MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel Customer Segmentation

Diogo, Marquês, number: 20230486
Diogo, Pimenta, number: 20230498
João, Lopes, number: 20230748
João, Maia, number: 20230746
Pedro, Ventura, number: 20230728

Group K

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

This report presents the findings of a comprehensive analysis conducted on customer segmentation within the hospitality industry. The analysis utilized clustering techniques to categorize customers into distinct groups based on their booking behaviors, demographics, and spending patterns.

The report begins with an overview of the initial setup and data understanding phases, outlining the steps taken to import, clean, and prepare the data for analysis. This includes identifying inconsistencies, handling duplicates and missing values, removing outliers, and performing feature engineering.

Next, the clustering process is detailed, starting with principal component analysis (PCA) for dimensionality reduction. The report then describes the methodology used to determine the optimal number of clusters, assess cluster quality, and interpret the clustering results.

Based on the clustering analysis, five distinct customer segments were identified:

1. **Well Planned Spenders**: This segment represents customers who exhibit responsible booking behavior, plan their stays well in advance, and prefer higher spending options.
2. **Value-Conscious Last-Minute Travelers**: Customers in this segment prioritize last-minute bookings and are value-conscious, seeking deals and discounts.
3. **Well Planned Big Spenders:** Like the first segment, these customers plan well in advance but display different preferences in terms of nationality and revenue generation.
4. **Low spenders with no planning:** This segment contain customers who prefer last-minute bookings, exhibit lower spending patterns, and demonstrate varied geographic distribution.
5. **Meticulous Long-Term Patrons**: These customers are characterized by long-term planning, higher spending, and a preference for personalized experiences.

Finally, the report concludes with tailored marketing recommendations for each customer segment to help hospitality businesses better target their offerings and enhance customer satisfaction.

Overall, this analysis provides valuable insights into customer segmentation within the hospitality industry, enabling businesses to optimize their marketing strategies and improve customer experiences.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1 Background

Hotel H, located in Lisbon, Portugal, is a member of the independent hotel chain C. In 2018 the hotel chain created a marketing department and hired a new marketing manager, A, for the hotel H. Hotel H currently employs a customer segmentation approach based on the sales origin of the customer. This strategy categorizes customers only by one characteristic, sales origin. However, the new marketing manager has recognized that the current segmentation method may not provide sufficient insights for the hotel's marketing department. This segmentation method failed to capture important customer characteristics. Instead, for a better customer segmentation it should include the geographic origin of the customers, demographic characteristics (e.g., age), and behavioral attributes (e.g., number of stays).

The recognition of the limitations of the current segmentation strategy highlights the need for Hotel H to implement a more advanced and comprehensive segmentation plan. Incorporating various customer attributes, such as behavioral, demographic, and geographic aspects, allows the hotel to better understand its clientele and better target its marketing campaigns.

2.2 Business Objectives

To enhance the marketing team's understanding and effectively target different customer groups, an exploration of the hotel's data is essential. By doing this our main goal is to identify and understand the various needs and preferences of Hotel H´s customers. Through various aspects of the data such as customer behavior, purchasing patterns, demographic attributes, and other pertinent factors, it becomes possible to identify underlying patterns and similarities. These insights help the aggregation of customers into distinct segments based on shared characteristics.

Through customer segmentation, the marketing team can develop their strategies to serve the unique needs and preferences of each segment. This may involve tailoring advertising campaigns, optimizing distribution channels, and making targeted promotions. By aligning marketing efforts with the specific traits and behaviors of each customer segment, the hotel can:

- Enhance customer engagement,
- Improve customer loyalty,
- Improve business performance.

2.3 Business Success criteria

Increased Customer Satisfaction: Evaluating whether the updated segmentation strategy leads to enhanced customer satisfaction levels. This can be measured by analyzing feedback channels, like surveys and reviews before and after the implementation of the new strategy.

Higher Revenue and Profitability: Measuring the impact of the improved segmentation strategy on revenue generation and profitability. Control changes in critical metrics such as average revenue per customer, customer lifetime value, and profit margins to assess the financial efficacy of the project.

Enhanced Customer Retention: Analyzing how well the segmentation plan is working to keep and grow your consumer base. Study the loss of customers and retention rates in various market segments to determine whether the new strategy leads to improved customer retention results.

Optimized Marketing: Assessing the feedback of marketing activities conducted under the new segmentation strategy. To find opportunities for optimization and improvement, compare the efficiency and cost-effectiveness of marketing efforts aimed at various customer segments.

Alignment with Business Objectives: Ensuring that the project's results align with the Hotel's business objectives and strategic priorities. Analyze how much the improved segmentation approach helps the company achieve important goals like revenue growth, brand positioning, and improved customer experience.

2.4 Situation assessment

To facilitate effective communication within our project team, it's essential to familiarize ourselves with key terms related to the hospitality industry. Acquiring a comprehensive understanding of this industry will enhance clarity in our discussions, collaborations, and precision in analyzing the hotel´s data.

Regarding the data, it was given to us a csv file by the hotel with information about the bookings that the customers made until 2018. The dataset contains 28 features and 111733 rows or customers. The dataset includes several key variables that are essential in understanding and categorizing our clients effectively. "*Nationality*" provides valuable insights into cultural preferences and travel habits, which can significantly influence marketing strategies and service offerings tailored to specific demographics. "*Age*" serves as a fundamental demographic variable, allowing us to discern distinct preferences and behaviors among different age groups, guiding personalized marketing campaigns suited to different age groups. "*DaysSinceCreation*" and "*AverageLeadTime*" offer critical temporal perspectives, indicating the duration since a customer's first interaction with the hotel and the average time between booking and arrival. These metrics inform strategic decisions regarding customer engagement, loyalty programs, and promotional timing. "*DistributionChannel*" highlights the platforms through which customers discover and engage with our hotel, guiding marketing allocation and channel optimization efforts. Lastly, "*LodgingRevenue*" serves as a direct indicator of customer spending patterns, enabling segmentation based on profitability and guiding revenue management strategies to maximize profitability. By analyzing these variables effectively and many others in the dataset, we can create customer segments tailored to diverse preferences and behaviors, enhancing the efficacy of our marketing campaigns, and improving overall customer satisfaction and retention.

Throughout the project we will face a few obstacles and constraints. To counter these risks, we will do a few steps to minimize them:

Data inconsistencies – We assume that there will be inconsistencies in the provided dataset which could arise from various sources such as human error during data entry, differing recording practices, or system-related issues. By anticipating these issues, we can verify the existence of them and appropriately rectify them. After that we ensured that the data was clean of inconsistencies before beginning our analysis.

Time constraints - Since we have a large dataset it is expected that we might face some time constraints when running clustering algorithms. To minimize this, we need to consider the number of data points, the dataset size, and the clustering method's complexity.

2.5 Determine Data Mining goals

The project aims to facilitate the identification of different customer segmentations in the most effective approach. This involves utilizing data mining techniques to uncover hidden patterns and preferences within the dataset:

Feature engineering – Creating new features or modifying existing ones to add more information to that provided by the dataset. These engineered features may also increase our outcomes throughout the clustering phase.

Clustering to identify customer segments - Clustering will be a robust technique to effectively identify distinct customer segments within the hotel's customer base. The goal is to have a well separated and homogenous cluster to get a good visualization of the different segmentations. Once the clusters have been identified, we will profile each segment by analyzing the typical traits, preferences, and behaviors of customers within that segment. To determine what distinguishes one segment apart from another, we will examine the mean values of relevant features within each cluster.

Visualization – Since we have high–dimensional data, we must transform this into a lower–dimensional space with Principal Component Analysis (PCA). PCA will help to uncover patterns and relationships between data points that may not be visible in the original high-dimensional space.

Cardinality and Magnitude – The evaluation of our clustering results will be evaluated through the correlation of the Cardinality (number of data points in each cluster) and Magnitude (average distance to the centroids of each cluster). If there is a positive correlation, it indicates that clusters with higher magnitudes tend to have higher cardinalities, which suggests a more homogeneous distribution of data points within the clusters. On the other hand, a negative correlation may indicate anomalies in the clustering so it could be needed to revisit the preprocessing or adjusting the clustering parameters to improve the correlation.

3. METHODOLOGY

In order to present a solution for this case, we followed the CRISP-DM methodology. Accordingly, the work is divided into six parts. Firstly, the Business Understanding phase aims to articulate the problem under study. In the Data Understanding phase, we analyze and comprehend the provided data. Moving on to Data Preparation, we process our data, eliminating inconsistencies and creating new features. The fourth stage involves Modeling, where models are applied to our data. Subsequently, Evaluation arises, which entails assessing the model previously prepared. Finally, the sixth and last phase is Deployment, where our model is applied and put into practice.

3.1 Data understanding

To better explore the dataset, we used descriptive statistics and data visualization tools (pandas profiling) and other manual methods.

First, all the packages needed were imported and the dataset was opened and set "*ID*" as index, there were 28 variables involving this dataset. The variables were the grouped in segments of numerical, binary, and categorical features.

After studying the variables, it is evident that all variables starting with the prefix "SR," which are binary variables, exhibit extreme imbalance in their values, except for the variables "*SRKingSizeBed*" and "*SRTwinBed*." These exceptions show a higher frequency of 1, indicating that clients tend to request these special options more frequently.

Regarding the "*Nationality*" variable, the information provided indicates that clients come from 199 countries, with a significant percentage having European nationalities. Therefore, our analysis will focus on understanding how clients from select European countries (with a higher percentage in the dataset), as well as other European and global nationalities, interact with the hotel. (Figure 1)

In the "*BookingsCheckedIn*" variable, a notable quantity of 0 is observed, suggesting that clients have never visited the hotel and only have an affiliation with it. This information is crucial for the analysis, considering the potential impact on other variables, given the absence of actual stays.

The remaining data exhibit normal values considering the representation of each variable.

3.2 DATA PREPARATION

3.2.1 Inconsistencies/Incoherences

As mentioned in the data understanding phase, the variable "*BookingsCheckedIn*" exhibits 33,198 rows with the value 0. Given that this value indicates that the client has never visited the hotel and considering that our study primarily focuses on the hotel and its visitors, we have opted to exclude these corresponding rows, furthermore the hotel can only confirm a client's data after the client has visited the hotel at least once.

Regarding the "*Age*" variable, there are instances where the client's age is below 18 years, suggesting that it is likely an adjacent room to the parents. Given the context of our analysis, we cannot include these clients in our study. Consequently, we have removed the respective rows. Additionally, we encountered negative *"Age"*, for which we assigned a *NaN* value. Subsequently, we applied the KNNImputer method to impute these values, including the NaN values already present in the dataset. This imputation is based on similarity between their data and that of other clients.

Regarding the variables "*NameHash*" and "*DocIDHash*," we identified duplicates in both, needing confirmation of whether they correspond to the same individual. To address this situation, we decided to create a variable named "*CustomerHash*" combining these two variables along with the "Nationality" variable for verification purposes.

This consolidation confirmed the presence of repeated clients, suggesting multiple visits to the hotel where new customer records were created instead of utilizing existing ones. To address this, we devised a function to aggregate these duplicate entries for each client into a singular entry. For the variables "*LodgingRevenue*", "*OtherRevenue*", "*BookingsCanceled*", "*BookingsNoShowed*", "*BookingsCheckedIn*", "*PersonsNights*", and "*RoomNights*", a sum of all the records was applied. The "*AverageLeadTime*" variable was computed as the mean value, while for all other variables, the hotel's latest recorded value for the respective client was used. The *KNNImputer* also successfully imputed values for the 13 *NaN* values in the variable "*AverageLeadTime*". Following these transformations, our dataset now contains 75,102 rows.

3.2.2 Outliers

To understand the spread of numerical data, we used boxplots (Figure 2) and histograms (Figure 3) to see the distribution and spot potential outliers for removal from the dataset. We decided to employ five outlier removal methods, namely manual filtering, Z-Score, IQR, and a combination of Z-Score and IQR, each one in conjunction with manual filtering. After analyzing the results, the IQR and Z-Score methods exhibited a high percentage of data removal and proved to be ineffective for the variables "*BookingsCanceled*" and "*BookingsNoShowed*". Given that these variables had most identical values, these methods tended to eliminate the different values belonging to the minority, making them impractical for our study. Following this analysis for these two methods and considering that the combination with manual removal also exhibited weaknesses and high removal quantities, we chose to go with manual outlier removal for all variables.

3.2.3 Feature engineering

In terms of feature engineering to enhance our analysis, we have devised new ratios derived from combinations of the original variables. These ratios hold the promise of offering deeper insights into our dataset, potentially providing more understanding of the underlying patterns, and facilitating more precise segmentation.

We created the following:

- We created the variable "*RevenuePerRoomNight*" which calculates the average revenue per night spent at the lodging establishment, using the sum of "*LodgingRevenue*" and "*OtherRevenue*" and dividing by" *RoomNights*".
- We retained the four most significant European countries, namely "*FRA*" (France), "*DEU*" (Germany), "*PRT*" (Portugal), and "*GBR*" (United Kingdom). The remaining countries were categorized into "*Rest of Europe*" and "*Rest of the World*" for analytical purposes. This classification was applied within the "*Continent*" variable, ensuring clarity and consistency in our data analysis approach.
- The *"CancelationOrNoShowRate"* variable represents the rate of booking cancellations or no-show incidents relative to the total number of bookings checked in. It is calculated by summing the number of bookings canceled and the number of bookings for which guests did not show up, and then dividing this sum by the total number of bookings checked-in by guests. This metric provides insights into the efficiency of booking management and customer attendance. A higher value indicates a higher proportion of cancellations or no-show incidents relative to the total number of bookings checked in.
- The "$TotalSR$" variable aggregates the presence of these room preferences or amenities across all the binary variables. It represents the total count of room features or amenities selected by guests during the booking process. These binary variables include: "*SRHighFloor*", "*SRLowFloor*", "*SRAccessibleRoom*", "*SRMediumFloor*", "*SRBathtub*", "*SRShower*", "*SRCrib*", "*SRKingSizeBed*", "*SRTwinBed*", "*SRNearElevator*", "*SRAwayFromElevator*", "*SRNoAlcoholInMiniBar*", "*SRQuietRoom*".

3.2.4 Binning and One Hot Enconding

For better data interpretation, we decided to partition certain variables using the binning method, followed by performing One Hot Encoding. Here are the modifications made during the binning process:

- Age: Ages were categorized into decades, starting from individuals under 20 years old (with the minimum age being 18 years), progressing through intervals of 10 years up to 60 years, with the final age group being individuals over 60 years old.
- DaysSinceCreation: Data was segmented by years, indicating how many years ago the client created their account at the hotel.
- AverageLeadTime: Bins were created for the first year: up to one month, between one and six months, between six months and one year, and a final bin representing an AverageLeadTime greater than one year.
- BookingsCanceled: Three segments were defined: clients who never canceled any booking, those who canceled between one to three times, and those who canceled more than three times.
- BookingsNoShowed: Three segments were defined: clients who appeared in all bookings (zero no-shows), those who did not show up once, and those who did not show up twice.
- BookingsCheckedIn: Three segments were defined: clients who checked in once, between two and six times, and more than six times.
- RoomNights, PersonNights, and RevenuePerRoomNights: These three variables were divided into quartiles (Q1, Q2, Q3, Q4).
- TotalSR: The special requests made by clients were divided by quantity: zero requests, one request, two requests, more than two requests.

3.2.5 Feature Selection

For the selection of variables for cluster creation, we began by dropping the "*CustomerHash*" variable from the dataset as it represents a variable with maximum cardinality, indicating that it would not contribute to the study of clients. The final selection of variables used for cluster creation was obtained through trial and error, combining variables that the group deemed relevant for the study. Proceeding to enumerate these variables:

Numeric and categorical variables that underwent One Hot Encoding: "Age", "AverageLeadTime", "Continent", "TotalSR", "DaysSinceCreation", "BookingsCanceled", "BookingsNoShowed", and "BookingsCheckedIn".

Variables that did not undergo One Hot Encoding: "LodgingRevenue", "OtherRevenue", "RoomNights", "RevenuePerRoomNights".

The variables "DistributionChannel_Corporate", "DistributionChannel_Direct", "DistributionChannel_GDS Systems", "DistributionChannel_Travel Agent/Operator" demonstrated poor segmentation for study purposes as they are highly imbalanced variables.

3.2.6 Feature Scaling

Since the k-means method operates based on distances between points, it is crucial that all our data have the same scale. To achieve this, we employ the MinMax Scaler to standardize our entire dataset.

3.2.7 PCA

To reduce the dimensionality of our dataset without losing important information for our analyses, we employed Principal Component Analysis (PCA). We chose to select 16 principal components that explain 96% of the variance in our dataset.

3.3 Modeling

Customer segmentation was performed using the K-means++ algorithm, which allowed for the division of data into distinct groups. To determine the optimal number of clusters for the algorithm execution, two models were employed: the Elbow method and the Davies-Bouldin Score. After analyzing the results of both models, it was concluded that the optimal number of clusters would be K=5.(Figure 4)

Following the execution of the K-means++ algorithm with five clusters, as previously determined, the clusters were visualized using colors to analyze their distribution. The clusters are well-distributed, with almost no color overlap between clusters (Figure 5)

3.4 Evaluation

To assess the quality of the clusters, we measured the cardinality and magnitude of each cluster, along with the regression between them.

As illustrated in Figure 6, the clusters exhibit satisfactory cardinality, indicating that there is an adequate amount of data in each cluster to perform analysis. In the analysis of magnitude (Figure 7), similar to cardinality, the clusters exhibit good magnitude, indicating that the points within the clusters are close together and have a high level of similarity between them.

A positive relationship is observed between cardinality and magnitude. This indicates that as the number of points in a cluster increases the total magnitude of the variables associated with that cluster also increases. Therefore, it can be concluded that the variables within the clusters are highly related to the clusters themselves. This strong association reinforces the internal consistency of the clusters and the reliability of the analysis results.(Figure 8)

4. RESULTS EVALUATION

With the clustering analysis, we identified five clusters. We will detail them by illustrating their size proportion, age distribution per cluster, average lead time per cluster, revenue, continent distribution, and days since creation.

Cluster 0 – Well Planned Spenders

This cluster represents 22.6% of the customers.
Exhibiting an even distribution across age groups, with a slight dominance of bookings in the Established Professionals (40-49 years old) and Approaching Retirement (50-59 years old) categories. Young Adults (20-29 years old) and Young Travelers (< 20 years old) have a lower presence in this cluster.
The dominance of Advance Bookers (1-6 months lead time) suggests that most customers within this segment plan their bookings well in advance. Last Minute Bookers (up to 1 month) and Long-Term Planners (6 months - 1 year) are absent, with a negligible presence of Very Long-Term Planners (over 1 year).
Responsible booking behavior is evident in this cluster, with minimal cancellations and no-shows.
Exhibiting a diverse mix of nationalities, with a significant presence from several European countries. France and Germany contribute the highest proportions of bookings within the cluster, with other European countries collectively forming a substantial portion.
This cluster shows an even distribution across booking creation times.
Likely representing a segment of customers with a higher spending propensity, characterized by either staying for more nights at the hotel or staying in more expensive rooms.
Unique compared to others; this cluster consists entirely of bookings with no special requests.

Cluster 1 - Value-Conscious Last-Minute Travelers

This cluster represents 13.6% of the customers.

Exhibiting a balanced distribution across age groups, with Established Professionals (40-49 years old) and Approaching Retirement (50-59 years old) segments holding notable presence, each comprising around 24% and 22%, respectively. Despite these groups dominating, the distribution remains even across other age groups.

Characterized by a strong preference for Last-Minute Bookings (up to 1 month lead time), with nearly all bookings (99.88%) falling within this category. Virtually absent are Advance Bookers (1-6 months) and Long-Term Planners (6 months to 1 year), suggesting uncommon behavior. Additionally, Very Long-Term Planners (over 1 year) have a negligible presence (0.12%) within this cluster.

Responsible booking behavior is a hallmark of this cluster, with exceptionally rare cancellations and no-shows, with nearly all bookings (99.9%) resulting in guests honoring their reservations.

Displaying a diverse guest base with a strong European presence, with France (15.11%) and Germany (6.78%) contributing the highest proportions of bookings within the cluster. Other European countries collectively account for a substantial share (31.55%).

This cluster exhibits an even distribution across booking creation times, with a sizable portion of bookings (around 54.34%) falling within the past 2 years, indicating a mix of recent and older bookings.

Notably, this cluster exhibits lower revenue generation compared to most other clusters, possibly due to booking rooms with lower rates or staying for less time. However, all bookings in this cluster share a unique characteristic of having one special request.

Cluster 2- Well Planned Biggest Spenders

This cluster represents 28.2% of clients.

Exhibiting a balanced distribution across age groups, the Approaching Retirement segment holds the highest representation within the cluster at 24.39%. It is closely followed by Established Professionals (21.35%) and Retirees (22.81%). Young Professionals (19.78%) and Young Adults (10.89%) also have a moderate presence, while Young Travelers (under 20 years old) are negligibly represented.

A characteristic of Advance Bookers, this cluster demonstrates a strong preference for booking in advance, with almost all bookings (99.92%) made 1-6 months before the stay.

Exceptionally low cancellation rates and absence of No-Shows align with responsible booking behavior observed within this cluster.

With over 81% (81.1979%) of bookings originating from European countries, this cluster exhibits a diverse source of bookings, displaying a strong European presence.

Bookings are evenly distributed across different periods, with a sizable portion (around 54.34%) falling within the past 2 years (combining the first two categories). Bookings created between 2 and 3 years ago (28.00%) and those created over 3 years ago (19.16%) also hold a notable presence.

Notably, this cluster stands out for its superior revenue generation compared to other clusters. Possible explanations include the absence of bookings falling under the "No Special Requests" category and a significant majority (almost 81.48%) of bookings having one special request. While a smaller portion has two or more special requests, their presence indicates that specific needs are prevalent within this cluster, contributing to the higher value generation.

Cluster 3 – Low spenders with no planning

This cluster accounts for approximately 21% of the client base.

It demonstrates a diverse age distribution. The highest proportion of customers falls within the 40-49 age group (28.95%), closely followed by the 50-59 age group (24.06%). Significant representation is also observed among customers aged 30-39 (22.86%). Those aged 20-29 constitute a smaller portion (9.44%), while individuals under 20 and over 60 are the least represented (3.94% and 14.30%, respectively).

Predominantly favoring last-minute bookings, nearly all bookings (around 99.80%) are made within a lead time of up to one month. Rarely do bookings occur more than a year in advance (0.20%), with no observed bookings between one to six months or six months to one year in advance.

The cluster displays responsible booking behavior, with a negligible cancellation rate (0.24%) and almost all bookings (99.98%) honored by guests without any no-shows. Most bookings (93.26%) result in guests checking in once, while a smaller proportion (6.66%) show guests checking in between two to six times. Multiple no-shows are extremely low (0.01%).

Geographically diverse, the cluster sees high representation from customers booking from Portugal (24.95%) and the Rest of Europe (29.81%). France contributes significantly (14.17%), followed by the Rest of the World (15.92%), while Germany (8.51%) and the United Kingdom (6.64%) also exhibit a presence, albeit to a lesser extent.

Account creation spans various time frames, with approximately 28.84% within the last 1-2 years and 28.84% within the 2-3 years bracket. Customers within the 0-1 year and 3+ years since account creation categories constitute around 23.48% and 20.29% of the cluster, respectively.

Compared to other clusters, this one shows lower values across revenue-related metrics, all denoted in euros. Lodging revenue stands at €339.04, with other revenue at €65.85. Consequently, revenue per room night is also lower at €364.14. Moreover, the number of room nights is considerably low, recorded at only 2.56, indicating lower overall revenue generation.

Regarding special requests, most bookings (84.30%) do not have any special requests. A smaller portion of bookings (14.42%) have one special request, while a small fraction (1.29%) have two or more special requests.

<u>Cluster 4 – Meticulous Long-Term Patrons</u>

This cluster represents 14.6% of the client base.

It is characterized by a predominant presence of older customers, with approximately 44.44% aged 60 and above, followed by 22.61% in the 50-59 age group. Customers in the 40-49 age range represent about 16.43% of the cluster, while those under 20 are the least represented at approximately 0.79%.

Primarily favoring long-term planning, most bookings (92.03%) are made between six months to one year in advance. Bookings made more than a year in advance also show some presence, accounting for about 7.97% of the cluster. However, there are no observed bookings made within one month or between one to six months in advance in this cluster.

This cluster demonstrates extremely responsible booking behavior, with virtually no cancellations (99.99%) and all bookings (100%) resulting in guests checking in as expected. Additionally, there are no instances of no-shows recorded in this cluster.

Exhibiting a diverse geographic distribution, Germany (31.49%) contributes the highest proportion of customers, followed by the rest of Europe (24.50%) and France (9.69%). Portugal also shows significant representation, accounting for approximately 6.97% of the cluster. The United Kingdom (15.31%) and the rest of the world (12.03%) make up smaller proportions within this cluster.

In terms of the time since account creation, this cluster displays a varied distribution. Approximately 32.65% of customers have accounts created within the last 2-3 years, followed closely by 25.30% within the 1-2 years bracket. Customers with accounts created within the last 0-1 year constitute around 24.24% of the cluster. Those with accounts older than 3 years but less than 4 years represent the smallest proportion, at approximately 17.81%.

Regarding revenue metrics, this cluster ranks third among all clusters in terms of lodging revenue and revenue per room nights. However, it stands out with the highest value in terms of other revenue among all clusters. Despite having a lower number of room nights than some of the higher value clusters, the higher revenue generated from other sources contributes to a notable overall revenue performance in this cluster.

Moreover, this cluster predominantly consists of bookings with one special request, accounting for approximately 51.33% of the cluster. Additionally, bookings without any special requests represent around 38.90%. There is a smaller proportion of bookings with two special requests, comprising about 8.72% of the cluster. Finally, bookings with three or more special requests are the least common, constituting only about 1.05%.

4.1. Marketing strategies for each cluster

Following our clustering analysis findings, we have crafted specific marketing recommendations for each cluster to better target their preferences and behaviors:

Cluster 0 – Well Planned Spenders:

Personalized Planning Assistance: Offer tailored recommendations and exclusive offers to assist customers in maximizing their advanced bookings.
Early Bird Discounts: Introduce incentives like early bird discounts or loyalty rewards to encourage continued advanced bookings.
Special Request Enhancements: Provide additional benefits or upgrades for customers making special requests to enhance their experience.

Cluster 1 - Value-Conscious Last-Minute Travelers:

Last-Minute Deals: Promote last-minute deals and flash sales to appeal to customers booking close to their stay dates.
Flexible Cancellation Policies: Emphasize flexible cancellation policies to alleviate concerns about last-minute changes.
Quick Booking Incentives: Offer incentives like loyalty points for customers making spontaneous bookings.

Cluster 2 - Well Planned Biggest Spenders:

Exclusive Booking Packages: Create tailored booking packages bundled with upgrades or vouchers.
Early Access Benefits: Provide early access to limited time offers for customers who book in advance.

Personalized Recommendations: Use data analytics to offer personalized booking suggestions based on past preferences.

Cluster 3 – Low spenders with no planning:

Value-Driven Packages: Offer affordable packages with bundled amenities to appeal to price-conscious customers.
Last-Minute Discounts: Provide discounts or promotions for spontaneous bookings.
Streamlined Booking Process: Simplify the booking process for quick and hassle-free reservations.

Cluster 4 – Meticulous Long-Term Patrons:

VIP Loyalty Programs: Implement VIP programs with exclusive benefits for long-term patrons.
Tailored Experiences: Curate personalized amenities and services to cater to specific preferences.
Extended Stay Discounts: Offer special incentives for extended stays to encourage longer bookings.

5. DEPLOYMENT AND MAINTENANCE PLANS

Deploying the model involves strategic planning, seamless implementation, rigorous testing, and careful monitoring. It begins with understanding its purpose and assessing compatibility with existing systems. Once integrated, thorough testing ensures reliability and accuracy. Deployment includes documentation and training for users. Ongoing monitoring and maintenance, along with security measures, ensure compliance and performance optimization. Assigning clear roles, gathering feedback, and continuous improvement are essential.

After deployment, continuous monitoring of the model's performance is essential. This involves tracking key metrics in real time and setting up alerts for anomalies. Regular maintenance tasks, including software updates and data retraining, ensure its effectiveness over time. Additionally, implementing a feedback loop allows for adjustments based on user insights, ensuring the model remains aligned with evolving business needs and objectives.

6. CONCLUSIONS

In conclusion, our project has provided valuable insights into the diverse segments of our customer base, each with unique characteristics and preferences. By identifying distinct clusters such as "Well Planned Spenders," "Value-Conscious Last-Minute Travelers," and "Meticulous Long-Term Patrons," we can tailor our marketing strategies to better target and serve each group.

Moving forward, deploying our data mining model involves strategic planning, seamless implementation, rigorous testing, and careful monitoring. This encompasses understanding its purpose, ensuring compatibility, thorough testing, and documentation. Ongoing monitoring and maintenance, coupled with security measures, will optimize performance and ensure compliance. By assigning clear roles, gathering feedback, and continuously improving, we can adapt to evolving business needs effectively.

After deployment, continuous monitoring is crucial, involving real-time tracking of key metrics and proactive anomaly detection. Regular maintenance tasks, including updates and retraining, will sustain the model's effectiveness over time. Implementing a feedback loop ensures adjustments based on user insights, maintaining alignment with business objectives.

In summary, our project provides a comprehensive approach to understanding our customer base and deploying our data mining model effectively, ensuring sustained value delivery and informed decision-making for the organization.

6.2 CONSIDERATIONS FOR MODEL IMPROVEMENT

Firstly, we have identified that certain variables, such as the number of people per night and the total number of rooms per night of their stay in the hotel, could be presented in a more effective format. Additionally, we recognize the need for a more comprehensive approach to identifying the most important Special Requests that each group commonly makes.

Furthermore, there are still inconsistencies present in our data, particularly concerning outliers. While acknowledging these issues, we note that time constraints prevented us from thoroughly addressing them. However, for future research or implementation, addressing these inconsistencies should be a priority.

Moreover, we believe there is room for experimentation with different variables and combinations for clustering analysis. Exploring alternative approaches could yield more insightful results and enhance the analysis's effectiveness.

7. REFERENCES

sklearn.metrics.davies_bouldin_score. (n.d.). scikit-learn: Machine Learning in Python. Retrieved March 13, 2024), from Scikit-learn

Google Developers. (n.d.). Interpretar agrupamentos. Retrieved March 13, 2024, from Google Developers

8. APPENDIX

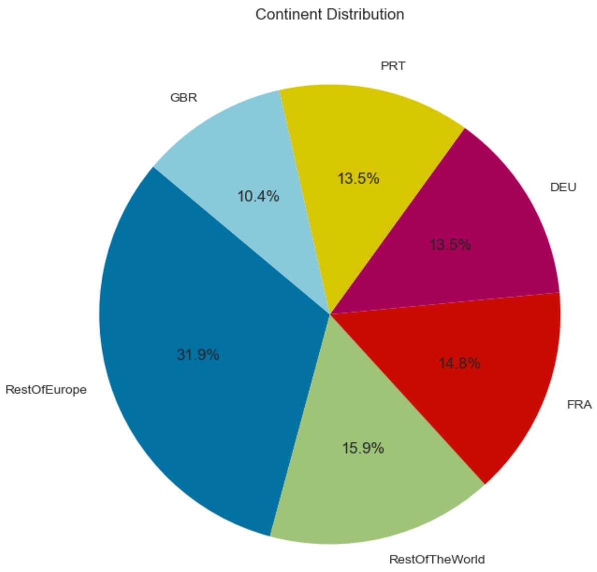

Figure 1 – Continent distribution

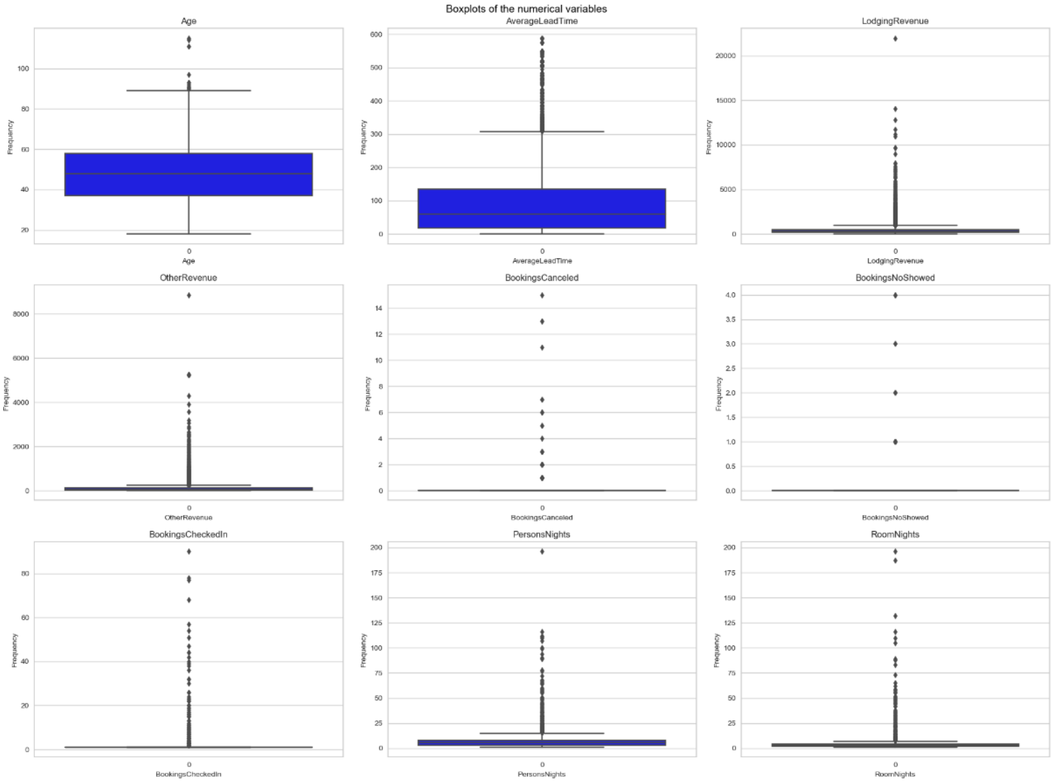

Figure 2 – Outliers Box Plots

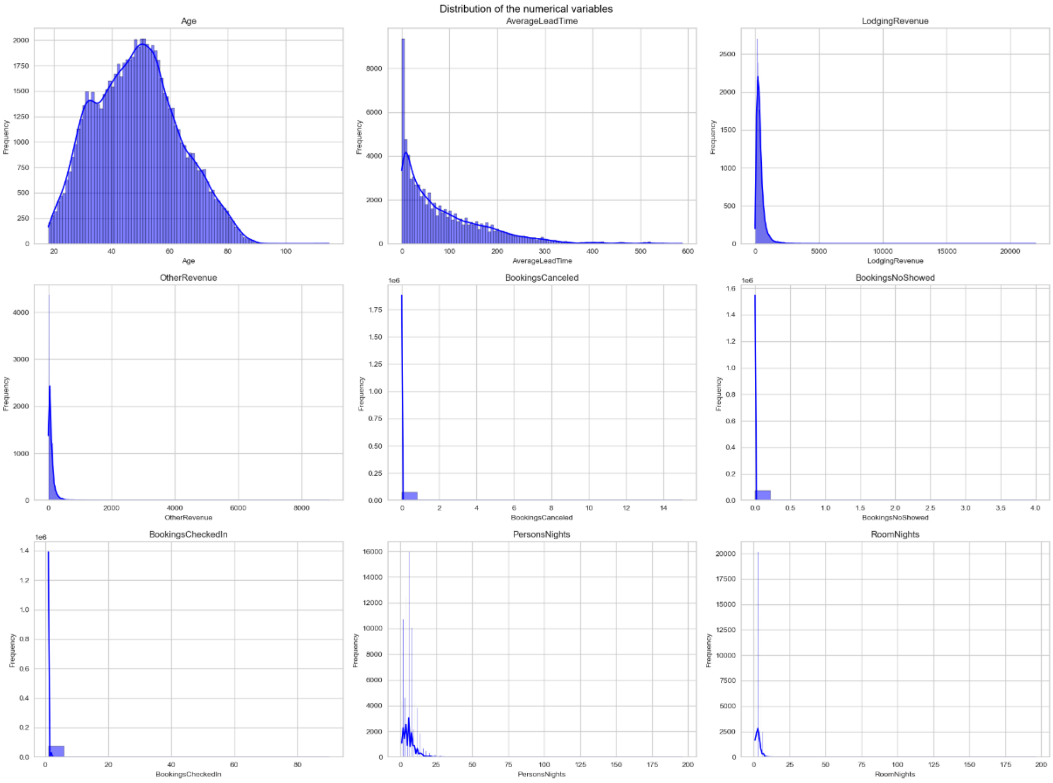

Figure 3 - Histograms

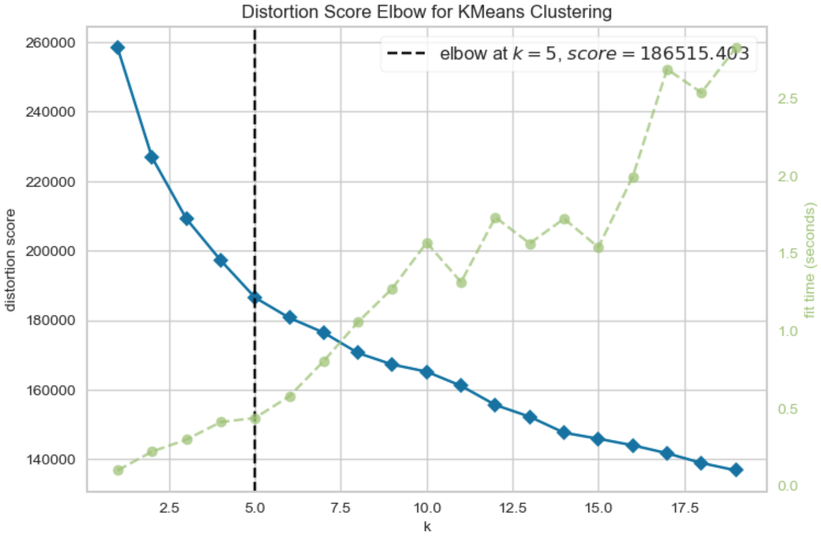

Figure 4 – Elbow Method

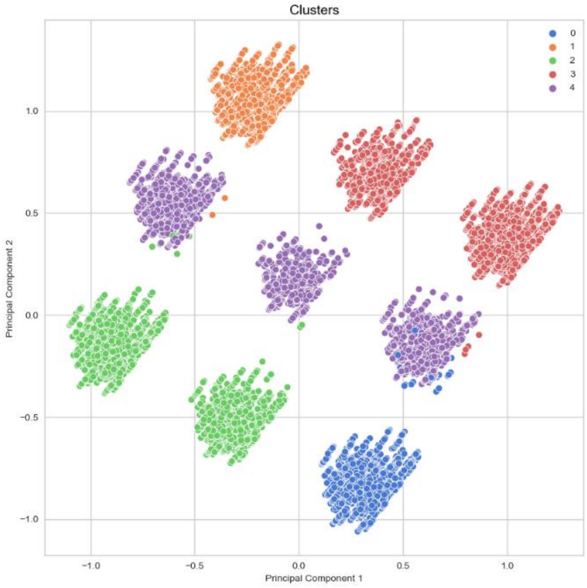

Figure 5 – K-Means++ algortihm

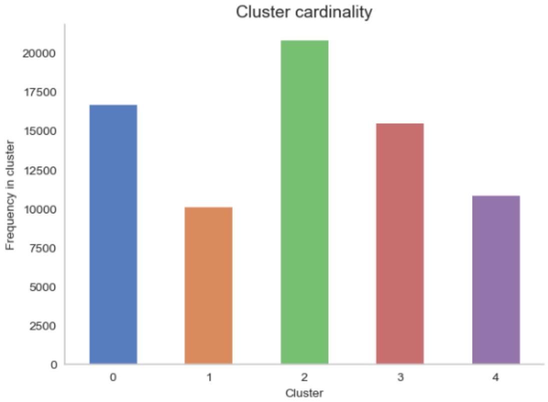

Figure 6 – Cluster cardinality

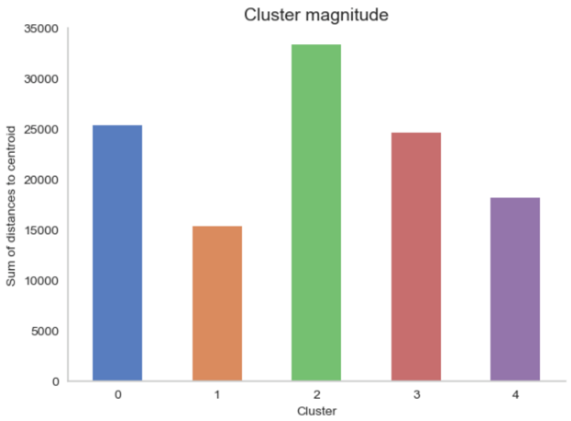

Fig 7 – Cluster magnitude

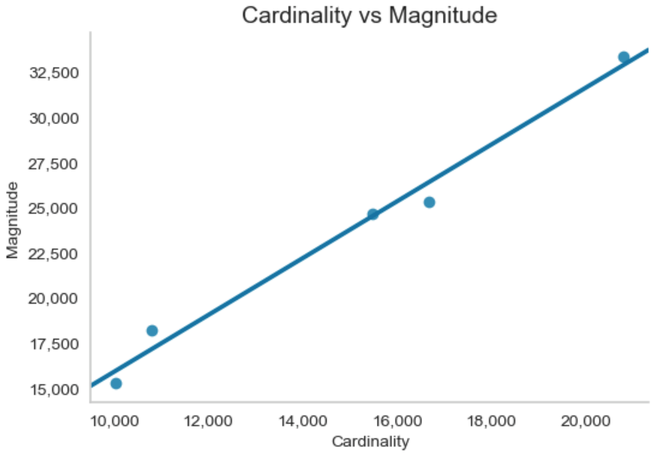

Figure 8 – Cardinality vs Magnitude