, through more effective marketing strategies, and promising a great return on investment and competitive advantage in the hospitality market.

2.5. Determine Data Mining Goals

The project consists of three main phases: the understanding and exploration of the data available and variable transformation; the use of a model to identify customer segments; and the suggestion of targeted business strategies and insights.

In technical terms, the objectives include the application of Principal Component Analysis to reduce the dimensionality of the dataset, improving the efficiency and accuracy of the clustering process. The selection of an appropriate number of clusters, K, involves the application of the Elbow method to find a balance between the level of detail in the segmentation and the practical use of the clusters to develop marketing strategies. Lastly, the utilization of an unsupervised classification algorithm, K-Means ++, to segment the hotel's customer base into groups.

Our data mining success criteria is measured by the silhouette score, which is a metric used to evaluate the quality of the clusters, based on within-cluster similarity and out-of-cluster discrimination. The success criteria are considered to be the average silhouette score equal to or above 0.25. [1] Moreover, the DM aim of this clustering analysis is to obtain separate, different segments, while utilizing demographic, geographic and behavioral characteristics of the customers.

3. METHODOLOGY

3.1. Data Understanding

Understanding customer data is necessary in the competitive landscape of hospitality industry in order to both identify new opportunities and improve analysis tools of recognizing existing clients as individuals belonging to different market segments. The provided dataset, a property of Hotel H, presents a comprehensive overview of its customer demographics, behaviors and preferences.

It consists of 111,733 rows, each of them representing a customer of the hotel H. The available data was collected for about 3 years.

In general, the dataset includes demographic information and one identifying the customers such as their ID number, name, age, or nationality. Moreover, it includes data regarding revenue and spending behavior, that is lodging another revenue revealing if they are more generous and unbounded while staying at the hotel or frugal.

[1] Cf. Shahapure, K. R., & Nicholas, C. (2020).

From the dataset one may conclude also about customer engagement and interaction. It provides an overview of when the customer was first registered in the system, indicating if a client is an old or more recent one, as well as an overview of their cancellations, no-shows, and checked-in stays. This knowledge gives insights if the customer has ever even stayed in the hotel, since having zero bookings checked in, which in this case appears to be a frequent problem as around 30% of observations have this value equal to 0, indicates no stays in the establishment ever. These customers may have potentially enrolled in the loyalty program and not necessarily intended to make a booking. They can be already treated as a separate part of the analysis and the marketing approach applied in their case should be different than for customers belonging to other market segments. The same is applicable to hotel customers who booked a stay, however, never made it due to cancellations or no-shows. Their some data such as age, nationality, or ID, as well as those of the previous ones, may not be correct, as the hotel collects this information after check-in, which never happened in the case of both groups, hence segmenting them might not be fully correct and justified.

In addition, this dataset allows one to draw conclusions about customers' booking habits, as it reveals how long before their stay they make a reservation, potentially indicating private, corporate, or just plain cautious customers. Also, looking into information about how many rooms, for how many nights or how many people are staying with the customer for how long may provide insights into the utilization patterns of the hotel's facilities and services.

Furthermore, the dataset includes current campaign data - market segment to which the customers were previously assigned and booking channels through which they usually book the stay.

Lastly, for each client one can check their preferences for room specifications, such as a king-size bed, quiet room, a crib and many more.

Initial insight into the data suggests that the mean age of customers is about 46 and on average they have existed in the database for about 1.5 years. They come from almost 200 different nationalities. Reservations are most often made by a travel agent or operator and are on average made 60 days in advance. Customer spending on lodging ranges from 0 to 21,781 but averages around 283. While most special requests are not very popular, the request for a king-size bed is a definite exception as more than 30% of registered customers in the database requested this type of bed.

It shall also be mentioned, that by looking deeper into statistics of the data and while evaluating IDs of the customers, an unusual occurrence has been found where over 3,000 rows were assigned the same ID. This was identified as different transactions performed by the hotel regarding customers. Therefore, these cases cannot be considered as real customers, and consequently they cannot be grouped together. It is important to keep this in mind for subsequent analyses as it may falsify the results and lead to misleading conclusions.

Another issue to consider are some inconsistencies and missing values that can be found in the data, which should be handled with special care when dealing with them, in order to obtain reliable results. In addition, some observations were identified in the data as duplicates since they presented exactly the same information, which may indicate an accidental mishandling of the data, as they present the same customers.

This dataset enables the identification of patterns and trends that will result in strategic decision making improving the marketing strategies for the hotel. By utilizing this data, the new marketing manager can go beyond traditional segmentation based on customer origin and take a more nuanced approach taking into account some of the aforementioned factors. Nevertheless, the data quality is of course not perfect, and the problems found hinder a perfect and error-free analysis. Nonetheless, these details are the key for tailoring marketing strategies and in long-term, attracting new customers while ensuring the satisfaction and loyalty of existing ones.

3.2. DATA PREPARATION

In order to get the data ready for the modelling phase and achieve the most accurate analysis of customer segmentation the data preparation has taken place.

3.2.1. Removal of Some Observations in the Data

In the beginning the duplicated rows suggesting accidental recording of the same customer were dropped, to clear the data and avoid redundancy. Next step consisted of handling around 3,000 rows, with the same value of encrypted document ID, indicating the transactions regarding the customers recorded by the hotel. These were removed as well, as they do not represent real people and therefore are not valuable for the analysis. Lastly, people with 0 bookings checked in were also removed from the dataset in order to retain only observations that have accurate and reliable information. This is because the cases in question, which accounted for about 30% of the data, may have had the wrong customer's nationality, age, and their document number in the first place, as this information is only collected by the hotel at check-in, not at booking. Additionally, if the customer has never stayed at the hotel, the facility also does not have access to their booking habits which would ultimately lead to erroneous conclusions.

3.2.2. Handling Missing Values

The previous step already led to a significant improvement against missing values. This was because afterwards they could be only found in the age variable and single instances in the document id. All observations related to negative age, were also changed to missing and then the missing data was imputed based on the completion based on similar observations from the dataset. 4 observations without an available document id were removed due to uncertainty related to the rest of the available information regarding them.

3.2.3. Dealing with Duplicated IDs

As it might seem that the document ID should be unique, this was not the case for this dataset. Some observations had the same document ID; however, they differed in terms of other variables. It was decided to group those observations that contained a common nationality, name, and document id, as they then clearly suggest the same person.

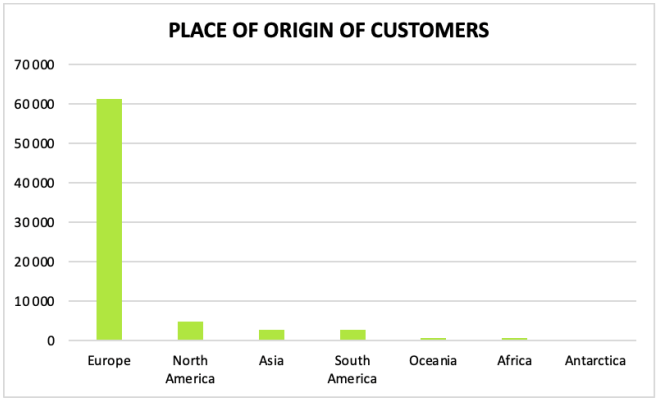
3.2.4. Other Inconsistencies

Then, for data preparation, the remaining inconsistencies were addressed. The age of those with positive income was changed to be consistent with their expenses. In addition, children were removed from the dataset as it would be impossible to approach them with marketing strategy anyway as they are underage. Also, individual negative average lead time values have been corrected.

3.2.5. Feature Engineering

A variable continent was introduced to reduce information on nations and group customers into common bunch. This step made it possible to see more clearly the distribution of the regions from which customers come. As expected, it can be seen that the vast majority of customers come from Europe.

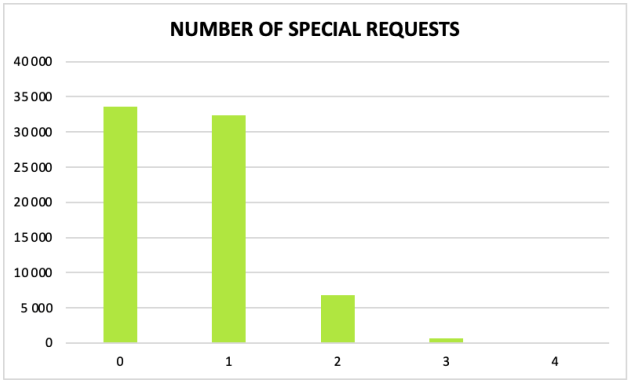
Figure 1: Customers' Continents of Origins



Source: Own Elaboration based on Jupyter Notebook.

Next, a variable on how many special requests is recorded for each customer during their stay was introduced, which showed that the most customers usually have no special requests, followed by customers with 1 each, and the rest between 2 and 4.

Figure 2: Number of Special Requests by Customers



Source: Own Elaboration based on Jupyter Notebook.

Then three new variables were introduced - Total Revenue which is the sum of Lodging and Other revenue, Average Daily Spend of the customer and a variable that reports Not Coming Ratio. Finally, some binary variables indicating special requests were removed, as a significant minority of customers requested them, so they would be redundant for the clustering analysis.

In the next step outliers were handled as well as highly correlated variables. Eventually, variables ‚LodgingRevenue' and 'OtherRevenue' were removed as their information is available in Total Revenue variable. Variables ‚PersonsNights' was also deleted as it was highly correlated with Room Nights which is a more used metric in the industry. Finally, ‚Market Segment' was dropped as well, as this was an old clustering result, whereas the goal of the current analysis is to assign a new one for the customers.

In addition, a new variable was included in the analysis testifying whether a customer comes from Portugal or is a foreigner. The latter represent the vast majority.

3.2.6. Encoding the Data, Normalization and Results

The next step of data preparation focused on encoding the data and its normalization. Finally, the clean data, consisting of 73,431 observations and 22 variables, and its statistics reveal that: the average age of Hotel H customer is 48.5, and they usually book their stay 685 days in advance. Moreover, an average customer makes around 0.65 special requests, spends in total around 470€ while staying, that is on average for the ordinary customer 83€ per day. The detailed statistics are presented below.

Table 1: Statistics of Customers Attributes after Data Preparation

Variables	mean	std	min	max
Age	48.49	14.71	18.0	89.5
DaysSinceCreation	685.36	375.25	36.0	1385.0
AverageLeadTime	88.61	90.09	0.0	588.0
RoomNights	3.15	1.58	1.0	7.0
NumberOfRequests	0.65	0.68	0.0	4.0
TotalRevenue	469.69	292.48	0.0	1152.0
AverageDailySpend	82.90	38.87	0.0	167.5

Source: Own Elaboration based on Jupyter Notebook.

3.3. MODELING

After preparing the data for analysis, the complex and diverse customer data was simplified through the application of a Principal Component Analysis (PCA) process. This crucial step resulted in dimensionality reduction of the data. Therefore, the information was clarified, enabling focusing on the most significant customer characteristics. In doing so, the foundations were laid for making strategic business decisions based on clear, actionable insights.

Identifying the right number of customer segments was essential. A method known as the Elbow method was deployed, complemented by a tool called *KElbowVisualizer*, to identify the optimal segmentation strategy. Moreover, the number of obtained clusters was confirmed through the Silhouette score assessment application. This careful analysis led to defining five distinct customer segments, enabling the balance of complexity with strategic focus. This segmentation is not just about organization; it is about better understanding hotel customers, allowing for more targeted and effective marketing as well as for engagement approaches.

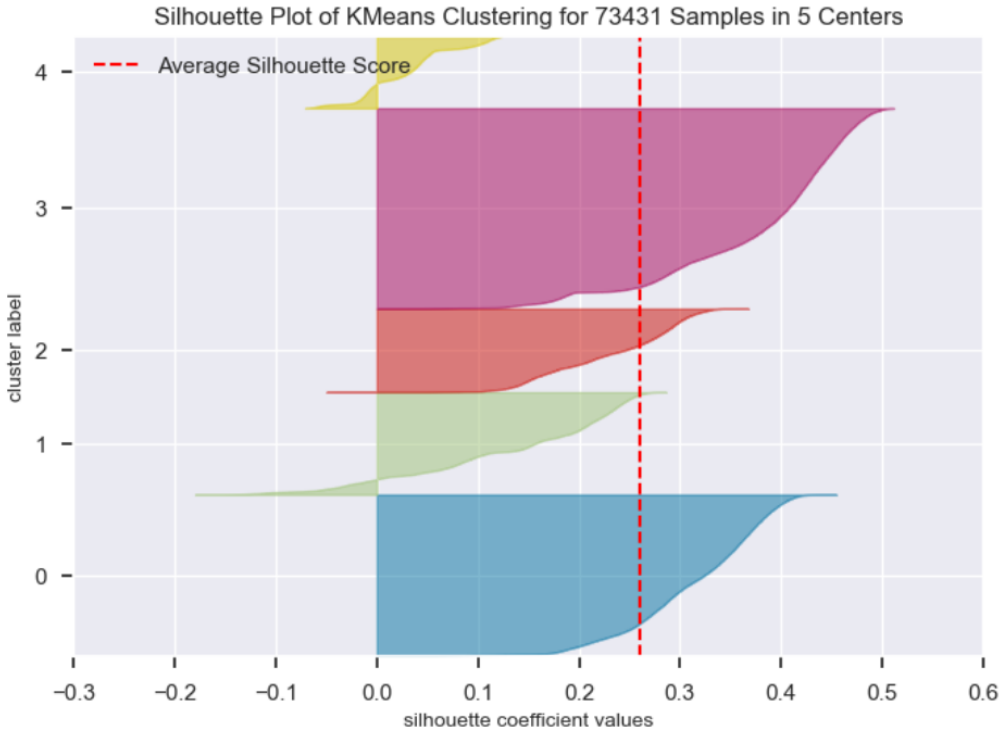
Finally, a clustering strategy using an algorithm known as K-Means ++, known for its efficiency, speed and great accuracy regarding the results, was implemented[2]. This approach segmented the customers into five unique groups, each defined by specific patterns, characteristics, and preferences. That may not only be described as just a technical exercise, but as a step towards a deeper understanding of the customers' values and the path to their satisfaction. By recognizing and appreciating the uniqueness of each customer segment, the services and marketing efforts can be adapted, resulting in an overall increase in customer satisfaction and profitability.

[2] Cf. Zhao, Z., Wang, J., & Liu, Y. (2017).

3.4. EVALUATION

The final results are composed of 5 different clusters. After modeling, several results were obtained with respect to the quality of the clusters and the average values for each cluster as well. There are three important measures to look at, these being: silhouette scores, cardinality, and magnitude.

Figure 3: Silhouette Plot of KMeans Clustering

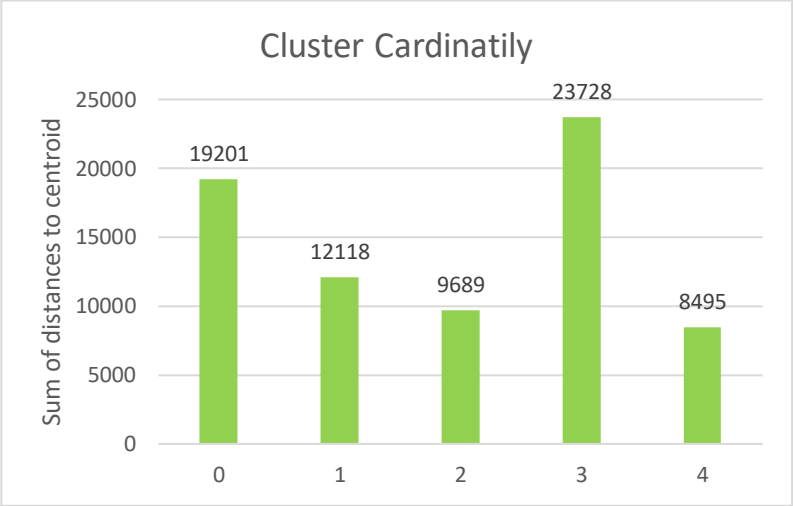


Source: Own Elaboration based on Jupyter Notebook.

The present graph (Figure 3) shows the silhouette score for each cluster and the respective size of the clusters. The silhouette score is a measure of the quality of the clusters by taking into account how similar the individuals in the same cluster are to each other and how distant they are to the next closest cluster. The X axis represents the silhouette score for each individual and the y axis represents the different clusters and the individuals that compose them. The colors demonstrate the five different clusters, and the area of each color demonstrates their size. Some observations about the graph:

- The average silhouette score for the segmentation is around 0.26: since this value can vary between -1 and 1 and the overall result is positive, it is possible to conclude that generally, each customer is in the correct cluster;
- Looking at each individual cluster there are some individuals that are incorrectly classified as there are some negative silhouette values although the percentage of them is much smaller compared to the ones with a positive silhouette score;
- The size of the clusters varies a lot, which will later be analyzed.

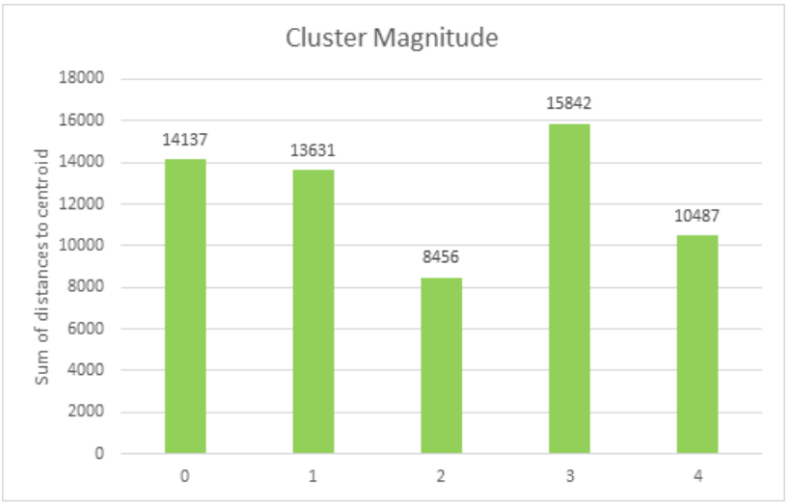
Figure 4: Cluster Cardinality



Source: Own Elaboration based on Jupyter Notebook.

The graph in Figure 4 represents the number of individual records (clients) in each of the cluster. Looking at the cardinality, cluster 0 is made up of 19,201 individuals, cluster 1 of 12,118, cluster 2 of 9 689, cluster 3 of 23,728, and cluster 4 of 8,495. There are some notable differences between the size of clusters and it is possible to identify two predominant clusters with bigger sizes and three smaller ones with relatively similar sizes between them.

Figure 5: Cluster Magnitude



Source: Own Elaboration based on Jupyter Notebook.

The magnitude of a cluster (Figure 5) represents the sum of the distances of each point to the center (centroid) of the cluster. This is a measure related to the differences between the individuals of each segment. Clusters with higher magnitudes are more dispersed and their individuals are more different between in each other, while in a cluster with a lower magnitude, the data points are closer together meaning they are more similar. It is important to notice that the bigger the size of the cluster, the bigger the magnitude as it is a sum. Some observations about the magnitude of each cluster:

- Clusters 0, 1, and 3 as the ones with high magnitude. This is in accordance with the size, for 0 and 3. However, cluster 1 is smaller in size, meaning that their individuals are significantly different, despite being in the same cluster (as seen in the silhouette graph, cluster 1 is the one with some of the lowest silhouette scores for some individuals);
- Clusters 2 and 4 have magnitudes of 8,456 and 10,487 meaning that their individuals are similar and there are fewer outliers present (very different individuals).

The following results (averages and percentages) for each variable per cluster were obtained:

Table 2: Characteristics of the Clusters

Characteristic	Feature	Cluster				
		0	1	2	3	4
Demographic	Age	47.2	46.7	51.0	48.6	50.9
Geographic	Asia	0%	2.4%	5.4%	0%	24.0%
	Europe	100%	85.3%	82.2%	100%	0%
	South America	0%	02.6%	3.8%	0%	25.2%
	North America	0%	7.2%	7.3%	0%	38.3%
	Other Continent	0%	2.4%	1.4%	0%	12.5%
	Non-Portuguese	90.9%	79.8%	94.2%	88.5%	100%
Psychographic	Average Lead Time in Days	90	59	121	92	81
	Average Daily Spend	80.77€	93.54€	78.82€	81.04€	82.49€
	Total Revenue	494.36€	490.68€	476.79€	444.32€	446.51€
Behavioral	Days Since Creation	702	645	647	724	643
	Room Nights	3.31	3.02	3.01	3.19	3.04
	Distribution Channel: Corporate	0.9%	13.0%	1.4%	0%	3.3%
	Distribution Channel: Travel Agent / Operator	98.7%	0%	98.6%	100%	95.6%
	Distribution Channel: Direct	0%	83.5%	0%	0%	0%
	Distribution Channel: GDS	0.3%	3.5%	0.1%	0%	1.0%
	Number of Requests	1.3	0.4	1.2	0.1	0.7
	Request: High Floor	6.6%	4.0%	4.6%	1.9%	4.3%
	Request: Crib	1.2%	1.8%	0.2%	0.8%	1.4%
	Request: Twin Bed	0.6%	6.9%	100%	0%	0.4%
	Request: Quiet Room	14.6%	3.5%	9.8%	6.5%	7.9%
	Request: King Size Bed	100%	17.7%	0.01%	0%	55.1%

Source: Own Elaboration based on Jupyter Notebook.

4. BUSINESS IMPACT

4.1. RESULTS INTERPRETATION

After assessing the quality of the clusters, it is necessary to identify the different business segments represented by them and their differences. To understand the results in the table 2 the reader should take into account that the non-percentage results are an average of all the clients inside the cluster.

The first segment represented in the first column of the table 2 is made up of European customers, with a significant percentage of non-Portuguese and ages being on average 47 years old. These customers have the highest amount of total revenue paid and also the highest room/nights the customer stayed in the hotel. This means they probably have longer stayed and are accompanied by more people. These clients made their bookings through travel agents. They have the highest number of requests being the most recurrent ones having the room on a higher floor, a quiet room, and a king-size bed. These customers are loyal, most likely families and couples coming to spend a vacation and represent the highest source of income for the hotel. This is also corroborated by the fact that they book their stays somewhat in advance compared to other segments.

The second group, as seen before, is the one with the biggest amount of difference when taking into account its small size compared to other client segments. These customers have the lowest age average and are mostly Europeans (with some being from other continents) and with the most Portuguese people since the foreigner rate is the lowest. These customers also represent a big portion of the hotel's revenue and have the highest daily expenses. This is following the fact that they are corporate customers, who probably order room service and enjoy other hotel amenities during their mostly short stays. These clients also book through GDS systems and directly with the hotel, which aligns with their corporate profile.

In contrast, this cluster has a very distinct characteristic which is the fact that it has the most amount for requests for cribs. This means that there are also parents with small children present in this group. These might make reservations directly with the hotel and justify why the age is the smallest of all. In general, this segment represents the corporate segment, with people staying for shorter periods while still spending a significant amount in the hotel.

The third segment represents the oldest people, with an average of 51 years old mostly European and foreigners. The bookings are made in advance, but the stays are shorter and with fewer people compared to other groups. These bookings are made with travel operators and with a considerable number of requests, all choosing twin beds. There is also the least number of requests for cribs meaning that there are very few children and a high number of requests for quiet rooms. These clients have the lowest daily spent but still represent a significant amount of the hotel’s revenue. This cluster is made up of essentially older friends who visit the hotel together for vacation and have little expenses during their stay.

The fourth client segment is made up of all Europeans with an average age of around 48 years old. These customers are the oldest in the hotel, however, they represent the lowest amount of revenue for the hotel. Their room/night value is one of the highest. This might mean that these people made very few bookings in the hotel, for longer stays, all through travel agencies. They have the least number of requests, which is explained by the fact that they probably are not regular clients and have only stayed a few times in the hotel. These are not the most important clients for the hotel. Although they are the easiest to please as they have few requests, it is important to investigate why they do not provide significant revenue for the hotel.

Finally, there is the fifth customer segment, which represents the most recent customers, with also one of the highest averages for age, and are all from countries outside of Europe. The bookings are made with travel agencies, and they ask for some special requests such as king-size beds and quiet

rooms. These are a recent addition to the customer portfolio; however, the hotel should invest in this segment as it can become a high source of income and lead to expanding the chain around the world.

4.2. RESULTS EVALUATION

The client segments obtained, allowed for the identification of numerous differences and the understanding of the independent profiles present in the Hotel's clientele. The model took into account people's demographic and geographic characteristics, as well as their behavioral characteristics. The development of an informed targeted marketing strategy will be addressed in the following chapter. With the information previously identified, it is possible to create a marketing strategy that includes more tailored services for each profile and address possible problems previously made. As a result, it is possible to conclude that the business goals were achieved.

In relation to the data mining aspects of this proposal, the clusters obtained are size significant, as they all represent considerable segments, and the silhouette score achieved was 0,26 which can be considered satisfactory.

With the implementation of the marketing strategies later defined, Hotel H will be able to keep the current customers while also attracting new ones, while also creating a competitive advantage in the market and increasing its revenue. Hotel H should continue to provide information to this model about the clients and the Marketing team should keep in mind the suggestions given in the following chapter. In the long term, Hotel H should look at the changes consequent to the implementation of these measures and review key indicators such as the number of bookings and ratings (and more) to see if they improved as a consequence of this project.

Overall, the results obtained are significant and can be an interesting source of information and guidance for the Marketing manager and eventually for an analytics team. The clusters highlight important details that can be explored, and it was possible to also identify some necessities in terms of needs for the future. These will be addressed in the conclusions and recommendations segment.

4.3. BUSINESS APPLICATIONS

As Hotel H continues its journey toward enhancing customer satisfaction and maximizing revenue opportunities, the findings from the customer segmentation analysis serve as a roadmap for targeted marketing strategies tailored specifically to its diverse clientele.

Segment 1: For Hotel H, European vacationers represent a cornerstone of its customer base, contributing significantly to revenue generation and brand loyalty. By understanding the preferences and behaviors of this segment, Hotel H can deploy targeted marketing initiatives such as loyalty programs, personalized offers, and family-friendly amenities. Strengthening relationships with European travel agents and emphasizing the hotel's appeal to families and couples on vacation will further solidify Hotel H's position as a preferred destination for European vacationers.

Segment 2: With an emphasis on convenience and business amenities, corporate travelers play a key role in Hotel H's revenue stream. By offering tailored business packages, conference facilities, and streamlined booking experiences, Hotel H can effectively cater to the needs of this segment. Special incentives such as corporate discounts and express services will enhance the hotel's attractiveness to corporate clients, fostering long-term partnerships and repeat business.

Segment 3: The mature vacationer segment presents Hotel H with an opportunity to appeal to older travelers seeking relaxation and comfort. By highlighting tranquil accommodations, senior-friendly amenities, and leisure activities tailored to their preferences, Hotel H can capture the loyalty of this demographic. Collaborating with travel agencies specializing in senior travel will enable Hotel H to reach this audience effectively and position itself as a preferred destination for mature vacationers seeking comfort and tranquility.

Segment 4: Hotel H's occasional European visitors represent a segment with untapped potential for repeat business and engagement. By incentivizing repeat visits through loyalty rewards, exclusive discounts, and personalized recommendations, Hotel H can nurture long-term relationships with this demographic. Strengthening partnerships with European travel agencies will expand the hotel's reach within this segment, driving sustained growth and revenue.

Segment 5: As Hotel H welcomes new international guests, it has the opportunity to showcase its global appeal and hospitality. Through targeted advertising campaigns on international travel platforms and personalized welcome packages, Hotel H can cultivate relationships with new clientele and foster brand loyalty. By leveraging its multicultural experiences and tailored services, Hotel H can position itself as a preferred destination for international travelers seeking authentic and memorable experiences.

Incorporating these targeted marketing strategies based on the insights gleaned from customer segmentation analysis, Hotel H can optimize resource allocation, enhance customer satisfaction, and solidify its position as a leading hospitality provider in Lisbon. Continuous evaluation and adaptation of these strategies will be essential to Hotel H's success in meeting the evolving needs and preferences of its diverse clientele.

5. DEPLOYMENT AND MAINTENANCE PLANS

The successful deployment and maintenance of the customer segmentation model are essential for maximizing business impact and ensuring its long-term effectiveness. To deploy the model into production for Hotel H, it is crucial to involve various stakeholders and ensure seamless integration with existing systems. The following steps outline the deployment strategy.

Collaboration with the IT department is vital to integrate the segmentation model into Hotel H’s existing systems, including the reservation management system and customer relationship management system. It is imperative to ensure that the model outputs, i.e. customer segments, are readily accessible to relevant staff members for decision-making purposes. Moreover, providing training sessions for hotel staff, especially those involved in marketing and customer relations, is crucial. This training aims to familiarize them with the segmentation model and its implications, educating them on how to interpret customer segments and leverage them effectively in marketing, product, and pricing strategies as well as customer interactions. To further ensure smooth operations and effective utilization of the segmentation model, comprehensive documentation is essential. This documentation should outline the model's functionality, inputs, outputs, and usage guidelines, serving as a reference point for future endeavors and training initiatives. It should include detailed step-by-step instructions, practical examples, and frequently asked questions to facilitate user understanding

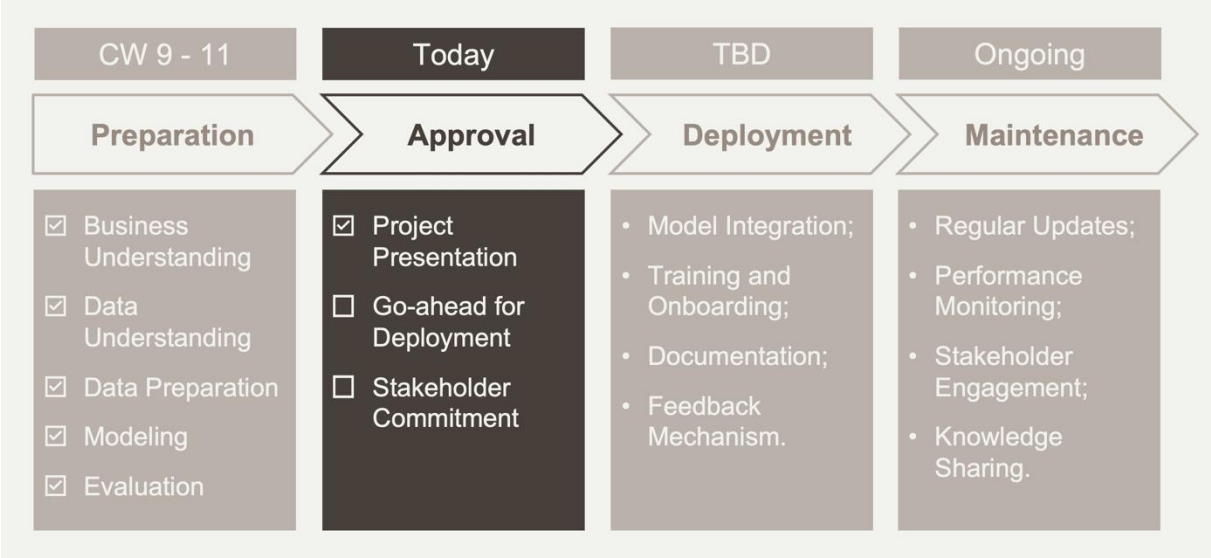
and adoption of the segmentation model. In addition to documentation, establishing a structured feedback mechanism is crucial for ongoing improvement and optimization of the model. This mechanism enables users to provide valuable insights regarding the model's performance, usability, and impact on business outcomes.

Maintaining the segmentation model involves ongoing monitoring, updates, and adaptations to ensure its relevance and effectiveness. The following measures support the model post-deployment.

Regular updates to the segmentation model should be scheduled to incorporate new data and insights gathered over time. Conducting model retraining using the latest available data ensures that it remains accurate and reflective of evolving customer behaviors and preferences. Another essential step is monitoring key performance indicators related to customer segmentation. This includes tracking the distribution of customers across segments, segment characteristics, and marketing campaign effectiveness. Setting up automated alerts to flag any anomalies or deviations from expected outcomes is crucial for prompt intervention and corrective actions. Furthermore, staying attuned to changes in the hotel industry landscape, market dynamics, and customer preferences is essential. Regularly reviewing and updating the segmentation model to accommodate shifting business priorities, emerging trends, and new market opportunities is necessary for its continued relevance and effectiveness. Critical for the model’s acceptance is also fostering ongoing collaboration and communication among stakeholders, including the marketing team, IT department, and senior management. Encouraging active participation in the maintenance process and soliciting feedback and insights from stakeholders to inform model refinements and improvements are vital for ongoing success. Lastly, maintaining comprehensive documentation of the segmentation model, including its algorithms, assumptions, and methodologies, is essential for knowledge sharing and continuity. Facilitating knowledge sharing among team members ensures institutional memory regarding the model's development, deployment, and maintenance.

In order to ensure a systematic deployment and maintenance plan for the segmentation model, a well-defined timeline is essential. Below is a brief overview of the proposed schedule.

Figure 6: Project Plan



Source: Own Elaboration.

By following these deployment and maintenance plans, Hotel H can effectively leverage the customer segmentation model to inform strategic decision-making, enhance marketing initiatives, and drive customer engagement and satisfaction.

6. CONCLUSIONS

Hotel H started by having an incomplete customer segmentation model, by only considering one of many possible characteristics. This new, presented model enabled identification of relevant customer groups, based on multiple characteristics while also personalizing marketing strategies, in order to improve client satisfaction, loyalty, and overall business performance.

Throughout the development of this model, the team faced some challenges, particularly in the data preparation that were solved and not impediment to the extraction of important information. The segmentation model successfully identified diverse customer groups, each with distinct behaviors and needs. These insights will allow Hotel H to tailor its services, ensuring every guest's satisfaction, which are, more importantly, European vacationers, corporate travelers, mature vacationers, and international guests.

Due to the nature of the hospitality industry, paired with the ever-growing customer expectations, Hotel H is advised to carefully monitor the segmentation model. There should be a focus on incorporating new data, market trends, and feedback from guests and staff. Maintaining a constant flow of information from the clients, while updating regularly marketing strategies is imperative to promote business growth and attract new customer relationships.

There are still more improvements to be made to the database as well as the model itself, however, the Hotel should primarily focus on implementing the recommended marketing strategies and prioritizing the different important aspects of the deployment phase as it is the most important phase in the short term. After assessing the impacts of this first trial, some further explorations into data analytics can be made. The hotel is equipped to make data-driven decisions about where to allocate resources to achieve the greatest impact and should keep this focus for the future.

In conclusion, there is significant potential in the future for Hotel H. By continuously harnessing the power of data and analytics, the hotel can not only enhance its operational efficiency and guest satisfaction but also explore new growth and innovation. The proposed project reflects only a small part of what can lead to a global hotel chain driven by customization but always prioritizing the client as they represent the hotel's foundation.

8.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

The following list contains some recommendations in order to improve the model created and optimize analytic performance for Hotel H:

- Remove transactions or other necessary corrections from the same dataset as these do not represent bookings or clients;
- Separate non-checked-in customers as the hotel cannot be certain about the information provided by these clients;

- Connect the parents with the children through a key so that the children are associated with their respective parents and requests made by them;
- To improve the demographic aspect of the segmentation, variables related to gender, who the person is accompanied by (family or friends), and others;
- It could be beneficial to collect information about the amenities used by the clients during their stay such as pool, spa, room service, gym, restaurant, etc;
- Add reviews/satisfaction scores to include the customer's perspective of the experience;
- Specify if the customer is part of the loyalty program;
- Keep track of the number of bookings made by the customer;
- Lastly, test different clustering algorithms more suited for mixed datatypes.

Disclaimer: A full understanding of data privacy issues is necessary to be able to implement these new changes, keeping in mind that the protection of the client data and their privacy is of the utmost importance.

7. REFERENCES

1. Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using Silhouette score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. https://doi.org/10.1109/dsaa49011.2020.00096

2. Zhao, Z., Wang, J., & Liu, Y. (2017). User Electricity Behavior Analysis Based on K-Means Plus Clustering Algorithm. Proceedings of the 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), 484-487. https://doi.org/10.1109/ICCTEC.2017.00111

8. APENDIX

The glossary below ensures clarity for the stakeholders and enhances understanding of key technical terms used throughout this report. These concepts are the core of the analytical approach used.

Table 3: Glossary

Term	Explanation
K-means Clustering	K-means is a popular unsupervised machine learning algorithm. It divides data into a specified number of groups, k, assigning each point to the group whose center is closest.
Elbow Method	Technique used in determining the optimal number of clusters in a dataset. It plots the explained variance as a function of the number of clusters and identifies the point where the rate of decrease strongly changes, the “elbow”.
Silhouette Score	Metric used to evaluate the quality of clusters created by a clustering algorithm. It ranges from -1 to 1, calculating how similar a point is to its own cluster compared to other clusters. A high value indicates that the point fits well within its own cluster and poorly matches its neighbors’ clusters.
Principal Component Analysis	Dimensionality reduction technique that reduces the complexity in high-dimensional data by converting it into fewer dimensions that still contain most of the information.

Source: Own Elaboration.