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

Student Number: 20230994 Pedro Adonis

Student Number: 20230742 Mohamed Taha Ben Attia

Student Number: 20230493 Márcia Lopes

Student Number: 20230474 Henrique Seganfredo

Student Number: 20231006 Deepak Lakhani

Group L

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. INTRODUCTION

Hotel H, located in Lisbon, Portugal, extends a warm welcome to visitors from around the world. As a proud member of chain C, the hotel is committed to delivering exceptional service. Previously, guest categorization relied solely on their place of origin. However, with A assuming the role of the new marketing manager, there's a recognition that this approach falls short. Aiming for a deeper understanding of guests, A acknowledges the need to transcend geographical boundaries. Given Lisbon's dynamic tourism landscape, Hotel H is in need of a fresh strategy to remain competitive. We'll examine Hotel H's approach to this problem, emphasizing the analysis of consumer behavior and visitor preferences.

Our objective is to assist Hotel H in offering distinctive experiences and establishing itself as a premier destination in Lisbon.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Until 2015, hotel chain C operated four hotels. However, with the acquisition of new properties, the board decided to increase investment in marketing. It wasn't until 2018 that the hotel chain established a marketing department and appointed a new marketing manager, A. A quickly realized that the existing customer segmentation was inadequate, as it only considered one characteristic: sales origin. It failed to account for geographic factors like country of origin, demographic aspects such as age, or behavioral traits like frequency of stays. Without proper segmentation, A found it challenging to develop strategies to attract new customers and retain existing ones.

2.2. BUSINESS OBJECTIVES

The main goal for profit should not forget serviceability and a warm welcome. Customers must feel at home and enjoy frictionless reservation and check-in procedures, providing them with all specific needs asked for in their appointment and following their profile. Second, we should reduce room non-occupancy through not only having good reviews from our customers, but also providing an intelligent pricing policy, which will directly lower any idleness and lower financial returns for our investors. We should also increase extra revenue sources from other hotel services such as food services and event catering.

2.3. BUSINESS SUCCESS CRITERIA

The success criteria for this project is achieving a more nuanced segmentation of the hotel's clientele, which is more reliable than mere regional classification. With a clearer understanding of client segments, marketing campaigns are anticipated to be more effective and can lead to an increase in operational revenue.

2.4. SITUATION ASSESSMENT

The hotel chain relies on a traditional booking system, available online for direct customer reservations but also for travel agents which will have more specific options tailored for their needs and allowing commission fees.

Data is stored on a relational database system with small IT teams providing custom changes on the reservation system, basic report design and other low level infrastructure operations in the on premises and cloud environment this system relies on.

There are a few caveats today on scalability issues and disaster recovery needs to achieve a higher service level. This could also be addressed on any new projects in the IT department. This will require some investment but will bring a more up-to-date standard for our current environment.

2.5. DETERMINE DATA MINING GOALS

The technical goals for this project are:

- An exploratory analysis that can provide a data understanding.
- A Preprocessing and feature engineering to handle correctly the outliers, create new features, drop the irrelevant ones and do a successful encoding and normalization for an effective modeling.
- Use the best algorithms for dimensionality reduction and clustering.

3. METHODOLOGY

We followed the CRISP-DM methodology, going back to previous steps whenever necessary, iterating through the various steps, in order to improve the final outcome.

3.1. DATA UNDERSTANDING

The dataset provided encompasses a comprehensive snapshot of customer transactions, comprising 111,733 entries across 29 columns. Each row encapsulates a distinct data point, shedding light on various aspects of customer interactions. Among the key variables, 'ID' serves as a unique identifier for individual customers, facilitating tracking and analysis. The 'Nationality' column provides insights into the diverse demographic backgrounds of patrons, crucial for understanding customer diversity and preferences. Additionally, 'Age' data unveils the age distribution of customers, aiding in targeted marketing strategies. 'DaysSinceCreation' offers a temporal perspective, delineating the duration since the inception of customer accounts, a vital metric for gauging customer loyalty and engagement. Moreover, attributes like 'AverageLeadTime', 'LodgingRevenue', and 'OtherRevenue' furnish valuable insights into booking patterns and revenue streams.

Summary Statistics:

The utilization of the 'describe' function yields a plethora of summary statistics, offering initial insights into customer behavior and revenue generation. For instance, the mean age of

approximately 45.64 years, coupled with a standard deviation of 17.24, shows that most of the customers are mature adults. Similarly, the average lead time for bookings, reflected by 'AverageLeadTime', provides crucial intelligence on booking trends and customer planning behavior, unveiling a tendency of bookings being made around 1 to 3 months before Check-In date. Furthermore, revenue-related metrics such as 'LodgingRevenue' and 'OtherRevenue' offer nuanced perspectives on revenue diversification and performance.

3.2. DATA PREPARATION

In the data preparation phase, our team took several measures to ensure the quality and relevance of the dataset for analysis. We began by addressing duplicate entries finding, and removing 112 instances, which cleansed the data of redundancies. We then focused on customers with incomplete profiles or no previous interactions with the hotel, specifically those who had not made any bookings and were only part of the hotel's loyalty program; these entries were excluded from further analysis.

We identified duplicate customer information based on 'Nationality', 'NameHash', and 'DocIDHash', and in these cases, consolidated the data to retain only a single, comprehensive record for each customer. This involved merging all related files, summing revenues and nights stayed, and recalculating average lead time where necessary.

The cleaning process also addressed nonsensical or illogical data entries. For instance, we found and corrected negative age values and adjusted average lead times that were mistakenly entered as -1, interpreting them instead as zero to indicate bookings made upon arrival. Additionally, we eliminated cases where the customer was under 18 years old, as being underaged doesn’t allow them to make or pay for their own reservations. Moreover, customers who were erroneously listed with multiple types of rooms or beds under a single booking were also eliminated scenarios that did not reflect plausible customer behavior.

To further refine the dataset, we set a unique identifier as the index of our data, enhancing the structure and accessibility of the records. The 'MarketSegment' column, deemed extraneous to our current analysis, was dropped. This left us with 25 features, categorized into 10 metric and 15 non-metric features.

Logical inconsistencies were also removed; for example, records showing fewer 'PersonsNights' than 'RoomNights' were excluded, as were records of clients who had never completed a booking. Such data points were considered irrelevant to our objectives and could potentially skew the results of the analysis.

The final dataset was carefully vetted to confirm the accuracy of customer preferences regarding rooms and beds. Records indicating improbable selections, such as booking more than one room or bed type, were removed. This thorough preparation ensured that the data was now primed for insightful analysis, with a robust structure supporting the integrity of our subsequent findings.

3.2.1. Missing Values

We identified and addressed missing or strange values in the dataset. A check for missing values revealed that the 'Age' column had 2,814 missing entries, which constituted approximately 3.76% of the data. To handle this, we filled the missing 'Age' values by calculating the mean age for each

'Nationality' group and applying this meaning to the respective missing entries. This approach is an adapted form of the nearest neighbors' algorithm, which instead of grouping the nearest rows it was decided it made more sense to group them by country of origin. This ensures that the age values are representative of the respective nationalities' average age.

After treating the missing values, we converted the 'Age' column to an integer data type to maintain consistency within the dataset. This step is crucial as it ensures the discrete values of age values.

3.2.2. Initial Feature Engineering

We refined the 'Nationality' feature by assessing the distribution of our customer base. The value counts indicated a vast range of nationalities, with the highest numbers coming from France, Germany, and Portugal. To simplify our analysis, we introduced a binary classification, designating a customer as a 'Foreigner' with a value of 1 if they were not from Portugal ('PRT') and 0 if they were. This feature was added to our non-metric features list.

Further enhancements to our dataset involved mapping nationalities to their respective continents, with a particular focus on Europe due to a higher customer volume from this region, as our dataset indicated. We established a 'Region' feature to reflect this geographical information more accurately. To do this, we renamed certain countries with small data representations into broader regional categories, ensuring a more streamlined analysis. For example, countries like Luxembourg, Cyprus, and Malta were reclassified under 'Western Europe' or 'Eastern Europe' as appropriate. Appart from that, certain countries with more clients stayed out of their certain region or continent. An example of that is Canada and USA that stayed out of America.

Upon completion of these updates, we observed in the figure provided that most of our customers hailed from European countries, with France, Germany, and Portugal being the most represented. The 'Region' column in our dataset now provides a clear image of our customers' geographical distribution as shown in Fig 1.

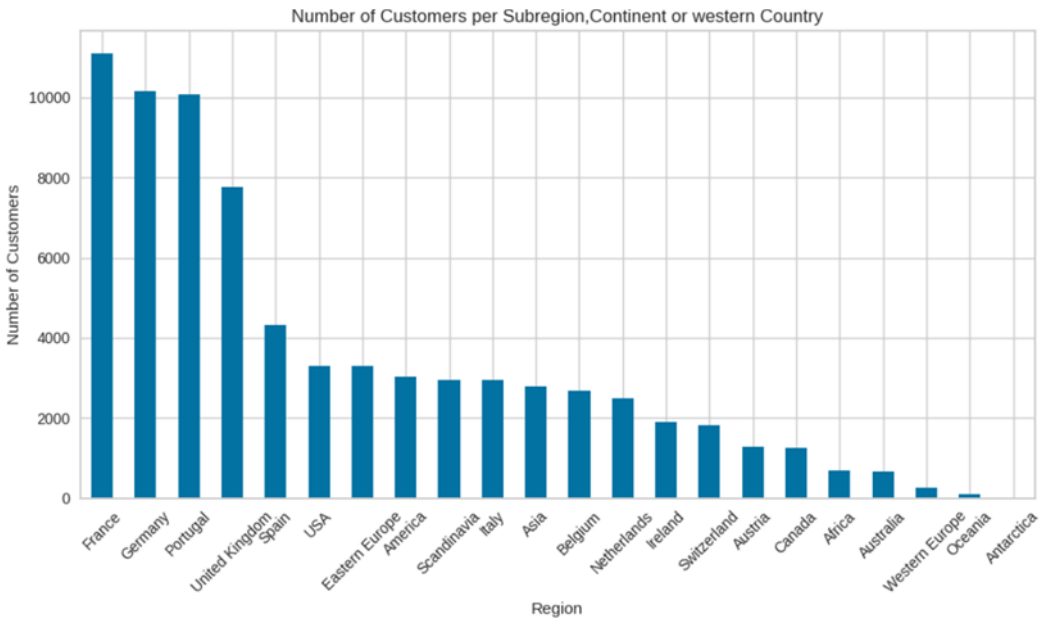

Fig 1

3.2.3. Data Visualization

We embarked on an exploration of the numerical variables within our dataset. We constructed pairwise relationships for these numerical features, opting for histograms along the diagonal for a clear distribution overview. This visualization aided in perceiving potential correlations, as seen between 'PersonsNights' and 'RoomNights' (Image 1), and distributions simultaneously, enhancing our understanding of the data's structure.

Further to this, we employed a boxplot and histogram pairing for each numerical variable to examine their central tendency and spread, alongside the identification of outliers. This dual approach of plotting allowed us to gain a deeper insight into each numerical feature independently which later was extremely important for choosing the approach to each feature outliers handling.

On the categorical side, we determined the absolute frequencies of non-metric variables through a series of bar plots. These plots were instrumental in revealing the most common category within the features related to customer usual preferences was the 0, the no preference (Image 2)

Moreover, to comprehend the interrelationships between our numerical variables, we calculated the Spearman correlation coefficients and depicted them in a heatmap. The heatmap illuminated significant correlations, such as a strong positive relationship between the number of room nights and the number of persons nights (0.86) indicating that an increase in room bookings typically corresponds to more people staying., as well as a moderate positive correlation between lodging revenue and other revenue (0.54). This suggests that as customers spend more on lodging, their expenditure on other hotel services or products tends to increase. Another notable relationship was the strong correlation between 'PersonsNights'/'RoomNights' and the revenue features, reinforcing the concept that more stays are associated with a higher spending. These correlations are shown in bolder orange in the heatmap below.

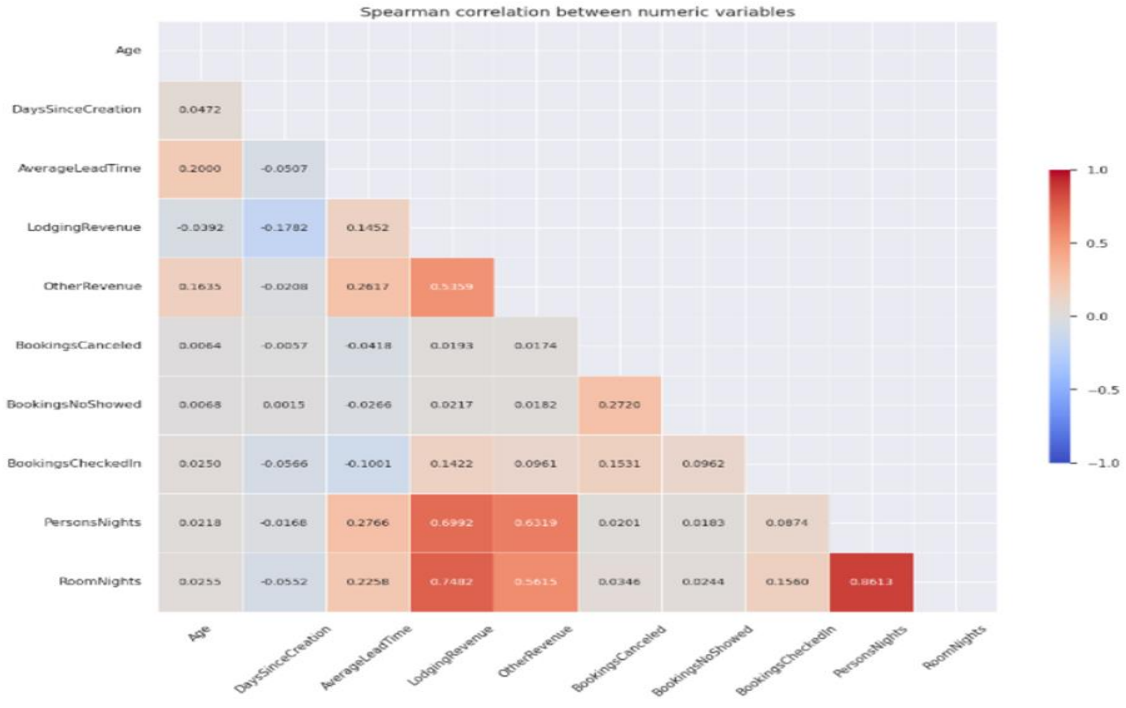

Fig 2

We also plotted a stacked bar plot of the features ‘Foreigner’ and Distribution Channel. Which showed a dominance of Portuguese customers in every distribution channel. In spite of this, among the customers who book a stay through a Corporate the proportions are more leveled, with the foreigner customers making more of 40%.

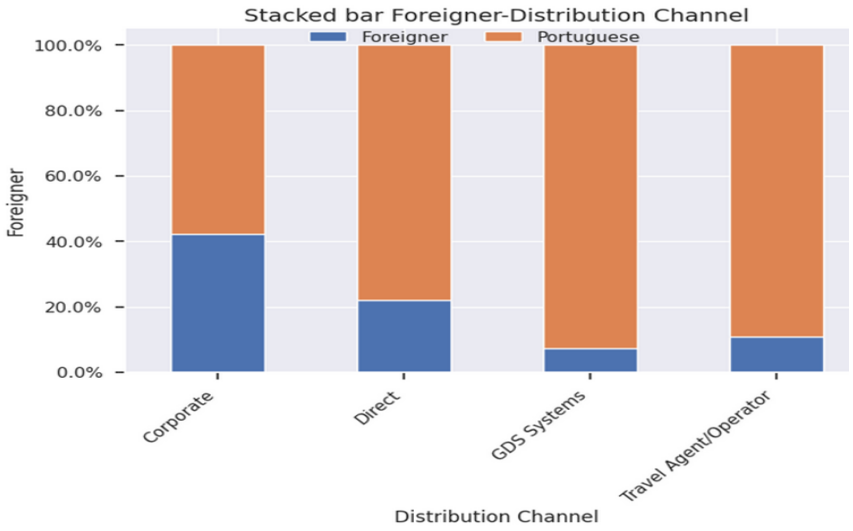

Fig 3

3.2.4. Outliers Handling

Of all the metric features, only ‘DaysSinceCreation’ didn’t had outliers, to handle them we have tried 4 approaches: Z-Score approach, IQR (Interquartile Range) approach, manual outlier handling , an adapted Winsorize approach and even resorted to the creation of new features less affected by outliers. As to reduce the loss of information we decided to use the adapted Winsorize which chosen a threshold, sets all greater values as equal to the said threshold, and created new features. The adapted winsorize was only used on ‘AverageLeadTime’, setting the threshold on 365 (1 year) and after on a newly created feature which still needed some outliers handle (‘TotalRevenue’).

3.2.5. Further Feature Engineering

We decided to create 9 new features with the help of the previous ones, and these are:

- Age Category - Bins were created for age groups classifying ages into categories such as '18-29', '30-39', '40-49', '50-59', '60-69', '>=70'.
- Total Revenue - A new column TotalRevenue is calculated by summing LodgingRevenue and OtherRevenue. It showed some outliers which were handled using the adapted Winsorize with threshold 5000 after its boxplot analysis.
- Premium Service Rate - A new feature that indicates the premium level of service based on revenue, calculated with the division of ‘OtherRevenue’ by ‘TotalRevenue’ and its set as 0 if ‘TotalRevenue’ is 0.
- Booking Success Rate - Computed as the ratio of BookingsCheckedIn to the total number of bookings (including checked in, canceled, and no-shows). As we perceived this column as a good indicator, the rest of the bookings related features were dropped.

- Preferred Room Location - A weighted score for room location preference is calculated using the existing features related to room location, with the consideration to potentially change it to a geometric sequence.
- Preferred Bed Type - Created using a lambda function to calculate a bed preference score based on bed preference and crib preference features.
- Preferred Bathroom Type - Represents the bathroom preference, scored by the preference of a bathtub or shower.
- Preferred Elevator Proximity - A numerical score representing elevator proximity preference, where being near or away from the elevator is factored. The used features were dropped.
- Quietness and Alcohol Preferences - Engineered by applying a lambda function to evaluate the preference for room quietness and alcohol in the room.
- Selective Customer - We categorized customers based on how selective they are with their room attributes, labeling them as 'Not Selective', 'Sometimes Selective', or 'Very Selective'. This was determined by counting how many specific rooms feature a customer has preferences for, such as room location, bed type, bathroom type, elevator proximity, and preferences for quietness and alcohol availability.
- Average Person per Room - Captures the average number of people per room, calculated by dividing 'PersonsNights' by 'RoomNights'. This feature is beneficial as it provides a more granular look at occupancy, is instrumental in identifying outliers and allowed for the elimination these very correlated features without information loss.

In each step, these new features are appended to metric_features or non_metric_features list as appropriate, and original columns used to create these new features are dropped from the data frame to avoid redundancy.

3.2.6. Data Transformation

We adopted one-hot encoding for various non-ordinal categorical features including 'Region', 'DistributionChannel', 'PreferredRoomLocation', 'PreferredBedType', 'PreferredBathroomType', 'PreferredElevatorProximity', 'QuietnessAndAlcoholPreferences', 'AgeCategory' and 'SelectiveCustomer'. This method transforms the categories into a binary format, avoiding any implicit hierarchy and making the information more digestible for our predictive models.

Following the encoding process, we shifted our focus to the scaling of metric features like 'DaysSinceCreation','AverageLeadTime','TotalRevenue','PremiumServiceRate','BookingSuccessRate', and 'AveragePersonRoom'. Utilizing MinMaxScaler, we normalized these features to a range of 0 to 1, thereby aligning their scales. This step is crucial to ensure that larger-valued features do not disproportionately influence the model.

3.3. MODELING

Before modelling we created a new separate data frame using only 5 metric features (AverageLeadTime , BookingSuccessRate, PremiumServiceRate, AveragePersonRoom, TotalRevenue) which we will consider afterwards to make a clustering on a behavioral perspective.

Then proceeded with the application of Principal Component Analysis (PCA). By employing PCA, we condensed the dataset to its most influential factors, representing the essential patterns of customer

behavior in a simplified manner. We first reduced the complexity of our data to just the two main components, which using a scatter plot then allowed us to inference potential clusters. We can perceive between 2 to 4 clusters.

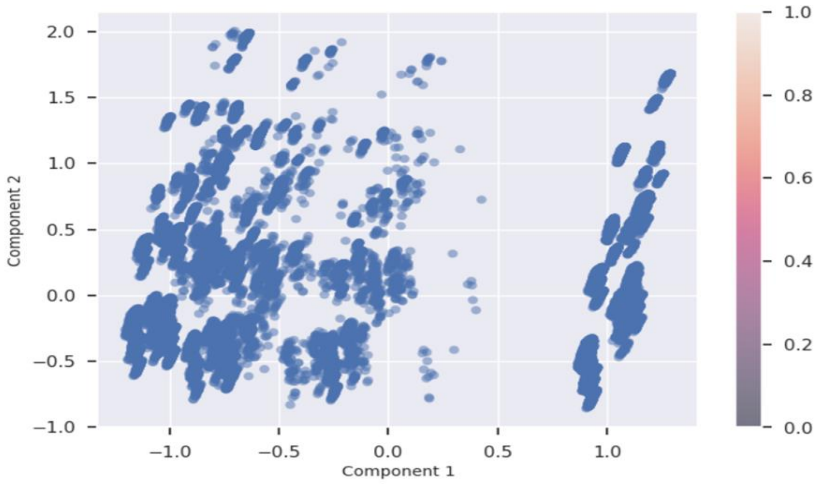

Fig 4

Next, we further refined our model to capture 95% of the variance by selecting an optimal number of principal components, in this case 26. This detailed PCA enabled us to understand which features contribute most to customer decisions and booking habits and which had little to no impact.

We then proceed to cluster the data using the 26 principal components previously chosen and using the K-means algorithm. Specifically, the K-means++ was used as it tend to present better results.

The “Elbow Curve” graph was instrumental in determining the optimal number of clusters for K-Means clustering. By identifying a suitable 'k', we can effectively group guests for targeted marketing and personalized service offerings, improving customer satisfaction and loyalty.

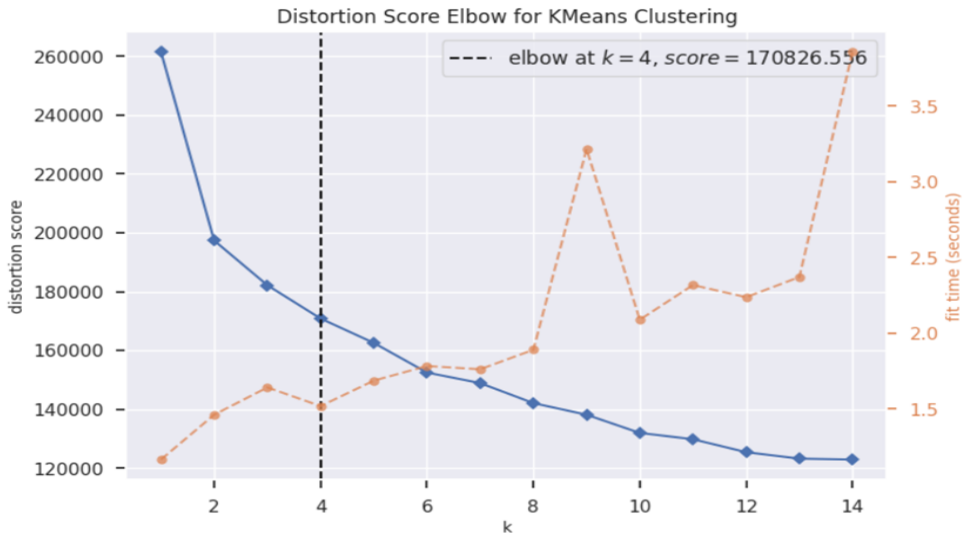

Fig 4

After thorough evaluation, we have determined the optimal number of clusters for our hotel business to be four. Together, these analytical approaches equip us with a comprehensive

understanding of our guests, enabling us to anticipate their needs, suggest relevant services, and personalize their stays, thereby elevating the overall guest experience in our hotels.

3.4. EVALUATION

The Cluster Cardinality chart indicates that the Cluster 0 and the Cluster 2 are the most populated meaning that these are the group of customers that the Hotel should be more concerned in address on their marketing investment.

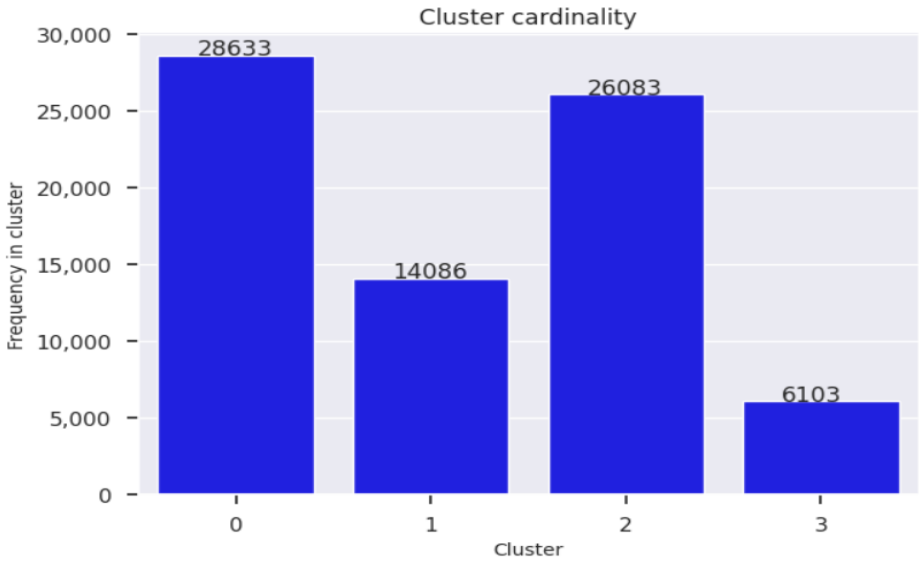

Fig 5

We leveraged the KMeans Intercluster Distance Map to depict the division of customer segments in a simplified two-dimensional PCA space (Fig 6). Despite the fact that these 2 components represent only about 32.5% of the variance, the map showcases a notable separation between clusters. This visual separation reinforces the presence of distinct customer segments, each potentially requiring a different approach in marketing and service design.

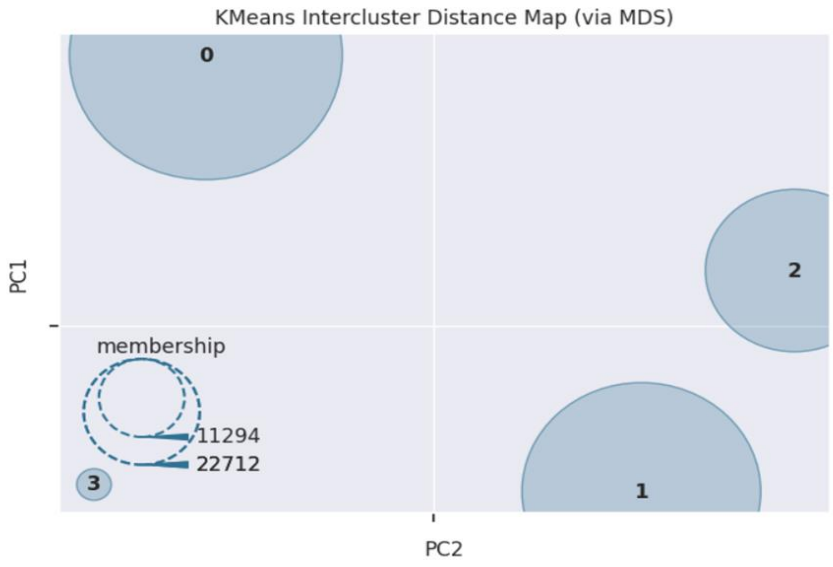

Fig 6

4. RESULTS EVALUATION

4.1. GENERAL SEGMENTATION

Our latest cluster analysis has provided key insights into our guest demographics. Below is an exploration of the identified segments to guide our service and marketing strategies.

Cluster 0: The Global Guests

Characteristics: These guests show a moderate average lead time, suggesting they plan their stays with a fair amount of foresight. They exclusively consist of foreign visitors, indicating a global clientele possibly traveling for leisure or business.

Behaviors: The use of corporate channels is minimal, with a higher inclination towards direct bookings and travel agents, suggesting a preference for personalized or third-party facilitated arrangements. Their DaysSinceCreation are quite significant, hinting at a mix of new and returning guests.

Preferences: Almost all have no specific room location or bed type preference, indicating flexibility in their choices. Their total revenue and premium service rates are moderate, suggesting balanced spending habits.

Cluster 1: The Planner Pack

Characteristics: This cluster demonstrates a longer lead time than Cluster 0 but shorter than Cluster 2, indicating guests who plan their stays well but not too far in advance. They have a mix of foreign and local guests, with a noticeable portion from Portugal.

Behaviors: Primarily booking through travel agents, this group seems to rely on external assistance for planning their stays. They generate a good amount of revenue, hinting at a willingness to spend on their accommodation.

Preferences: A significant preference for twin beds and a notable interest in premium services suggest they value comfort and are willing to pay for enhanced experiences.

Cluster 2: The Comfort Seekers

Characteristics: With a lead time similar to Cluster 0, these guests also plan ahead but not excessively. The cluster has a mix of foreigners and locals, with diverse regional representations.

Behaviors: Like Cluster 1, they predominantly use travel agents, indicating a preference for convenience in booking processes. Their revenue contribution is substantial, though not the highest.

Preferences: Unlike the other clusters, they show no clear preference for bed type, suggesting adaptability. However, their interest in premium services is akin to Cluster 1, indicating a desire for comfortable and high-quality accommodations.

Cluster 3: The Domestic Travelers

Characteristics: Exhibiting the shortest lead time, these guests are likely spontaneous or short-term planners, exclusively from Portugal, indicating a domestic market focus.

Behaviors: They show a varied use of booking channels, with a significant portion through travel agents, yet a considerable amount directly, suggesting a mix of planning preferences. Their total revenue is the lowest, which might indicate shorter stays or less spending.

Preferences: They predominantly have no specific preferences for room location or bed type, showcasing a flexible approach to their bookings. Their premium service rate is similar to other clusters, showing a consistent interest in additional services across different guest types.

The cluster tables with all the relevant mean cluster values per feature can be consulted on the notebook.

4.2. Behavioral Perspective Segmentation

These are clusters obtained using the reduced data frame with a behavioral perspective.

By understanding these clusters, we can tailor our marketing strategies and service offerings to better meet the distinct needs and preferences of each traveler type, potentially enhancing guest satisfaction and operational revenue.

With that in mind, the elbow method graph indicates that three clusters provide a reasonable balance between within-cluster variance and the number of clusters. The silhouette plot supports this, showing that while there is some overlap in characteristics, the average silhouette score is positive, indicating appropriately distanced clusters.

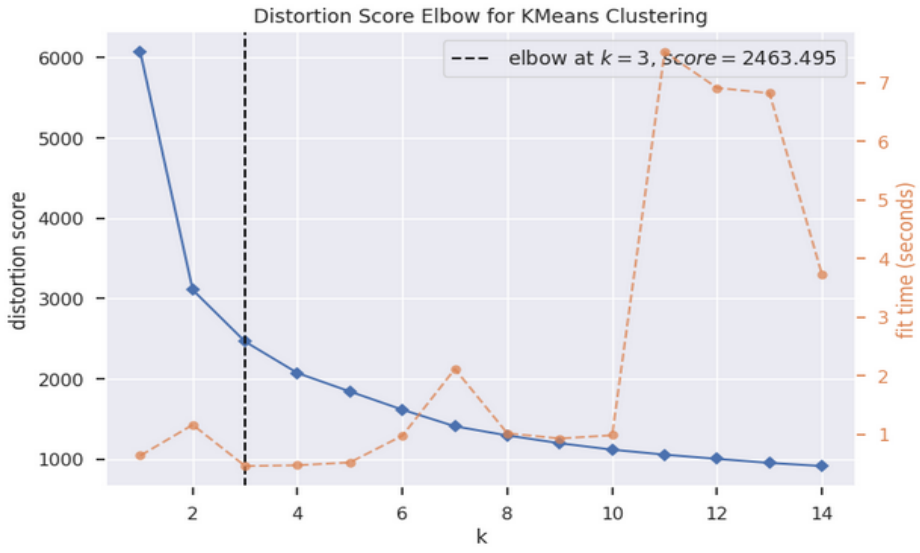

Fig 7

Cluster 0: Spontaneous Travelers

Characteristics: With the lowest average lead time, this cluster embodies travelers who make spontaneous decisions, suggesting a propensity for impromptu trips. The data doesn't specify age, but their behavior suggests they might prioritize flexibility and convenience over detailed planning.

Behaviors: Their booking success rate is nearly perfect, indicating a high commitment level once they decide to book. They contribute moderately to total revenue, suggesting their stays might be brief or less indulgent in additional services.

Preferences: The lowest premium service rate in this group hints at a preference for basic, no-frills accommodations, possibly driven by cost-effectiveness or a focus on simplicity and convenience.

Cluster 1: Advanced Planners

Characteristics: Exhibiting the highest average lead time, these travelers plan their stays well in advance, possibly to ensure availability or to meticulously plan their travel itinerary. They might represent a diverse demographic, focusing on a structured and well-organized travel experience.

Behaviors: Their exceptional booking success rate underscores a strong determination in their travel plans, likely translating into reliable revenue for the hotel. Their preference for premium services is notable, indicating a willingness to invest in enhanced comfort or luxury.

Preferences: The average number of people per room is higher than in other clusters, suggesting these travelers might often book in groups or families, valuing shared experiences or the need for multiple accommodations.

Cluster 2: Balanced Travelers

Characteristics: These travelers exhibit a moderate average lead time, indicating a balance between spontaneity and planned behavior. Their booking habits suggest a blend of young and mature travelers who seek a mix of adventure and comfort.

Behaviors: Their total revenue generation is the highest, reflecting either longer stays, a tendency to avail additional services, or a preference for higher-priced accommodations. Their near-perfect booking success rate illustrates a high level of commitment to their bookings.

Preferences: The preference for slightly more people per room than Cluster 0 could imply family travel or group bookings, indicating a social aspect to their travel or the value they place on shared experiences.

The cluster tables with all the relevant mean cluster values per feature can be consulted on the notebook.

The cluster cardinality bar graph displays the number of observations within each cluster, with Cluster 0 being the largest, followed by Cluster 2, and Cluster 1 being the smallest. This distribution suggests that while fewer customers may fall into the high-spending Cluster 1, they contribute substantially to the revenue.

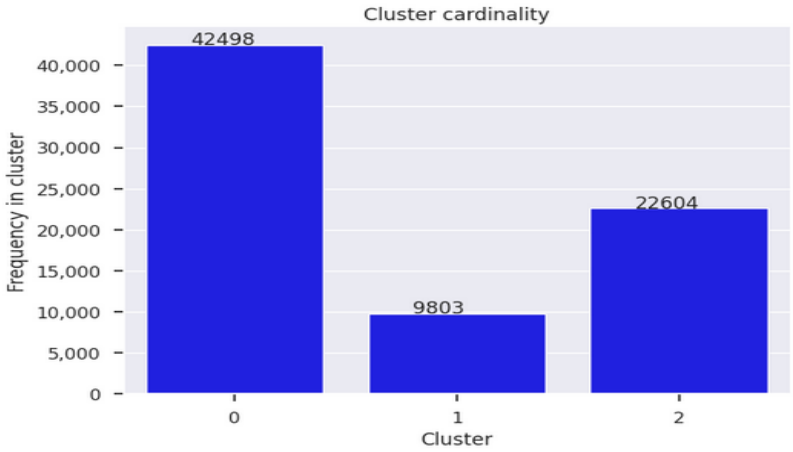

Fig 8

5. DEPLOYMENT AND MAINTENANCE PLANS

There are challenges to keep database integrity and avoid the redundancy of data. This would require a proper data warehouse environment with data quality procedures. Also, there should be assessed the need for purging or summarizing historical data that may help for report performance.

Also, the merge of known customer behavior in a well-designed schema will definitively help the build of specific customer key business indicators and predictors.

This data warehouse can be initially deployed on the cloud for testing purposes, being fueled by the proper pipelines doing all the necessary transformations and final report design. This will require the hiring of proper IT personnel with knowledge of data science and engineering which will keep such oversight in constant pace for the whole hotel chain for the internal reservation system. It could be also achieved through a contract with an external source of experts for the same purpose, with very clear tasks and deliveries.

6. CONCLUSIONS

We believe that the insights gained from our latest clustering analysis enable us to deepen our understanding of our guests' diverse needs, allowing us to tailor our services, optimize resource allocation, develop targeted marketing strategies, and ultimately deliver a more personalized and fulfilling guest experience. This refined approach to customer segmentation through data-driven analysis is set to enhance our operational efficiency and foster a more strategic, focused application of our marketing resources.

By delving into the distinct characteristics and preferences of each guest cluster, we can design superior experiences and more effective marketing initiatives. Here's how we envision the benefits:

Personalized Guest Experiences:

Recognizing the specific preferences identified in the new clusters enables us to fine-tune the guest experience significantly. For example, guests in Cluster 0, with a moderate average lead time, might value flexible booking options and personalized service offerings, whereas guests in Cluster 2, who

plan their stays well in advance, could appreciate detailed information on their accommodation and surroundings well before their arrival.

Enhanced Marketing Initiatives:

A clear understanding of each cluster's booking behavior and service preferences allows for highly customized marketing efforts. For instance, Cluster 2, with the highest premium service rate, is an ideal target for campaigns emphasizing luxury and exclusive offerings. Conversely, Cluster 1, with a lower booking success rate and premium service rate, may respond better to promotions that highlight value and convenience.

Revenue Optimization Strategies:

Insight into each cluster's total revenue helps refine our pricing and promotional tactics. While Cluster 0 contributes significant revenue with a high booking success rate, Cluster 2, with the highest premium service rate, offers opportunities for upselling high-value services. Discounts and value-focused packages might be more appealing to Cluster 1, aiming to enhance their booking success rate.

Customized Loyalty Program Design:

Our loyalty programs can be adapted to align with the behaviors and preferences of the most lucrative guest clusters. Tailoring rewards to the planning habits of Cluster 2 or the flexibility sought by Cluster 0 can enhance loyalty and encourage repeat visits.

Operational Efficiency:

Understanding each cluster's preferences aids in making informed operational decisions. For instance, the higher average person per room in Cluster 0 might indicate a need for family or group-friendly accommodations, while the detailed planning characteristic of Cluster 2 suggests a desire for comprehensive service offerings.

In conclusion, leveraging detailed data from our cluster analysis enables us to meet our guests' specific needs with precision, providing experiences that not only meet but exceed their expectations. This approach goes beyond merely responding to current trends; it's a proactive strategy to shape the future of our hospitality services, ensuring sustainable growth and a loyal customer base.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

The model's performance could potentially be enhanced through more extensive experimentation with preprocessing techniques and modeling algorithms. However, given the constraints of our limited timeframe for delivery, our team acknowledges that there could have been more exploration of various data mining techniques to achieve more effective results.

7. REFERENCES

Landmann, D. (2019, June 10). Market Segments for Hotels: Want to Know Which Generate the Most Revenue? Retrieved from https://blog.hqrevenue.com/hotel-market-segments

Bação, F. L. (2023). Data Mining: Partition-based algorithms (k-means). Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa. Available on: https://elearning.novaims.unl.pt/pluginfile.php/212994/mod_resource/content/1/MAA%20MD%20Session%209b%202324.pdf

Bação, F. L. (2023). Machine Learning II: Visualization of Multidimensional Data slides 11-27. Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa. Available on: c

Scaler. (2023). Outlier Analysis in Data Mining. Retrieved from https://www.scaler.com/topics/data-mining-tutorial/outlier-analysis-in-data-mining/

8. APPENDIX

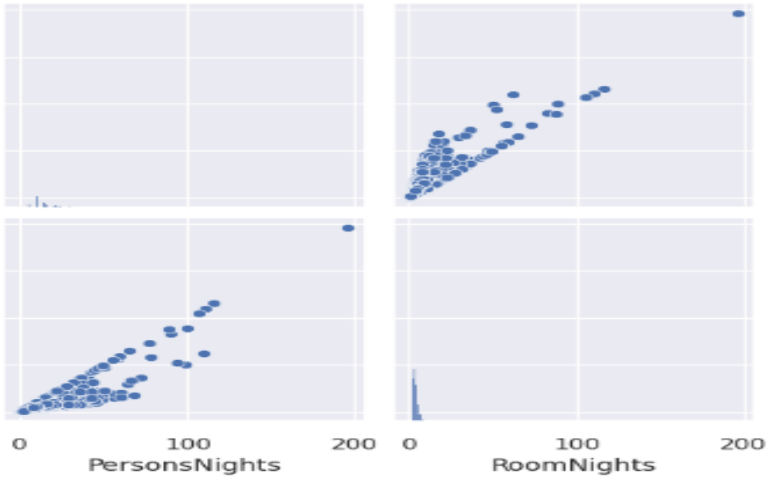

Image 1: Scatter plot “PersonsNights” x “RoomNights”

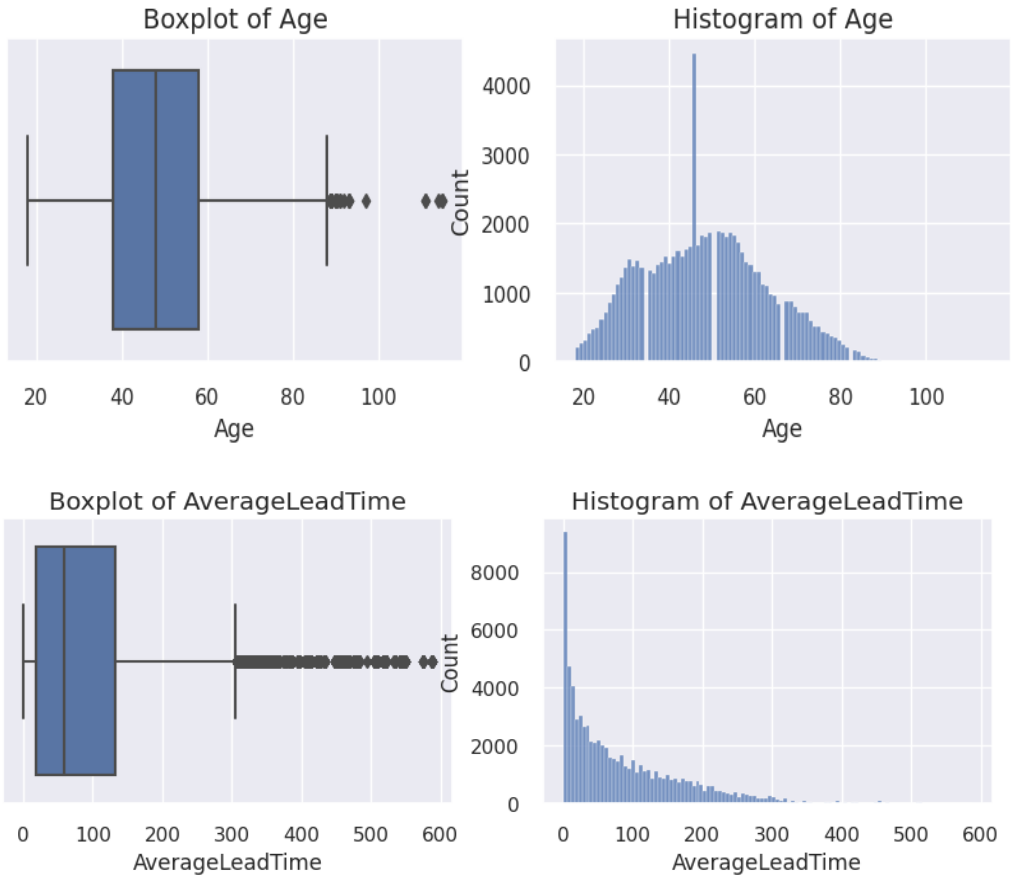

Image 2: Box plots for initial features

Image 2: Absolute frequencies of categorical variables

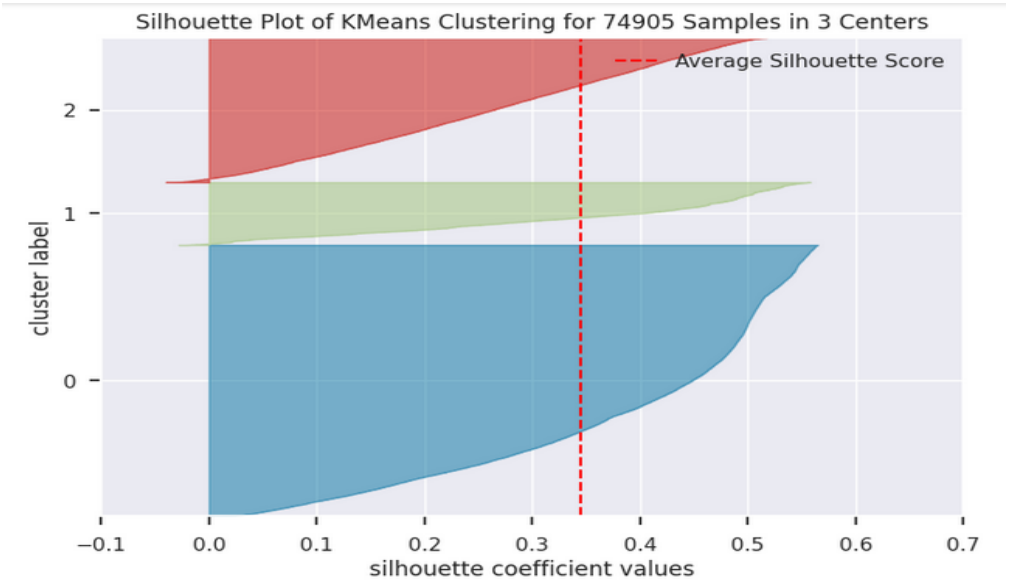

Image 4: Silhouette Plot of K-Means Clustering