NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case1: Hotel Customer Segmentation

Ana Miguel Monteiro, number: 20221645

Ana Rita Viseu, number: 20220703

Miguel Cruz, number: 20221391

Rodrigo Brigham, number: 20221607

Sara Galguinho, number: 20220682

Group G

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

Index

1. EXECUTIVE SUMMARY

This project aims to develop an appropriate customer segmentation for a Portuguese hotel located in Lisbon, hotel H. By finding the relevant customer segments related to this business, we were able to gain and communicate insights on the demographics, spending habits and usual preferences of each customer group, which in turn allowed us to suggest business applications and targeted marketing approaches.

We followed the well-known process model CRISP-DM, starting by gaining a broader understanding of the business. In order to do this, we explored Hotel H's background, clarified business objectives and defined success criteria for the problem at hand, which will allow us to measure the performance of our findings and suggestions, should they be implemented. We also assessed the kind of resources that were available, as well as risks/contingencies and costs/benefits that this project entailed. As a final step, before starting to work on the data provided by the hotel, we determined what were our data mining goals - objectives in technical terms - for this customer segmentation project.

After gaining a thorough understanding of the business and establishing clear goals and success criteria, we proceeded with data preparation and exploration. This involved cleaning and organizing the data provided by Hotel H, as well as identifying and addressing any missing values or outliers. We also conducted an exploratory data analysis to gain insights into the distribution and correlation of the data, which helped us identify relevant features for the customer segmentation model.

Using a combination of unsupervised machine learning techniques, we developed a customer segmentation model that effectively grouped Hotel H's customers into distinct segments based on their spending habits, demographics, and preferences. We then analyzed the results of the model and identified four key customer segments. For each segment, we provided detailed insights and recommendations on targeted marketing strategies, business applications, and potential revenue opportunities.

Overall, this customer segmentation project provided valuable insights and recommendations to Hotel H, allowing them to better understand their customers and tailor their marketing efforts accordingly. By following the CRISP-DM model, we were able to effectively navigate the data mining process and deliver actionable results that can drive business growth and improve the overall guest experience.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Located in Lisbon, hotel H is part of the independent hotel chain C. Up until 2015, this chain operated 4 hotels, but, with the acquisition of new hotels, the board decided to invest more in marketing.

In 2018, the hotel chain created a marketing department and hired a new marketing director, A, who recognized that the current customer segmentation used by Hotel H was not at all useful. The hotel uses a hospitality standard market segmentation based only on the origin of the customer - the sales origin. It does not reflect geographic, demographic, or behavioral characteristics, such as country of origin, age, or number of check-ins, respectively, which are present in their customers' dataset. It's difficult for the marketing department to define a strategy to reach new clients and continue to engage current customers without proper customer segmentation.

Thus, working with the key stakeholders of the hotel, we aim to develop an adequate customer segmentation that takes into account a broader set of the clients' characteristics. This will help the organization make better strategic choices about product "creation", price definitions, and other marketing tasks such as advertising and promotions according to the different channels and customer groups. This segmentation, along with the derived insights, will be especially useful to the marketing employees. Still, Hotel H will be able to better understand their customers, which in turn brings benefits to the whole organization.

2.2. BUSINESS OBJECTIVES

The central project objective is to gain a deeper understanding of the diverse needs and preferences of Hotel H's customers by analysing their general behaviours and traits, dividing them into distinct groups based on their similar geographic, demographic, and behavioural characteristics. This way, the marketing department will be able to develop efficient marketing campaigns directed at different target audiences, attracting new customers and adapting its services and amenities to better meet the needs of each group. Furthermore, by exploring patterns in customer data, the hotel can better predict what their clients are likely to do in the future, as well as identify the key drivers of customer behaviour.

Some of the expected benefits of this project include:

- Stronger relationships with customers and increased customer satisfaction;
- Attract new customers;
- Improved customer loyalty;
- Improved marketing strategies;
- Increased revenues (and at a lower cost);
- Improved operational efficiency.

2.3. BUSINESS SUCCESS CRITERIA

To evaluate the success of a segmentation project, we need to establish clear and objective criteria that allow us to measure its performance and impact. In this case, we consider it important to define several success criteria that can be assessed by the marketing director and the relevant stakeholders, of which:

- Customer segments that allow a good distinction of the customer base;
- Customer segments that allow for strong targeted marketing strategies;
- Increase customer loyalty by 1% (considering the hotel industry in general has a low customer loyalty);
- Increase customer lifetime value by 5%;
- Increase occupancy rate by 4%;
- Increase total revenue by 2.5%.

2.4. SITUATION ASSESSMENT

The assessment of the available resources is essential for better project planning and increasing efficiency.

For the development of this project, we used the file "Case1_HotelCustomerSegmentation.csv", a dataset with 28 features/columns and 111733 customers/rows provided by the hotel with the reservations made until 2018 (year of extraction of the dataset). These data are fundamental for identifying customer patterns and segments.

To prepare the data for analysis, we used software tools such as Python and some of its classic libraries: *Pandas*, *Scikit-learn*, *NumPy*, *ydata_profiling*, *math*, *pycountry_convert*, *SciPy* and *yellowbrick*. For data visualization, which is essential to communicate insights and results from our customer segmentation, we also used libraries like *Seaborn* and *Matplotlib*.

Hotel H's employees are an important resource as well, as they have essential skills and knowledge to develop marketing campaigns and perform personalized service delivery to customers. The project team should be familiar with terms related to the hotel market to ensure clear and effective communication.

Besides resource assessment, it's important to mention the multiple assumptions that we made about the provided database:

- We only considered the customers that checked-in at the hotel at least once: the customer retention rate of hotels is very low when compared to other industries and our focus is on segmenting the customers that have actually used Hotel H's accommodations.
- In order to calculate the company's KPIs (key performance indicators), we assumed a period of time beginning at the date of the last registered customer.
- When the personal document identification number of a customer wasn't provided, we assumed this customer was registered under a generic document ID hash (this assumption was made because there were a large number of records, nearly 2900, in the provided database under the same document ID).
- When there were multiple records with the same document ID (with the exception of the ones in the previous bullet point), we assumed they corresponded to the same customer but to different bookings, so we "merged" those records into one.

For the realization of the project, certain contingencies could be faced, such as resource constraints (limited computing resources) and time constraints. There are also some risks worth considering, as well as what was done to mitigate them:

- Data quality issues: data that is incomplete, inconsistent and/or that contains errors. To minimize this problem, we ensured the data was clean, complete, and accurate before beginning our analysis.
- Unrepresentative samples: the customer sample used for segmentation analysis may not be representative of the whole population. To help this potential problem, it's good to make sure the data is diverse and representative of the customer base.
- Problems with the implementation of some algorithms. To mitigate this risk, we made sure a thorough data preparation was made.
- The insights and recommendations generated by the customer segmentation analysis may not be accepted or adopted by the hotel. To ensure this doesn't happen, our findings should be communicated effectively and should generate interest for the stakeholders and marketing employees.

Finally, this project may involve direct and indirect costs, like the purchase of additional software, development time, implementation of marketing campaigns and the need to offer training to employees. However, potential benefits include increased revenue, reduced operating costs, more qualified staff and increased customer satisfaction, loyalty and lifetime value.

2.5. DETERMINE DATA MINING GOALS

With the aim of evaluating our customer segmentation solutions and choose the best one, it's important to determine the kind of data mining results we are expecting, meaning what are the technical objectives of this project.

Considering the already mentioned business objectives, the general goal of this project, and the available data provided by the hotel, we defined the following data mining goals:

- A solution with 2 to 6 clusters (customer segments): Having too many clusters would make it harder to interpret the results and draw meaningful insights from them as well as develop general marketing approaches. On the other hand, having too few clusters would fail to clearly distinguish customers and provide a comprehensive understanding of the customer base.
- High distortion score: We intend to find the tradeoff between the silhouette score of the clustering solution and an appropriate number of clusters.
- Correlation between Cardinality and Magnitude: We want the cardinality of the clusters to be correlated to their magnitude, since that indicates no major anomalies exist.
- High silhouette score: We aim to have a high silhouette score, which measures the degree of similarity between each data point in a cluster and the points in its neighboring clusters.
- Well distinct clusters with a good discrimination ability: By examining the average of the variables in each cluster when compared to the population mean we can draw insights and conclusions about each customer segment. Having well separated and homogenous clusters - which can be analyzed by calculating the distance or dissimilarity between clusters and within each cluster, respectively - allows an effective and reliable interpretation of each customer group.

3. METHODOLOGY

To present a solution for the proposed problem, we followed the CRISP-DM methodology, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The first phase (developed previously in chapter 2) involves defining the problem and project objectives. In the second phase, data is collected and analyzed to gain a better understanding of the available information. The third phase involves cleaning, transforming, and integrating the data to prepare it for modeling. In the modeling phase, various techniques are used to build and test predictive models. The fifth phase is evaluation, which involves assessing the quality and effectiveness of the models. Finally, in the deployment phase, the chosen model is put into action and monitored for performance. It's important to note that the sequence of these phases isn't rigid, and we moved back and forth between steps when necessary.

3.1. DATA UNDERSTANDING

Our first step was to perform an initial exploration of the available information and try to identify data patterns, structural problems and draw early conclusions. After uploading the provided data file into a pandas data frame, we started by becoming familiar with our customer data and identifying the existing features.

	FEATURES	DESCRIPTION
Sociodemographic	Nationality	Nationality of the customer in ISO 3166-1 (Alpha 3) format
	Age	Age of the customer
	NameHash	Hash of the customer's name
	DocIDHash	Hash of the customer's personal document identification number
Behavioral	DaysSinceCreation	Number of elapsed days since the customer was created
	AverageLeadTime	Average number of days before arrival date the customer makes bookings
	LodgingRevenue	Total amount of lodging revenue paid by the customer so far
	OtherRevenue	Total amount of other revenue (e.g., food & beverage, spa) paid by the customer so far
	BookingsCanceled	Number of bookings the customer made but subsequently canceled
	BookingsNoShowed	Number of bookings the customer made but subsequently made a "no-show"
	BookingsCheckedIn	Number of bookings the customer made, which actually ended up staying
	PersonsNights	Total person/nights the customer has stayed at the hotel so far Person/Nights = (sum of People in each booking) x (number of Nights of the booking)
	RoomNights	Total of room/nights the customer has stayed at the hotel so far Room/Nights = (number of Rooms of each booking) x (number of Nights of the booking)
	DistributionChannel	Distribution channel normally used by the customer to make bookings at the hotel
	MarketSegment	Current market segment of the customer
	SRHighFloor	Indication if the customer usually asks for a room in a higher floor (0: No, 1: Yes)
	SRLowFloor	Indication if the customer usually asks for a room in a lower floor (0: No, 1: Yes)
	SRAccessibleRoom	Indication if the customer usually asks for an accessible room (0: No, 1: Yes)
	SRMediumFloor	Indication if the customer usually asks for a room in a middle floor (0: No, 1: Yes)
	SRBathtub	Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes)
	SRShower	Indication if the customer usually asks for a room with a shower (0: No, 1: Yes)
	SRCrib	Indication if the customer usually asks for a crib (0: No, 1: Yes)
	SRKingSizeBed	Indication if the customer usually asks for a room with a king size bed (0: No, 1: Yes)
	SRTwinBed	Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes)
	SRNearElevator	Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes)
	SRAwayFromElevator	Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes)
	SRNoAlcoholInMiniBar	Indication if the customer usually asks for a room with no alcohol in the mini bar (0: No, 1: Yes)
	SRQuietRoom	Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes)

Table 3.1 – Features/Columns in the provided dataset

Afterwards, we used descriptive statistics and data visualization tools (like pandas profiling) to generate an overall report of the data and discover inconsistencies and potential insights such as (but not limited to):

- Duplicated rows (rows with the same values for every variable);
- A lot of rows with the same value in *DocIDHash* (the same customers were introduced in the database as a new client more than once);
- Variable *Nationality* has a high cardinality (a lot of distinct values), which could make distinguishing customers by their nationality more difficult;
- Most binary variables are highly imbalanced (meaning they have either a high percentage of 0's or 1's), more specifically *SRHighFloor*, *SRLowFloor*, *SRAccessibleRoom*, *SRMediumFloor*, *SRBathtub*, *SRShower*, *SRCrib*, *SRNearElevator*, *SRAwayFromElevator*, *SRNoAlcoholInMiniBar* and *SRQuietRoom*, which could also complicate the distinction of customer groups;
- A significant percentage of customers (29,7%) never checked in at the hotel (variable *BookingsCheckedIn* has a lot of 0's);
- Missing values in variables *DocIDHash* and *Age*;
- "Travel Agent/Operator" corresponds to 81,46% of the observations in the variable *DistributionChannel*;
- Most of the variables contain extremely severe outliers, which lead to skewed distributions.

3.2. DATA PREPARATION

Regarding data preparation, we started by removing duplicated rows which lead to the removal of 0,1% of the original data.

3.2.1. Inconsistencies/Incoherences

To begin this process, and as previously mentioned, we only considered customers that had checked-in at hotel H at least once, so we removed the rows that had no check-ins done. We also removed rows that had missing values in the variable *DocIDHash*. As stated before, there were customers identified multiple times (multiple rows with the same *DocIDHash*) and we merged all the information contained in those lines into a single one, thus retaining the integrity of the data. The 2893 rows that have the same *DocIDHash* were removed.

We also deleted rows where we have a higher value of *RoomNights* when compared to *PersonsNights*, where *BookingsCheckedIn* is higher than *PersonsNights* and where *BookingsCheckedIn* is higher than *RoomNights*. This aims to guarantee data quality and coherence.

After deleting and merging observations, we kept around 65% of the original data.

In the variable *Age*, we turned the values above 100 to *NumPy Nans* because it seemed unreasonable to have customers older than 100 years old, even taking into consideration the *DaysSinceCreation* variable. We did the same for customers with negative ages and for customers that had an age below their *DaysSinceCreation* in years. For the negative values in *AverageLeadTime*, we did the same thing. This was made in order to later impute those missing values, instead of removing these observations.

3.2.2. Data Imputation

We only had missing values in the *Age* and *AverageLeadTime* variables. For imputation, we used the Iterative Imputer with a Linear Regression.

3.2.3. Feature Creation

To conduct a better cluster analysis, we created new features based on the combination of the original ones. These new variables are possibly more relevant for the problem at hand and allow a more effective segmentation.

VARIABLE	DEFINITION
TotalRevenue	*LodgingRevenue + OtherRevenue*
AverageRevenueBooking	*TotalRevenue / BookingsCheckedIn*
AverageRoomNights	*RoomNights / BookingsCheckedIn*
AveragePersonsNights	*PersonsNights / BookingsCheckedIn*
RatioPersonsPerRoom	*PersonsNights / RoomNights*
CancellationRatio	*BookingsCanceled / (BookingsCheckedIn + BookingsNoShowed + BookingsCanceled)*
AverageRoomRate	*LodgingRevenue / BookingsCheckedIn*

Table 3.2 – New Features Created

3.2.4. Feature Encoding and Binning

During outlier detection, we realized many of the observations identified as outliers by the algorithms we used weren't actually outliers, leading us to remove important data. Therefore, we tried applying discretization (binning) to some of the variables. Binning not only helps dealing with outliers, but also improves simplification and interpretability. It is also useful when variables have a high cardinality.

Since the variable *Nationalities* had too many distinct values, we kept all nationalities that had a representation above 5% on the dataset and encoded all others as "Other".

Then, we created the following age groups: *<20*, *20-29*, *30-39*, *40-49*, *50-59* and *>=60*. With this, we binned the variable *Age* accordingly. We proceeded similarly for the variables *DaysSinceCreation*, *AverageLeadTime*, *BookingsCheckedIn*, *BookingsCanceled* and *BookingsNoShowed*. For *LodgingRevenue* and *OtherRevenue* we binned the data according to the quantiles. After, we encoded these variables using *OneHotEncoder*.

For the nationalities, we created a new variable, *Continent*, with three different values: "Portugal", "Europe" (except Portugal) and "OutsideEurope".

Finally, we renamed all the necessary variables appropriately.

3.2.5. Company's KPIs and Analysis of the Customers

In order to measure the current overall performance of Hotel H, we calculated the company's hotel standard industry KPIs, like the average revenue per client, average revenue per year and average revenue per booking.

Considering the provided dataset is now “clean” and we have the newly created variables to help in our analysis, we performed an exploratory analysis to provide a general description of the customer database.

Through this analysis we determined that Hotel H’s customers have a mean age of 48 years, with the most significant group being 50-59 (Figure 3.1, Appendix), although they are all very similar until ages below 29. Around 83% of the clients are from Europe, most of them being from France, Germany, Great Britain, Portugal, and Spain, our Top 5 nationalities (Figure 3., Appendix). In (Figure 31.3, Appendix) a distribution of the revenue by the top five nationalities can be seen. When it comes to bookings, 95% of the clients have only checked-in once, around 99.99% have never cancelled and more than 99.99% have never made a no-show. The most requested services by Hotel H’s customers are king sized beds and twin beds (Figure 3.4, Appendix).

3.2.6. Feature Selection

With the aim of not having irrelevant or redundant features and retain only the ones that effectively distinguish the customers, we chose to eliminate some of the following features (Table 3.3).

VARIABLE	REASON FOR REMOVAL
NameHash DocIDHash	These variables do not add any relevant information to the modelling
SRLowFloor SRAccessibleRoom SRMediumFloor SRBathtub SRShower SRNearElevator SRAwayFromElevator SRNoAlcoholInMiniBar	These variables are extremely imbalanced between 0's and 1's, they lack value for accurately segmenting the customers and also lack of correlation with any other variables
MarketSegment	We are creating a new segment because the previous segmenting strategy was not adequate. Therefore, we shouldn't use this variable considering that we would have to segment the customers according to it
TotalRevenue AverageRevenueBooking AverageRoomNights AveragePersonsNights RatioPersonsPerRoom CancellationRatio AverageRoomRate	These variables lack value for accurately segmenting the customers and have strong correlations to some of the existing variables

Table 3.3 – Reason for the Removal of each Feature

3.2.7. Feature Scaling

It is essential to scale our data when applying distance-based algorithms, like K-means, so we used three different techniques – MinMax Scaler, Standard Scaler and Robust Scaler – to test potentially different solutions. In our final solution, we applied the MinMax Scaler.

3.2.8. PCA

As a final step, we performed a Principal Component Analysis in order to understand which variables were essential to keep and had a higher explanatory power. We decided to keep 23 principal components which explain 99% of our dataset’s variance.

3.3. MODELING

To determine the relevant customer segments, we utilized a K-means clustering algorithm. However, before implementing the algorithm itself, we needed to decide on the number of clusters that would best fit our model. To achieve this, we used two evaluation metrics: the distortion score and the silhouette score. We employed the elbow method, which involves plotting the distortion score for various values of k (number of clusters) and identifying the "elbow" point where the change in distortion score begins to level off (Figure 3.5, Appendix). Using this method, backed by a higher silhouette score on that value for k (Figure 3.6, Appendix), we determined that 3 clusters would be the optimal choice.

We then trained a K-means model using the selected value of k=3, and generated visualizations to analyze our results. We used the Silhouette Visualizer (Figure 3.7) to further evaluate our clusters and obtain a better understanding of the quality of our model of how well it was able to group the data points into distinct clusters. This allowed us to identify any potential issues with our model, such as clusters that contained data points that were poorly matched to the other points in the cluster or clusters that contained overlapping data points. Overall, the Silhouette Visualizer provided us with a more nuanced evaluation of our clustering model and helped us to fine-tune our approach to obtain more accurate and meaningful customer segments.

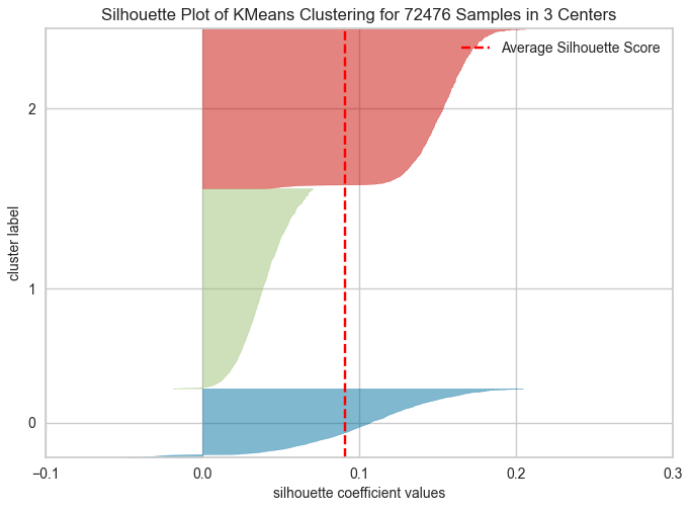

Figure 3.7 – Silhouette Visualizer

Upon analyzing the results, we discovered that our clusters contained an uneven proportion of customers, ranging from 16% to 47% of the available data. We also computed the weights that each variable held for each principal component and computed the centroids of our clusters for the most important variables.

To ensure that our model was robust, we tested it for different values of k, scaling techniques, and groups of features. After evaluating our model using various K-means evaluation metrics, we landed at a final solution that was both interpretable and optimal.

3.4. EVALUATION

The distortion score measures the sum of the squared distances between each data point and its assigned cluster centroid. A lower score indicates that the data points in each cluster are more tightly packed and better represented by their assigned centroid. Our final clustering model produced a distortion score of 225000, which suggests the model was able to effectively group the data points into three distinct clusters with minimal error.

We also evaluated our model using the silhouette score, which measures the degree of similarity between each data point in a cluster and the points in its neighboring clusters. Our silhouette score was calculated to be 0.09, indicating that our clusters were not very well-separated and that there may have been some overlap between the data points in adjacent clusters. This suggests that there may be room for improvement in our clustering model to better differentiate the data points and produce more well-defined clusters.

Moreover, we observed a positive correlation between the magnitude and cardinality of our clusters, which means that, as the number of data points in a cluster increased, so too did the total magnitude (or "weight") of the variables associated with that cluster, as we can see in the figure 3.8. This information can be useful in identifying which variables are most strongly associated with each cluster and can aid in developing targeted strategies for each segment.

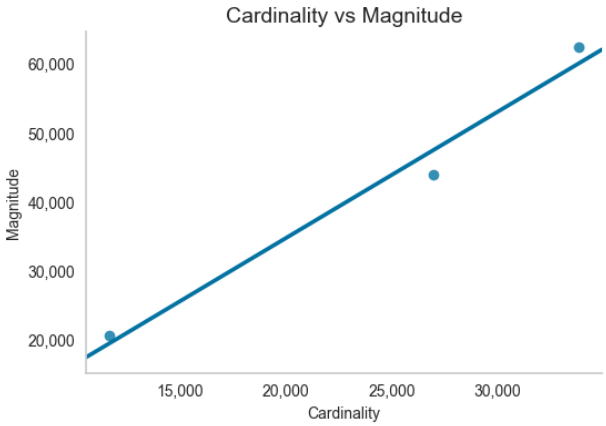

Figure 3.8 – Cardinality vs. Magnitude

4. RESULTS EVALUATION

Our final model effectively identified three distinct customer segments that allow for a good distinction of the customer base and provided detailed insights and recommendations for targeted marketing strategies that can be tailored to each segment.

With the respect to our data mining goals, we achieved several of the defined success criteria: we reached a solution with an appropriate number of segments, a good trade-off between a high distortion score and the number of clusters, a high silhouette score and a correlation between cardinality and magnitude. Finally, we have clusters with a good discrimination ability, which is essential to achieve the focal point of this project.

Although the model has not yet been applied in a real environment, it provides a strong foundation for future testing and refinement. By evaluating the model's performance metrics, such as the silhouette score, against new data and business success criteria, Hotel H can assess the effectiveness of the model in a real-world environment. Additionally, by monitoring the key performance indicators that we defined as success criteria - such as occupancy rate, customer loyalty, and total revenue - Hotel H can evaluate the impact of the model on business outcomes and make adjustments as needed. Overall, while further testing and refinement may be necessary, the developed customer segmentation model has the potential to meet the business objectives outlined for this project and promote growth for Hotel H.

Once we reached an optimal final solution, we conducted an evaluation and interpretation of each cluster:

"The Loyal Segment" (cluster 0):

- Cluster with the least number of customers (16%);
- The customers are distributed among the various nationalities;
- The main distribution channels are direct (85%) and corporate (11%);
- Most of the customers are above 30 years old (around 89%), and, compared to the other clusters, they have a smaller proportion of people above 60 and a higher proportion between 40-49;
- The total revenues have an asymmetric pattern: the cluster contains a bigger proportion of the worst 25% clients and the top 25%;
- In terms of average lead time, it tends to be smaller compared to the other clusters, where 76% makes their bookings 0-3 months ahead;
- This is the segment with the most loyal customers, where around 13% has done more than 1 check-in;
- More cancelations (0.7%) and no-shows (0.3%) were made compared to the other clusters;
- Consumers tend to be more recent when compared to the other clusters: around 59% had their profile created in the last 2 years, with 28% the last year;
- Usually ask more for a crib (2%) and less for a king sized bed (19%), twin bed (8%) and quiet room (4%) compared to the other clusters.

"The Typical Segment" (cluster 1):

- Cluster with the highest number of customers (47%);
- All customers are either Portuguese, French, Spanish, German or from Great Britain;
- The main distribution channel is Travel Agent (99%);
- Most of the customers are above 30 years old (88%) but, compared to the other clusters, there's a slightly higher proportion of people between 50-59 and above 60;
- When it comes to revenues, they tend to be slightly worse customers since the cluster has a 26% proportion of the worst 25% and a 21% proportion of the top 25%;
- In terms of average lead time, it tends to be higher compared to the other clusters, where 57% makes their bookings 0-3 months ahead, 23% 3-6 months and 19% 6 months and above;
- Compared to the other clusters, it has the highest proportion of customers that have only checked-in once (97%);

- Cancelations and no-shows were very rare (0.04% and 0.006%, respectively);
- Customers tend to be older in terms of when the profile was created when compared to the other clusters. Around 48% had their profile created in the last 2 years, with 22% the last year;
- Around 4% of customers asks for a high floor, 0.9% for a crib, 36% for a king size bed, 16% for a twin bed and 9% for a quiet room (values very similar to the next cluster, 2).

"The Outsider Segment" (cluster 2):

- Cluster with 37% of the customers of the hotel;
- None of the customers are from the 5 most present nationalities in the hotel (Germany, Spain, France, Great Britain and Portugal);
- They have Travel Agent as main distribution channel (98%);
- Most of the customers are above 30 years old (88%);
- When it comes to revenues, they tend to be slightly better customers since the cluster has a 22% proportion of the worst 25% and a 26% proportion of the top 25% customers;
- In terms of average lead time, 61% makes their bookings 0-3 months ahead, 26% 3-6 months and 13% 6 months and above;
- Around 3% have done more than one check-in;
- Cancelations and no-shows were also very rare (0.03% and 0.007%, respectively).
- Around 54% had their profile created in the last 2 years, with 25% the last year;
- Around the same values when it comes to high floor, crib, king size bed, twin bed and quiet room, compared to cluster 1;
- In general, considering most variables, cluster 2 tends to have values in the "middle" of the proportions of cluster 0 and 1, but tends more to the side of cluster 1.

5. DEPLOYMENT AND MAINTENANCE PLANS

Deploying a segmentation model into production requires a well-organized plan to ensure a smooth transition. The first step involves working with the IT department to assess the compatibility of the model with the existing infrastructure. The next step involves testing the model in a controlled environment before deploying it in production. The model's performance metrics must be validated before releasing it to the users to. The model must also undergo integration and system testing to ensure it is compatible with other existing systems, and the necessary changes are made.

After deployment, the model's performance must be monitored continuously to ensure that it is performing as expected. This type of analysis should be done regularly, for example, every 6 months, to always be aware of the changes in the customers. Users should look for changes between analysis. In addition, the production system must be regularly updated with new data to ensure the model remains relevant and effective. A team of data scientists must monitor and maintain the model to identify and correct any performance issues promptly. The team should also update the model periodically (for example, 6 months) to address any changes in the data or the business environment, such as new customers or market trends. From these insights, we will also be able to extract and potentiate new marketing strategies to keep us competitive.

6. CONCLUSIONS

To conclude, we defined marketing strategies for each segment.

For "The Loyal Segment", we have three main ideas: loyalty programs, personalized offers and early booking initiatives. As mentioned in the Accenture 2022 article, travel companies have a high loyalty dividend, compared to other consumer industries. Therefore, the development of a loyalty program, based on immediate rewards, flexibility and experiences is essential for the improvement of the customer retention rate. This way, we will keep our loyal customers engaged with a gamified system, integrating the three ideas: specific offers for each customer considering their own tastes and early booking initiatives to promote a higher lead time.

For "The Typical Segment", we want to focus on: targeted marketing, benefits to travel agents and discounts. Considering they come only from the 5 nationalities where we have more customers (Portugal, Spain, France, Germany and Great Britain), we could target them with specific advertisements adapted to their language and culture. As well as that, we could adapt the marketing channels to each country and actively measure the results to continuously improve our customers' network. In addition, we could provide benefits to travel agents that are able to find new customers from these countries. Finally, discounts could be applied in order to further engage with new customers and continue to grow our client base.

For "The Outsider Segment", our three main plans are: local tourism, luxury packages and hotel experience. We could find local partners to create culturally enriching programmes, enticing people from all around the world to visit and stay in hotel H. These programmes could be part of luxury packages, designed to provide the optimal experience in Lisbon. To complement, the hotel experience should be developed as to provide a window into the local culture.

With these strategies, we hope to achieve and surpass our business objectives and success criteria, boosting our understanding of the customers and redefining the path to new connections.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

To potentially improve our model for customer segmentation, there are several considerations to keep in mind. One way to enhance the model is to extract more variables. By incorporating more variables into the model, we can capture more nuances in customer behavior and preferences. This can be particularly useful for identifying new customer segments or fine-tuning existing ones.

Furthermore, historical booking data can provide valuable insights into seasonality, dynamic pricing, and sentiment analysis, which can help us better understand customer behavior and tailor our marketing efforts accordingly. Another important factor to consider is which hotels each customer checked into, as this can provide insights into travel patterns and preferences. Finally, incorporating guests' feedback into the model can help us understand customer sentiment and improve the overall customer experience.

In addition to these considerations, improving the silhouette score is also a crucial factor for model improvement. The silhouette score measures how well customer segmentation is separating different customer groups based on their similarities and differences. By improving the silhouette score, we can

increase the accuracy and effectiveness of the segmentation model. To achieve this, we may need to adjust the clustering algorithm, refine the feature selection process, or consider different evaluation metrics. Overall, these considerations for model improvement can help our hotel better understand their customers, tailor their marketing efforts, and improve the overall guest experience.

7. REFERENCES

Shopify Staff (2022). Understanding Average Customer Retention Rate by Industry. Retrieved March 3, 2022, from Understanding Average Customer Retention Rate by Industry (2023) - Shopify Canada

Accenture (2022). Travel Rewards that are truly rewarding (Loyalty programs that reward travelers). Retrieved March 7 from Travel Rewards That Are Truly Rewarding | Accenture

8. APPENDIX

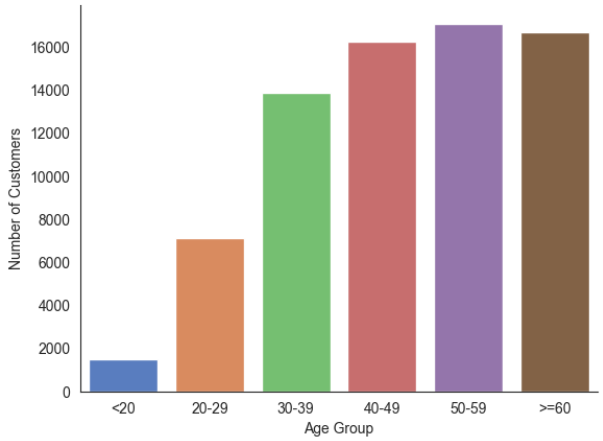

Figure 3.1 – Distribution of Customers by Age Group

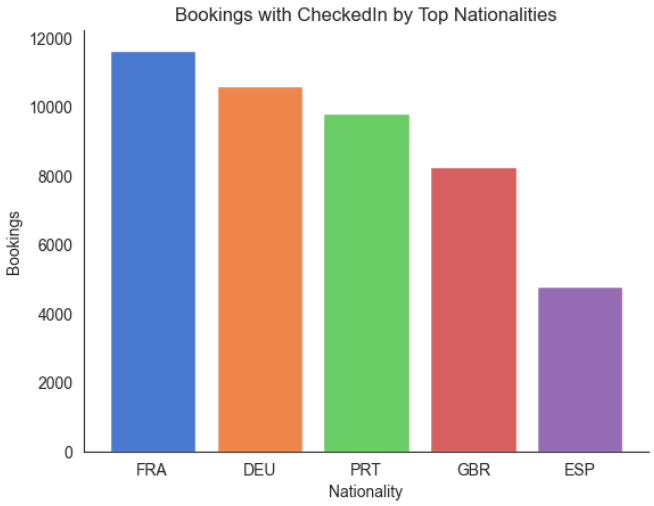

Figure 3.2 – Bookings Checked-In by Top Nationalities

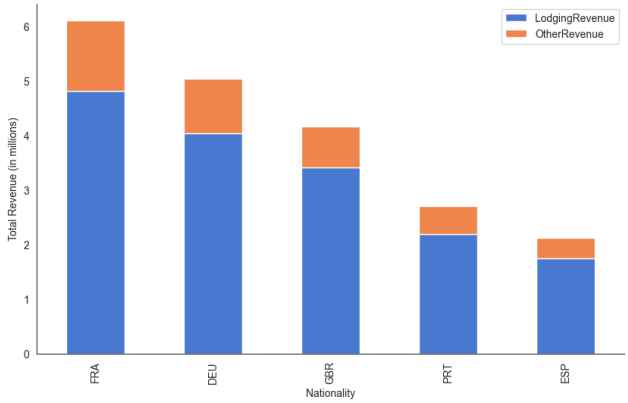

Figure 3.3 – Revenue by Top 5 Nationalities

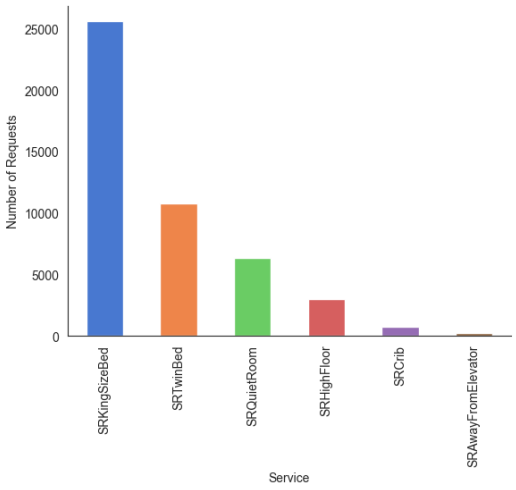

Figure 31.4 – Top 6 Requested Services

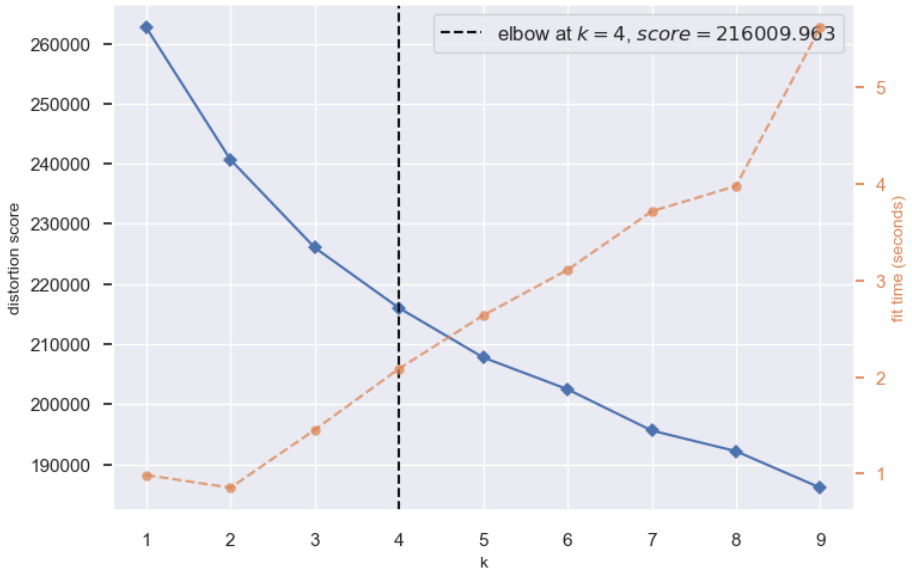

Figure 3.5 – Distortion Score Elbow Plot for K-Means Clustering

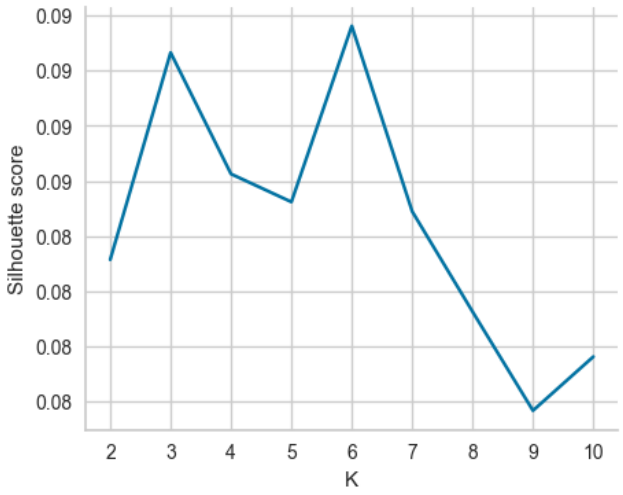

Figure 3.6 – Silhouette Method