Business Cases with Data Science

Master's Degree Program in Data Science and
Advanced Analytics – Major in Business Analytics.

Hotel Customer Segmentation:
A data-driven approach using K-Means.

(Case 1)

Group V:

- **André Moreira, 20222132;**
- **Beatriz Teixeira, 20211313;**
- **Catarina Nunes, 20230083;**
- **Pedro Catarro, 20230463;**

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

Contents

Keywords: Hospitality; Market segmentation; Data science; Principal Components; K-Means.

1 Introduction

In our rapidly evolving digital age, the world generates vast amounts of data every second, the hotel industry is no exception. This data transcends mere numbers and statistics; with the proper lenses, it embodies valuable information that can provide us with tools and insights into various aspects of our business and society. However, it is no secret that the volume and complexity of data require the use of appropriate tools and techniques to extract patterns that can draw us to conclusions and knowledge.

"Change is an inevitable force, and this is particularly true in the constantly evolving world of hospitality," states Daniels (2023). Recent events, such as the pandemic and the exponential growth of AI, have further emphasized this fact. The dynamic nature of this industry presents a challenge to our agility and innovative spirit. However, if we play our cards right, we can stay ahead of the competition and successfully navigate challenges and opportunities while keeping up with the latest hospitality trends (EHL Insights, 2024).

2 Business Understanding

2.1 Background & Context

This analysis is conducted as part of the "Business Cases with Data Science" course, a component of the Master's program in Data Science and Advanced Analytics at the Nova Information Management School. It addresses an issue identified by a new marketing manager within the Portuguese hotel industry.

The Lisbon hotel under analysis has a new Marketing Director who wants to upscale the customer segmentation due to the oversimplified current approach to customer segmentation, which only considers the sales origin of its guests. This method neglects crucial customer attributes such as geography, age, and behavior, all of which are essential for customizing marketing strategies to meet diverse strategy needs. Moreover, the existence of various booking channels, each appealing to distinct customer demographics with unique booking preferences, adds another layer of complexity to effectively reaching and engaging with the right audience. The lack of detailed segmentation hampers the chain's ability to fine-tune its marketing initiatives, thereby impacting its capability to attract new patrons and maintain loyalty among existing ones.

2.2 Business Problem

The hotel chain is currently facing a significant challenge due to its oversimplified approach to customer segmentation, which only considers the sales origin of its guests. This method neglects crucial customer attributes such as geography, age, and behavior, all of which are essential for customizing marketing strategies to meet diverse strategy needs. Moreover, the existence of various booking channels, each appealing to distinct customer demographics with unique booking preferences, adds another layer of complexity to effectively reaching and engaging with the right audience. The lack of detailed segmentation hampers the chain's ability to fine-tune its marketing initiatives, thereby impacting its capability to attract new patrons and maintain loyalty among existing ones.

2.3 Methods & Resources

In tackling our complex business challenge, we embrace the CRISP-DM methodology. We apply Principal Component Analysis (PCA) to reduce dimensionality and K-Means Clustering to generate new market segments.
The dataset includes information about hotel customers, including identification, demographic data, booking habits, and personal preferences. It records each customer's ID, nationality (in ISO 3166-1 Alpha 3 format), and age, along with the time passed since their profile was created. Anonymized hashes for names and personal document IDs provide privacy. Metrics like average booking lead time, total lodging, and other revenues, and booking outcomes (cancellations, no-shows, and completed stays) are tracked. The dataset also distinguishes between person nights and room-nights and notes the customer's preferred distribution channel and market segment, and personal preferences regarding room features.

2.4 Success Criteria

The main objective is to analyze and provide a new and efficient customer segmentation that captures the wide array of characteristics defining the hotel's customers that will guide our Marketing strategy. With a more accurate and multifaceted understanding of its customer base, the hotel can then customize its marketing efforts to resonate with the specific demands and preferences of various customer segments. By aligning product offerings, pricing strategies, and promotional activities with the identified needs of each distinct group, the hotel aims to not only draw in new customers but also upscale satisfaction and loyalty of its current guests, thereby fueling overall business.

- Update the current market segmentation and make it suitable for the current strategy needs.
- Clear and concise segments with clear guidelines to make the implementation easier
- Optimize the number of segments, an efficient trade-off between number of segments and personalization.

3 Data Understanding

3.1 Descriptive summary

Get into more detail, our dataset *Case1_HotelCustomerSegmentation.csv* includes 111773 costumer's records represented in 29 distinct variables for each. Each row should represent a unique guest, and each column represents guest characteristics The characterization of each variable is in the table A in the Appendix.

3.2 Exploratory Data Analysis: Univariate EDA

The univariate analysis offers an individual examination of each variable, making it one of the most straightforward methods of data analysis. Its primary objective is to provide a summary of the descriptive characteristics of a given variable, including measures of central tendency, dispersion, and distribution shape.

Numerical variables:

Analyzing our numerical variables, we can identify that the majority of them are higly skewed and only "DaysSinceCreation" is aproximately normal. This shows the need for further transformations on the data preparation step like transform outliers.

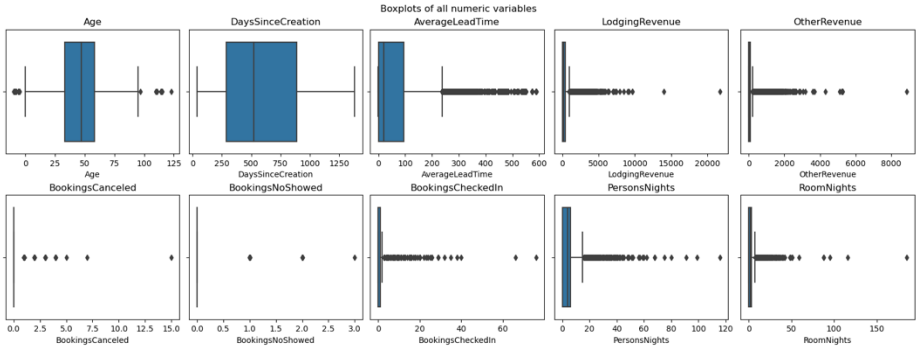

Figure 1. Numeric Features Box Plots

Categorical variables:

Regarding the distribution channels, there is a clear first option, a travel agent operator, and one with almost no significance.

Our customers – Who are they?

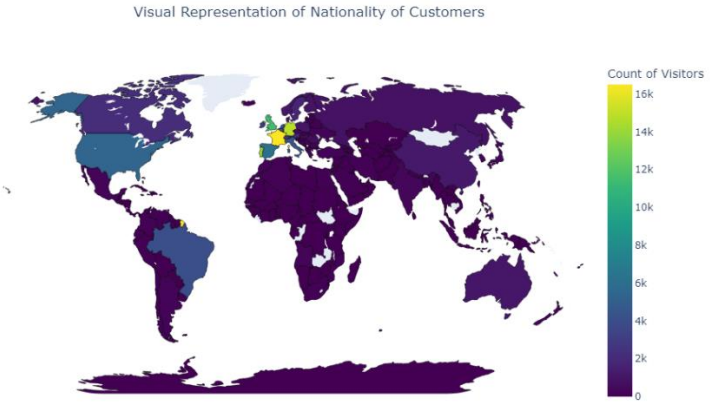

Figure 2. Nationality distribution

The Hotel Customers records show a variety of countries around the world. France leads with 15%, , followed by Germany (13%) and Locals/Portugal (13%). The United Kingdom is also a significant source of clients, representing 10% of the total. Other countries such as Spain, the United States, Italy, Belgium, Brazil, and the Netherlands also contribute to our client base, each representing between 3% to … of the total clients.

The exploratory analysis also reveals a strong concentration of customers in the age groups of 45-55 years (25%) and 35-45 years (20%), indicating a predominance in the middle age bracket. Age groups 55-65 years and 25-35 years show similar shares, each representing 18% of the total customers, while the extreme age groups, 65-100 years and 18-25 years, have smaller proportions, at 14% and 6% respectively. This distribution suggests the need for tailored marketing strategies and products to cater to different demographic segments.

Boolean variables:

After analyzing the graphic below, we are sure that the most relevant variables are the ones regarding the bed size, King-size with 36% Yes and Twin bed with 16%. After that, we can see some preferences for a

quiet room (9% Yes) or a room on a higher floor (4% Yes). There are also some cases where the customer asks for a crib /2%) indicating couples with children. The rest of the variables have little to no importance, and in future steps will be grouped.

3.3 Exploratory Data Analysis: Bivariate EDA

Bivariate data involves pairs of linked numerical observations, each pair representing two related variables. This type of data is essential for exploring relationships, correlations, and causal connections between variables.

How do the variables correlate?

Age vs Revenue: After analyzing the graphic, we can identify a lower spending group namely 20 and the rest of them are approximate to each other. The group that ends the most is the one between 40 and 60.

Potential Customers: Our segments will not take into consideration inactive 'customers', but that does not mean we can't do anything about it. We analyze records of individuals who have never checked in relative to the duration since their record creation to discern potential interest. For instance, they may have created a profile for a loyalty program or could be interested in the hotel for a future trip. We categorize these records into three groups: less than one year, between one and two years, and over two years to define potential loss of interest.

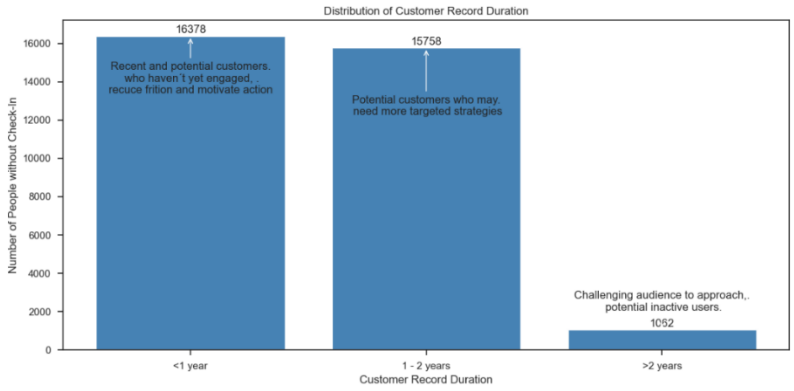

Figure 3. Startegies for Inactive Records

Distribution Channel and:

- **PersonsNights** / **TotalRevenue** / **BookingsCheckedIn**: Travel agents/operators predominantly influence these variables, while GDS systems have minimal impact.
- **BookingsNotAttended:** This column represents the quantity of both no-shows and cancellations across all bookings. The distribution channel **'Corporate'** bookings have the highest rate of unattended bookings, which poses operational challenges. Corporate bookings can involve large groups, resulting in numerous room reservations, as well as bookings for meeting rooms and restaurant tables. This increases the risk of inefficient utilization of **fixed capacity and revenue loss**, including missed opportunities for other clients to use these resources (potential gains). To address this issue, measures such as improving communication and revising policies like implementing cancellation costs can be considered.

Nationality and:

- **TotalRevenue / PersonsNights**: We categorize nationalities into gross categories primarily to understand the behavior of the top 10 nationalities versus others. France and Germany (excluding others) represent the largest sources of revenue and guests as individual countries.
- **BookingsCheckedIn / BookingsNotAttended**: It's notable that despite France and Germany having more appearances (indicating a larger variety of clients), Portuguese clients account for the highest number of check-ins and unattended bookings. This suggests that national clients tend to frequent the hotel multiple times, with individual clients making multiple bookings. In light of this observation, implementing loyalty programs and revising policies, such as introducing cancellation fees or offering the option to pay more for flexibility in rescheduling, could be beneficial.

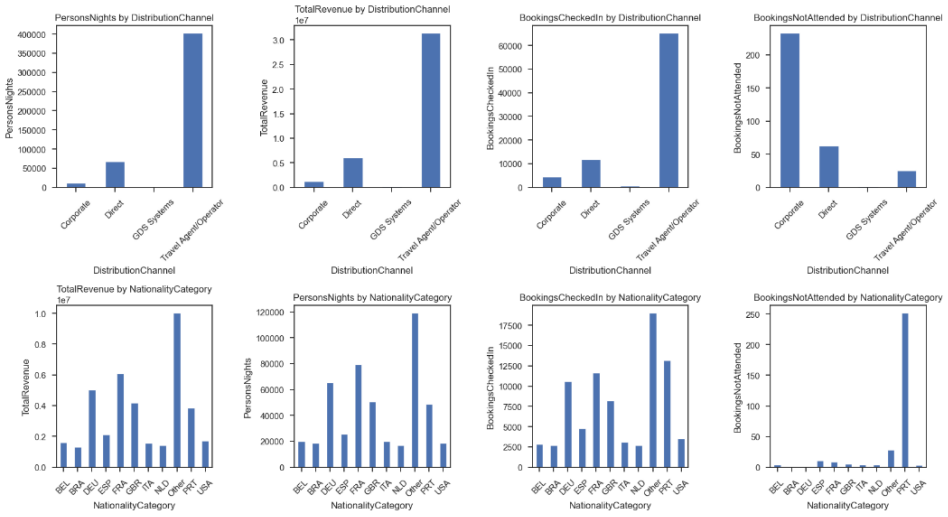

Figure 4. Bivariate EDA: Categorical - Numeric Variables

3.3.1 Old Segmentation: Why is it no longer serving us?

In today's increasingly competitive landscape, segments should provide us with more than just information about sales origin; they should also offer insights into who our clients and markets are. This allows us to gain knowledge and align our strategy accordingly. However, even if that information were sufficient, the following section will demonstrate that the old segment is outdated and no longer serves our needs.
Firstly, we observed that the most used channel is the Travel Agent and Operator. However, the old segment consistently exhibits preponderance in an undifferentiated segment called 'Other'.

Market Segment and: (Tables in Appendix)

- **TotalRevenue / PersonsNights / BookingsCheckedIn**: The data shows that the 'Other' segment has the most customers who generate the highest revenue, guest numbers, and check-ins. Contrary, the categories 'Aviation' and 'Complementary' have very few customers, suggesting that this segmentation may no longer be relevant and requires updating to better suit the current hotel industry. Furthermore, about **BookingsNotAttended**, the corporate segment still appears to have the highest rate of unattended bookings, mirroring the logic mentioned earlier.
- When analyzing the distribution of **Nationalities** by market segment, "Portugal" is mostly represented in the "Complimentary" segment. ‘Complimentary’ by hospitality definition services are provided at no charge, and Portuguese guests may be receiving more of them, indicating they are

influencers or VIPs. The corporate segment also has many Portuguese guests, as shown in the bivariate analysis.

4 Data Preparation

To implement the K-Means clustering there's the need to execute some data preparation steps before. The K-means model is an unsupervised learning algorithm that is intolerant to missing values, requires normalization, and needs a specific handle of categorical variables.

4.1 Our customer definition (partial outliers):

We use a combination of the variables **Nationality**, **NameHash**, and **DocIDHash** to identify unique individuals. However, inconsistencies arise when some companies provide company identifications instead of individual names, leading to multiple entries for the same person under different names but with the same DocIDHash. Without DocIDHash, accurately identifying individuals becomes challenging. For example, two people from Portugal may share the same name (e.g., Catarina), potentially resulting in aggregated revenue for multiple individuals. To maintain data integrity, we choose to drop all rows where DocIDHash is missing (NaN) because this is essential to our individual customer definition.

4.2 Inconsistencies related to the dataset:

The inconsistencies related to our dataset were treated during this phase, here are the most important ones:

- **Handling Age Inconsistencies:** Records with ages below 18 or above 100 were removed, as individuals in these age groups are unlikely to be potential customers, and marketing initiatives targeting minors require special considerations.
- **Handling Non-Booked Customers:** Records where customers neither canceled, not showed, nor checked in were considered adjustments or refunds and where removed after duplicate treatment. The rest represent customers, such as recent loyalty program members, and requires a targeted marketing approach to better understand their preferences, as we reefer previously.
- **Addressing Negative Revenues:** Records with negative lodging or other revenue were identified as adjustments or refunds and removed after duplicate treatment.
- **Outliers:** we tried three different approaches being them interquartile method, z-score method and an ad-hoc method being the last one the most effective way to deal the data. The first method led to a 17.38% data loss, the second to a 5.24% loss and the third one can be adapted for our analysis scope.

4.3 Feature Engineering

During this process, after analyzing the most important KPIs in the hotel industry, we have added to the dataset 7 more variables to help with the business perspective being them:

- ***BookingSuccessRate***: It indicates the ratio of successful bookings (checked in) to the total number of bookings (checked-in, canceled, and no-show bookings).
- ***TotalSpecialRequests***: Represents the total number of special requests made by guests, including various types such as high floor, low floor, accessible room, etc.
- ***RevenuePerPersonNight***: Represents the average revenue generated per person night, calculated by dividing total revenue by the total number of person nights.
- ***TotalRevenue***: Represents the overall revenue generated from bookings, including lodging revenue and other revenue.

- ***SpendingPerBooking***: Indicates the average spending per booking, calculated by dividing total revenue by the total number of bookings.
- ***BookingsNotAttended***: Represents the total number of bookings that were not attended by guests, including both canceled and no-show bookings.
- ***Age Group***: Categorizes age into specific groups (Young Adult, Adult, Middle Age, Senior) for contextual purposes after clustering.

The next steps involved selecting the most relevant variables based on their correlations. For example, the variable RoomNights was removed in favor of PersonNights to better capture group dynamics. Variables with minimal incidences, like BookingsCanceled and BookingsNoShowed, were combined into BookingsNotAttended. Nationalities were grouped together for simplicity and then subjected to one-hot encoding for categorical variables. Numerical variables were scaled using Min-Max scaling to ensure optimal performance of PCA and K-Means algorithms. For more details and explanations, please refer to the **Notebook** of this analysis.

4.4 Actions to improve data quality: Missing Values

The last step before moving to PCA is to handle the missing values. During this phase, By observing the missingno matrix, it's clear that missing DocIDHash records correspond with missing Age records. This suggests a consistency in the missing data patterns.

The consistent missing data patterns indicate potential issues with data quality, particularly regarding the absence of mandatory ID identification during check-in processes. This shows the importance of implementing consistent rules and standard procedures for guest check-ins to ensure accurate data collection.

It's also noted that access to clients' personal information may not be available until they physically arrive for check-in, possibly due to limitations imposed by the agency or privacy concerns. This lack of direct access to client information can impact data accuracy and may need to be addressed through improved communication or collaboration with relevant parties.

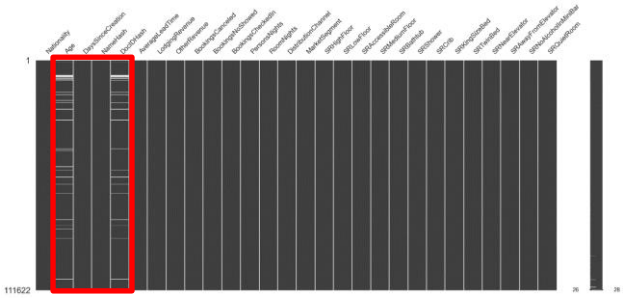

Figure 5. Missing Values Matrix

5 Principal Component Analysis

"PCA and k-means clustering are both unsupervised machine learning techniques used for data analysis, but they have different goals and methods. PCA is used to reduce the dimensionality of the data, while k-means clustering groups data points together based on similarity." (IBM, 2023)

To figure out how many components will be used in the PCA we decided to set a minimal **threshold of 0.95** for the cumulative explained variance to retain at least **95% of the data variance** after dimensionality reduction. It's a good rule that minimizes information loss.

The picture below shows that the ideal number of components is between 5 and 8, making a 95% marker a good trade-off between having fewer components and improving the algorithm. Now it's established that the ideal number of components is 5.

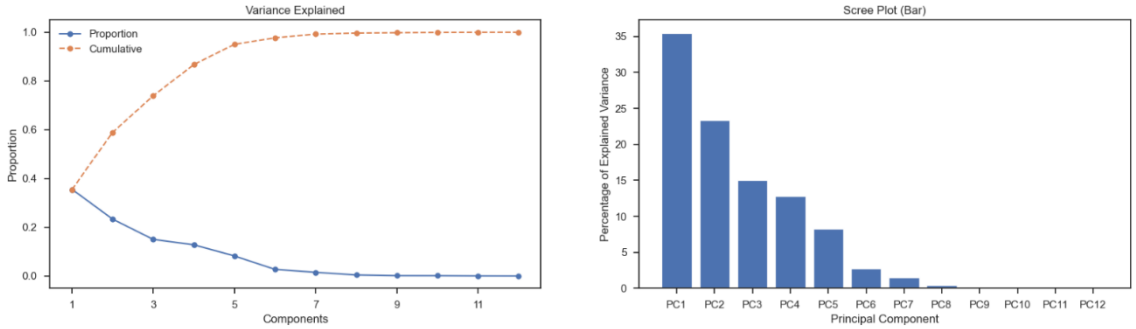

Figure 6. Ideal number of PCA components

A quick analysis of the explained variance of each variable for the corresponding Principal Component reveals that BookingsCheckedIn, RevenuePerPersonNight, and BookingsNotAttended lack representation, as expected, due to their low variance. In this step, we decided to test the K-Means Model using only **PCA_Component_1**, **PCA_Component_2**, **PCA_Component_3**, and **PCA_Component_5**, since PCA_Component_4 primarily explains the variance of Age, a role already covered by **PCA_Component_3**. After testing with both four and five components, we achieved better results with only four components.

6 Modeling: K-Means

The k-means clustering model iteratively assigns data points to the nearest centroid, which is initially random. Through each iteration, it recalculates the centroids until they no longer change, effectively defining the center of each cluster. To determine the optimal number of clusters (K) for the **4 PCA components established earlier,** we use methods such as the elbow method (that need to be complemented), silhouette score, and silhouette diagram.

The choose method was the silhouette coefficient, which measures the proximity of points within a cluster to points in neighboring clusters (ranging from –1 to 1), indicates an optimal number of clusters equal to **K= 4.**

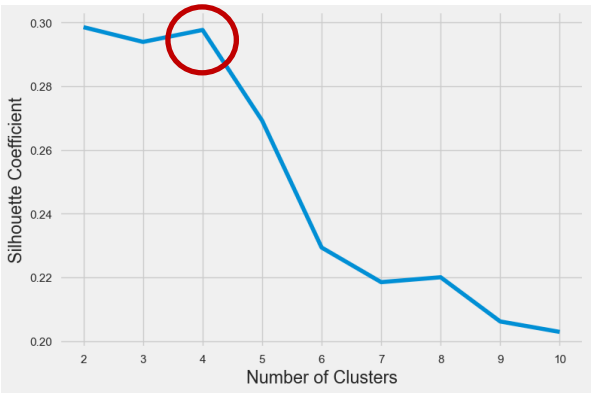

Figure 7. Ideal number of Cluesters.
For n_clusters = 4 The average silhouette_score is : 0.29739229848396

Having a maximum silhouette score of approximately 0.30 is suboptimal, as ideally, scores should exceed 0.5 for effective cluster separation. However, with the current PCA variables, this is the result we have obtained. Additionally, it's noteworthy that the data points appear to be closely clustered together. This

suggests that further refinement and optimization of the clustering model may be necessary to achieve better separation between clusters.

The result of r-squared value calculated by the r2 function is 0.9161, which is quite high, indicating that the clustering solution explains a significant portion of the variability in the data.

Below there's a representation of the inter-cluster distance concluding that all clusters are sufficiently apart from each other:

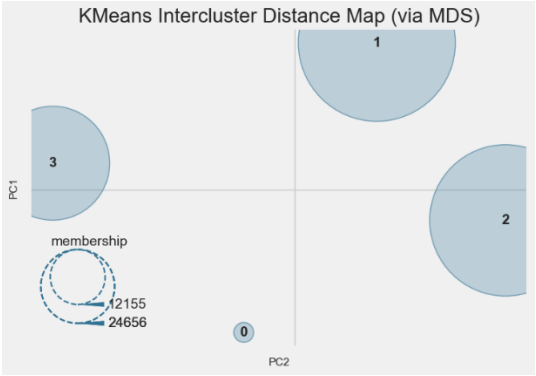

Figure 8. InterCluster Distance Map

- **Cluster Sizes:** While cluster sizes vary, this is not necessarily problematic. However, imbalances, such as a smaller cluster size compared to others (e.g., Cluster 3), may require further investigation as it could represent a niche segment.
- **Average Centroid Distances:** Lower centroid distances suggest tighter clusters, indicating greater similarity among data points within each cluster. The observed distances appear reasonable, with Cluster 1 and 2 exhibiting the lowest average distance, suggesting cohesive clusters.

7 Evaluation

The customer segmentation analysis reveals distinct distribution percentages across different age groups and regions. Among adults aged 26-40, 44% are in Cluster 2, while 41% are in Cluster 1. Middle-aged individuals (41-60) constitute 41% of Cluster 1. Seniors (61-100) account for 41% of Cluster 3. Notably, 44% of young adults (18-25) are in Cluster 1, which is the most well-represented in Cluster 0, the younger cluster.

Cluster 3, predominantly occupied by senior customers, presents challenges in marketing due to potential limitations in online engagement and a preference for personalized attention. Personalized marketing strategies focusing on quality service and personalized experiences may be necessary to effectively reach and engage this demographic.

When analyzing top countries, France has a higher incidence in Clusters 1 and 2, with Germany showing a dominant presence in Cluster 3. Portugal exhibits a higher incidence in Clusters 1 and 2, especially in Cluster 2.

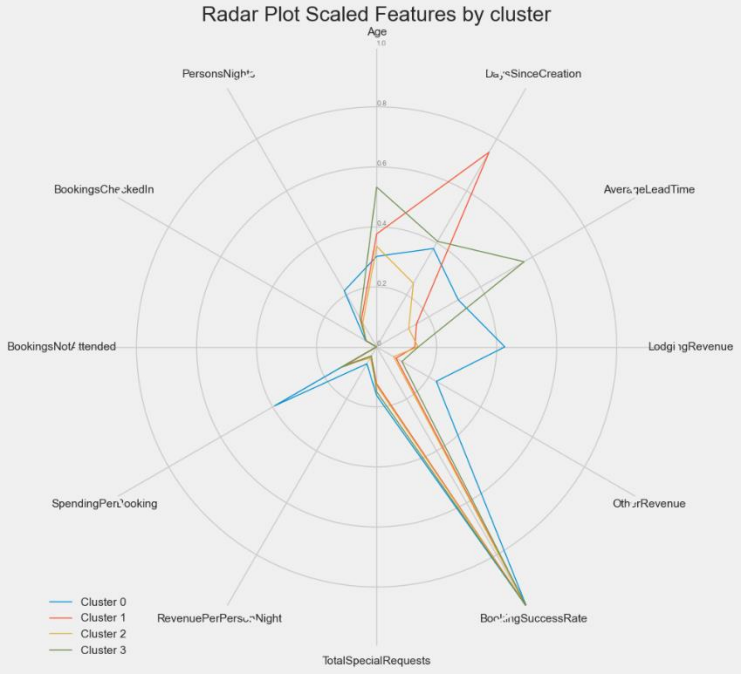

Figure 9. Cluster Radar Map

During this phase, we can analyze and describe each cluster based on the mean values for each variables per cluster:

7.1 Cluster 0: Young Elite

- This cluster represents the youngest and most affluent clients.
- They are highly loyal, with a tendency to make large group bookings, indicating they may be traveling for leisure or events.
- They have specific preferences for room types, such as King Size Beds and Quiet Rooms.
- They are willing to spend more on additional services and have a high overall spending per booking.
- They predominantly originate from Europe, with a significant portion booking directly.

7.2 Cluster 1: Established Guests

- This cluster consists of middle-aged clients who have been with the hotel for a considerable time.
- They exhibit a moderate level of loyalty and tend to book with shorter lead times.
- Although they have the lowest cancellation rate, they make fewer special requests, suggesting they may have more standardized preferences.
- They have a strong presence of clients from Portugal and often book through corporate channels.

7.3 Cluster 2: New Corporate Clients

- This cluster includes middle-aged clients who are relatively new to the company.
- They book with the shortest lead times and have a preference for corporate channels and GDS Systems.

- Similar to Cluster 1, they make fewer special requests, indicating they may be more business-oriented travelers.
- They have a significant presence of clients from Portugal, suggesting potential business ties with the region.

7.4 Cluster 3: Experienced Planners

- This cluster comprises the oldest clients who plan their trips well in advance.
- They are affluent, with high lodging and other revenues, and have specific preferences for room types.
- Despite their age, they have the lowest rate of bookings not attended, indicating reliability.
- They originate predominantly from North America and Europe, often booking through travel agents or operators.

8 Deployment

8.1 Strategies to approach the different Clusters

Cluster 0: Young Elite

For this type of customer, we will have to approach them through web platforms mainly the ones that deal with group holidays, create partnerships with different festivals near Lisbon in a way that the hotel can accommodate their participants and establish a booking price to avoid cancelations.

- Emphasize unique and enticing experiences to capture their attention and encourage repeat visits.
- Collaborate with local businesses, cafes, bars, restaurants, and cultural events to offer exclusive deals and packages, attracting their interest.
- Implement loyalty programs tailored to their preferences, such as discounts on future bookings or complimentary amenities for frequent stays.
- Offer fun and relaxing experiences within the hotel premises to cater to their preferences.
- Provide Portuguese wine experiences, such as wine tastings or wine pairing dinners, showcasing the local wine culture and flavors.
- Arrange local food experiences, including pastry and cheese tastings, allowing guests to indulge in authentic Portuguese flavors and delicacies.
- Incorporate recreational activities like poolside parties, live music events, or themed evenings to create a vibrant and enjoyable atmosphere.
- Highlight these unique experiences through targeted marketing efforts to attract the attention of young and affluent clients, encouraging them to choose the hotel for their stay in Lisbon.

Cluster 1: Established Guests

This group of customers must be convinced to book in advance and for that reason the hotel should create a discount if the booking is done at least a month before increasing the lead time, since this group gathers portuguese clients, the campaigns should be done through billboards around big portuguese cities like Porto or Faro and as the clients usually pretend a more standardized service, the hotel can adapt the rooms to have a more cost-efficient service since the customer doesn’t have extra preferences.

Cluster 2: New Corporate Clients

This type of corporate client must be led to come back to the hotel in the near future for leisure purposes and that can be accomplished by creating a discount on a second booking improving recurrence, another recommendation would be, as the customers are mostly corporate, contact directly their companies and make deals with them, mostly companies who aren't in Lisbon.

Cluster 3: Experienced Planners

Target senior clients who are affluent tourists, less tech-savvy, but highly loyal. Personalized marketing strategies focusing on quality service and personalized experiences may be necessary to effectively reach and engage this demographic. When communicating with this group of people, web ads should be avoided choosing a more direct approach like phone calls:

- Recognize their resistance to online approaches and instead focus on personalized and friendly customer service.
- Provide pre-made itinerary options based on their preferences, gathered through a questionnaire upon arrival, to show genuine concern for their trip.
- Offer curated experiences and recommendations tailored to their interests and needs, enhancing their overall stay in the city.
- Ensure clear and accessible communication channels for assistance and inquiries, such as in-person concierge services or dedicated phone lines.
- Leverage their loyalty by offering exclusive benefits or rewards for repeat visits and referrals.

8.2 Smart Helper

The Client Assignment and Marketing Strategy Input System provides a user-friendly interface for assigning clients to clusters and suggesting corresponding marketing strategies. Users input client information, including numerical and categorical variables, and the system assigns them to a cluster, offering tailored marketing strategies. While the system serves as a basic example, it can be refined and expanded over time to accommodate additional variables and improve decision-making in targeted marketing campaigns. It operates based on the Euclidean distance between the client and the average values of each cluster.

```
Enter value for Age:  35
Enter value for DaysSinceCreation:  500
Enter value for AverageLeadTime:  60
Enter value for LodgingRevenue:  800
Enter value for OtherRevenue:  100
Enter value for BookingsCheckedIn:  2
Enter value for PersonsNights:  8
Enter value for SRHighFloor:  0
Enter value for SRCrib:  0
Enter value for SRKingSizeBed:  1
Enter value for SRTwinBed:  0
Enter value for SRQuietRoom:  1
Enter value for BookingSuccessRate:  0.9
Enter value for TotalSpecialRequests:  2
Enter value for RevenuePerPersonNight:  100
Enter value for SpendingPerBooking:  900
Enter value for BookingsNotAttended:  0
Enter value for Origin_North_America: (0/1)  0
Enter value for Origin_Other_Europe: (0/1)  1
Enter value for Origin_Others: (0/1)  0
Enter value for Origin_Portugal: (0/1)  0
Enter value for Origin_South_America: (0/1)  0
Enter value for DistributionChannel_Corporate:  0
Enter value for DistributionChannel_Direct:  1
Enter value for DistributionChannel_GDS Systems:  0
Enter value for DistributionChannel_Travel Agent/Operator:  0
Cliente atribuído ao Cluster: 2
Marketing Strategies:
- Offer discounts on second bookings to encourage recurrence
- Establish direct deals with corporate clients' companies
- Highlight cultural diversity in marketing materials
- Offer multilingual customer support
```

Figure 10. Smart Helper Example

9 Conclusion

In conclusion, while the current market segmentation marks a significant improvement, there's still potential for refinement. Incorporating additional variables such as age documentation, nationality, duration of stay, and time variables like month of stays could enhance clustering accuracy. These improvements would provide more nuanced insights into customer behavior, enabling more targeted marketing strategies. Thus, there's an opportunity to enhance segmentation by refining variables to better meet customer needs.

9.1 Limitations and Potential Improvements

To improve data quality and gather more comprehensive insights, the following suggestions are recommended:

- Implement a policy requiring the collection of age documentation and nationality for all customers. This will ensure more accurate demographic information and facilitate better segmentation analysis.
- Incorporate additional information such as the number of nights spent or the average duration of stay. This data can help distinguish between individuals who are truly part of a specific group and those who may have irregular stays at the hotel.
- Introduce time variables, such as the month of stays, to enable the analysis of seasonality trends. Time series analysis can provide valuable insights into booking patterns and peak periods, aiding in strategic planning and resource allocation.

By implementing these improvements, the dataset will become more robust, allowing for more accurate analysis, especially when defining segments.

References

EHL Insights. (2024, January). *Hospitality industry trends (2024 update)*. Hospitalityinsights.ehl.edu. https://hospitalityinsights.ehl.edu/hospitality-industry-trends?hs_amp=true#Artificial-intelligence-and-technology

Daniels, A. (2023, August). *Embracing Change: A Key to Thriving in the Hospitality Industry*. Www.linkedin.com. https://www.linkedin.com/pulse/embracing-change-key-thriving-hospitality-industry-adam-daniels/

IBM. (2023, December). *What is principal component analysis? | IBM*. Www.ibm.com. https://www.ibm.com/topics/principal-component-analysis

Appendix A

Table A 1. Nomenclature and Categorization of Variables

xi; "*variable name*"; "type of variable"			
x1: ID	C	x19: SRAccessibleRoom	B
x2: Nationality	C	x20: SRMediumFloor	B
x3: Age	N	x21: SRBathtub	B
x4: DaysSinceCreation	N	x22: SRShower	B
x5: NameHash	C	x23: SRCrib	B
x6: DocIDHash	C	x24: SRKingSizeBed	B
x7: AverageLeadTime	N	x25: SRTwinBed	B
x8: LodgingRevenue	N	x26: SRNearElevator	B
x9: OtherRevenue	N	x27: SRAwayFromElevator	B
x10: BookingsCanceled	N	x28: SRNoAlcoholInMiniBar	B
x11: BookingsNoShowed	N	x29: SRQuietRoom	B
x12: BookingsCheckedin	N	x30: BookingSucessRate	N
x13: PersonNights	N	x31: TotalSpecialRequests	B
x14: RoomNight	N	x32: RevenuePerPersonNIght	N
x15: DistributionChannel	C	x33: SpendingPerBooking	N
x16: MarketSegment	C	x34: BookingsNoAttended	B
x17: SRHighFloor	B	x35: Age Group	C
x18: SRLowFloor	B	x36: Origin	C

Where, N - Numerical; C – Categorical and B – Boolean.

	AgeGroup	Customers	Percentage
0	(45, 55]	24546	25%
1	(35, 45]	20440	20%
2	(55, 65]	17753	18%
3	(25, 35]	17749	18%
4	(65, 100]	13903	14%
5	(18, 25]	5736	6%

Figure 11. distribution of records per Age Group

	MarketSegment	TotalRevenue	Percentage
5	Other	23023016.41	59%
3	Direct	6105310.29	16%
6	Travel Agent/Operator	4644151.25	12%
4	Groups	3887811.19	10%
2	Corporate	1082017.43	3%
0	Aviation	176023.60	0%
1	Complementary	24428.75	0%

	MarketSegment	BookingsNotAttended	Percentage
2	Corporate	208	65%
3	Direct	35	11%
1	Complementary	32	10%
0	Aviation	28	9%
6	Travel Agent/Operator	9	3%
5	Other	7	2%
4	Groups	3	1%

Figure 12. Bivariate EDA: Market Segment vs.

Cluster	0	1	2	3
Age	41.849560	47.733523	44.493595	60.112631
DaysSinceCreation	546.711906	1046.839486	366.185067	584.165364
AverageLeadTime	114.307218	55.913324	44.848759	206.662526
LodgingRevenue	1067.918461	318.851460	344.242532	347.332668
OtherRevenue	227.190697	74.245210	64.651526	96.208072
BookingsCheckedIn	1.115824	1.063502	1.060391	1.034389
PersonsNights	12.297964	5.821934	5.270441	6.390950
SRHighFloor	0.059840	0.044172	0.039301	0.040230
SRCrib	0.022517	0.008668	0.013344	0.003538
SRKingSizeBed	0.401450	0.351200	0.364495	0.365117
SRTwinBed	0.175663	0.117937	0.133598	0.221884
SRQuietRoom	0.133560	0.079930	0.083266	0.103743
BookingSuccessRate	0.999618	0.999413	0.999507	0.999918
TotalSpecialRequests	0.809685	0.613041	0.644427	0.746030
RevenuePerPersonNight	143.335807	79.877961	92.987167	76.162170
SpendingPerBooking	1232.665355	378.901472	393.700309	434.490763
BookingsNotAttended	0.002468	0.003010	0.001703	0.000165
Origin_North_America	0.055830	0.046747	0.081643	0.084492
Origin_Other_Europe	0.815238	0.719700	0.700316	0.787988
Origin_Others	0.050740	0.064118	0.059945	0.043603
Origin_Portugal	0.034547	0.131537	0.118876	0.044673
Origin_South_America	0.043646	0.037898	0.039220	0.039243
DistributionChannel_Corporate	0.015577	0.035432	0.035813	0.010777
DistributionChannel_Direct	0.197101	0.122652	0.166207	0.079720
DistributionChannel_GDS Systems	0.004010	0.006165	0.015290	0.000494
DistributionChannel_Travel Agent/Operator	0.783313	0.835751	0.782690	0.909009

Figure 13. mean of the variables by cluster