MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel Customer Segmentation

Carolina, Costa, number: 20220715
João, Gameiro, number: 20221364
Martim, Santos, number: 20220540
Rodrigo, Silva, number: 20221360
Rúben, Serpa, number: 20221284

Group C: Young Talent Consulting Group

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

1. EXECUTIVE SUMMARY

Through the power of data, the Young Talent Consulting Group made an understanding of the existing customers and identified unique characteristics among different segmentations, discovering valuable insights for hotel H to thrive. The recommendations and timelines are provided in detail, at the end of the report.

In sum, there was a lack of appropriate analysis of data over the years, which led to Hotel H to not position itself in the most data-driven way possible and recently creating many difficulties for marketing manager A in defining suitable marketing strategies. The original segmentation only took into consideration one customer characteristic, which was sales origin. It didn't take into consideration other important characteristics. The team saw a golden opportunity and built a bridge between data and valuable information, by creating a more complex approach to customers' behaviours and characteristics.

By conducting a detailed analysis with the support of outlier detection algorithms, Principal Component Analysis (PCA) and implementing algorithms such as K-Means and K-Prototypes a new three customer segment label list was created in accordance with the findings and a list of business applications were meticulously thought out based on outcomes. Gen X Customers with Potential for Future Growth, baby boomer customers with high average lead time, and Baby Boomer Customers with High Revenue Contribution are the final three clusters analysed.

It is believed that by implementing the given recommendations and with appropriate model maintenance, Hotel H will be able to keep track of its customers origins and deliver the right strategies to acquire and retain more customers with much less resource and time spending involved.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

In 2018, the travel and tourism sector represented 6,1% of the overall Portuguese GDP (Travel and Tourism Account as a % of GDP, 2022). Being one of the major economic sectors in Portugal, it is crucial to have a good understanding of the market structure and customer composition in order to make good data driven decisions that will make our business thrive. During 2018, the hospitality segment accounted for 81,0% of the total guests, followed by local accommodation (15,6%) and tourism in rural areas and housing (3,4%) (Instituto Nacional de Estatística, I. P. [INE], 2019).

Hotel chain C operates various hotels, including hotel H, located in Lisbon, Portugal, a city that was responsible for 25,1% of Portugal's total stays, in 2018 (Instituto Nacional de Estatística, I. P. [INE], 2019). After 2015, Hotel chain C started their expansion from the four hotels chain structure with the acquisition of new hotels. As they expanded their business it was also important to maintain a business process that could hold that same growth.

According to Lisboa OFFICIAL Site (2020), Lisbon is a popular tourist destination with a thriving hotel scene. Visitors can choose from a variety of accommodations ranging from luxurious five-star hotels to budget-friendly hostels. Many hotels in Lisbon are in the city´s historic neighbourhoods, offering easy access to top attractions. Beyond accommodations, the city offers a vibrant tourist scene with a plethora of things to do, including visiting museums, exploring the city´s gastronomy, and taking a scenic tram ride.

2.2. BUSINESS OBJECTIVES

With such a wide range of possible customers, it is acknowledged the importance of a good segmentation that reflects, in a detailed way, customers characteristics, something that until the moment is not being made. In order to reach that it is necessary to: analyse customer data and booking patterns to identify the most profitable customer segments; maximize revenue for Hotel H by targeting the most profitable customer segments and optimizing pricing and marketing strategies accordingly and define the new segmentation strategies and implement them within the next six months and evaluate its effectiveness after one year.

It all starts with these simple questions: "Who are our current customers?"; "What patterns do they have in common?"; "How can we maintain a relationship with them?"; "Who are our potential customers?"; "How can we reach them?".

2.3. BUSINESS SUCCESS CRITERIA

The business success criteria is going to depend on the ability to define and segment, in a clear and concise way, hotel H customers. By doing so, Hotel H will be able to proper allocate advertising costs and offer the appropriate services at the right price.

With the purpose of achieving success, an extensive data exploration analysis is going to be made in hopes of better understanding which factors have a stronger weight in customer behaviour and grouping. To evaluate clusters and clustering performance, some methods are going to be applied, including the elbow method and a dendrogram plus the use of silhouette score to verify the existence of well-defined clusters, an important point regarding the uniqueness of each segment.

2.4. Situation assessment

In terms of personnel, Hotel chain C has recently invested in the marketing department, hiring marketing expert A in hopes to help deal with the lack of customer understanding. Although A has a strong strategy background, it lacks on the analysis expertise needed to create the basis of where every action is going to take form.

To make this project doable, a dataset was provided with geographic, demographic, and behavioural information regarding its customers.

The project itself doesn't have any major risks associated with it but some concerns may possibly arise regarding consultant's time needed to properly analyse the provided dataset as opposed to the proposed scheduled delivery.

Young Talent Consulting Group values data privacy. All personal data will be handled with care in hopes of preventing data leakage of any format.

2.5. Determine Data Mining goals

This project is going to follow a sequence of phases, industry known as CRISP-DM. To fulfil the proposed business objectives, three main technical checkpoints exist:

1. Capture active customers or customers that have had a transaction history with the company.
2. Understand which types of information are more useful to our analysis.
3. Use Clustering algorithms to try to unravel trends and patterns in the data.

3. METHODOLOGY

3.1. Data understanding

According to Kurgen and Musilek (2006), the step of data understanding overview involves the initial exploration and collection of data to establish a foundation for further analysis. The primary objective of this step is to detect any inconsistencies or problems in the data that must be resolved before proceeding to the next phase of the CRISP-DM process.

The analysis was initiated by importing all the necessary libraries and the dataset "Case1_HotelCustomerSegmentation.csv" which was provided by the independent hotel chain. Subsequently, the size of the dataset was verified, having 111733 rows and 29 columns, and a backup was created to preserve the original version unaltered.

During the process of exploring the available data, the subsequent step involved defining the ID as the index of the data frame, as well as verifying the data types, detecting missing values, and identifying any duplicated records for each variable. Moving on to the checking of missing values, after replacing by NaN the empty cells, it was possible to conclude that *Age* and *DocIDHash* variables had missing values, 4172 and 1001 respectively, which would be filled later in feature engineering, reducing bias. When looking for duplicates, it was discovered that there were 8252 duplicate records in the *DocIDHash* variable, which is not possible as each client can only have one ID. Additionally, the duplicated method was also used to identify 111 duplicate records in other variables of the dataset.

Upon conducting an analysis of the descriptive statistics of each variable, it was observed that the variable Age contained negative values as well as values exceeding 100, which appears to be suspicious. Regarding the variable named *DaysSinceCreation*, it was observed that it has a high standard deviation. Additionally, there is a noticeable difference between the minimum and the 25th percentile, as well as the maximum and the 75th percentile values of this variable. It was found that the variable *DocIDHash* contains only 103480 unique values, suggesting the possibility of repeated customers. In relation to the variable *AverageLeadTime*, it was observed that there are negative values present in the dataset. Additionally, a high standard deviation was found, along with a significant difference of 493 days between the 75th percentile and the maximum value. The variables *LodgingRevenue* and *OtherRevenue* exhibit a high standard deviation upon analysis. In addition, it was observed that the difference between the 75th percentile and the maximum on the variables *LodgingRevenue*, *OtherRevenue*, and *BookingsCheckedIn* are significant. All of the aforementioned variables, namely *DaysSinceCreation*, *LodgingRevenue*, *OtherRevenue*, *BookingsCanceled*, *BookingsNoShowed*, *BookingsCheckedIn*, *PersonNights*, and *RoomNights*, were verified and found to have no negative values. For the categorical variables, it is possible to check the corresponding number of classes. It was also checked the uniqueness of the *Nationality* of customers to see if there was any ISO code that was not in accordance with the ISO 3166-1 (Alpha 3) format.

Turning to data visualization, the dataset has been divided into metric and non-metric features as they require different visualization methods. Furthermore, the non-metric features include *NameHash, DocIDHash, Nationality, DistributionChannel, MarketSegment, SRHighFloor, SRLowFloor, SRAccessibleRoom, SRMediumFloor, SRBathtub, SRShower, SRCrib, SRKingSizeBed, SRTwinBed, SRNearElevator, SRAwayFromElevator, SRNoAlcoholInMiniBar,* and *SRQuietRoom*.

To explore the metric characteristics of the dataset, drivers such as histograms and boxplots were used. These drivers were used to visualize distributions and identify potential outliers. Upon analysis, it was concluded that *Age* appeared to have outliers on both the lower and higher ends, while all other variables (apart from *DaysSinceCreation*) appeared to have outliers on the higher end.

Additionally, bar charts were created for each non-metric characteristic, which did not reveal any inconsistencies. Furthermore, the visualizations indicated that most consumers are reached through a Travel Agent/Operator, do not request alcohol in the mini bar, and have no preference for the floor level they intend to sleep on. However, some customers showed a preference for making a reservation that includes a king-sized bed.

Afterwards, an analysis of the relationships between the characteristics was conducted by examining the heatmap of a correlation matrix. Based on the matrix, it can be concluded that there is only one significant correlation, namely between *PersonsNights* and *RoomNights*. However, it should be noted that these findings may be impacted by the presence of missing values and outliers.

3.2. DATA PREPARATION

3.2.1. Treating Inconsistencies

During this stage, a data coherence check was necessary. Upon it, nine coherence issues emerged. According to the study by Ramezani and Farsijani (2013), in which they used data mining techniques to identify the factors affecting customer loyalty in the hospitality industry, customers who did not check-in were less likely to become loyal customers compared to those who did check-in. To further support the idea that customers who do not check-in are less likely to be retained by the hotel, Alipour, Ahani, and Ebrahimi (2015) conducted a study which revealed that customers with a history of no-shows are also less likely to be retained by the hotel. Hence, it can be inferred that not doing a check-in may have a negative impact on customer acquisition in the hotel industry as well as not having a transaction history will not add value to our customer acquisition strategy creation. To deal with this issue, all customers that did not present any bookings were separated from the main database since they don't represent active customers. These customers summed to a total of 33.197. The coherence checks continued for the 78.536 remaining customers. Having a short look at the customers that haven't made a reservation yet, a filter was applied to this section of the database, by customers that were registered for less than a year. It was found a total of 16.314 potential customers that Hotel H could try to get to know better. For instance, most of these customers are registered through a travel agent or operator meaning they were probably cases of customers interested in booking a room in Lisbon and their data information was sent by these agencies to Hotel H.

As already noticed during the previous phase, we cannot have customers with a negative age or average lead time. There was a total of twenty five clients, twelve with negative *Age* and another thirteen with negative *AverageLeadTime*. After analysing each situation, it was possible to conclude that it is probably an error in the database, therefore it was decided to substitute the respective age values by an imputer during the next phase and remove the incoherent *AverageLeadTime* values. This resolution was also applied when analysing clients in which *DaysSinceCreation* was higher than *Age*, finding eight customers under this condition.

It is unrealistic to have a larger number of check-ins than the length of the stay. At best, these are equal. There was a total of five and six cases where customers had *BookingsCheckedIn* greater than *RoomNights* or *PersonsNights*, respectively. The rows in question were removed.

When customers have checked-in, it is important to check for the existence of three components: *Lodging Revenue, Person Nights* and *Room Nights*. During the check, it was found 554 rows of customers that did not have any Lodging expenditures even though they had spent at least one night at the hotel, so they were removed from the dataset and considered to be either errors or staff nights offered by the hotel and therefore not useful for clustering purposes. The other two components were coherent except for one situation. Considering that each Room can have one or more persons, it means that *RoomNights* can never be higher than *PersonNights*. It is unreasonable to book a room for no one to stay in. There were fourteen incoherent situations. To deal with this, it was decided to switch the columns in these observations.

3.2.2. Treating Duplicates

After treating inconsistencies, there were still fifteen duplicate values on the dataset compared to the initial 111 duplicate records. The observations were removed, and the focus moved to understanding the existence of duplicate values on the *DocIDHash* column. At this point there are still 5642 rows with duplicated *DocIDHash*. First, it is important to understand how many times each *DocIDHash* is repeated, because it is related to the database software changes. It is possible to check that one *DocIDHash* is repeated 2742 times, therefore, this should be treated separately from the other cases, as it is an absurd value. In this case, an aggregation was made were the value of *LodgingRevenue*, *OtherRevenue*, *BookingsCanceled*. *BookingsNoShowed*, *PersonsNights* and *RoomNights* were going to be the max value present in these rows. For the rest of the cases, the columns mentioned above use as aggregation metric the sum metric, considering the duplicates to be cases where the customer made different reservations in different moments in time and the registration of those was made separated.

3.2.3. Treating Missing Values

Missing values are a common problem in data analysis, and they can have a significant impact on the accuracy and quality of the results obtained. Therefore, it is crucial to address missing data appropriately to ensure that the dataset is suitable for modelling. As highlighted by Little and Rubin (2019), handling missing data is an essential step in the data preparation phase, where the primary objective is to convert raw data into a format that can be used for analysis. When checking for missing data values, there was a total of 232 cases where age was missing. It was decided to fill those values with a measure of central tendency, more specifically using the median.

3.2.4. Treating Outliers

Outliers are data points that deviate significantly from most of the data. In the context of hotel business segmentation, outliers can provide valuable insights into customer behaviour and preferences that are not represented by most of the data. Identifying and understanding these outliers can help hotels create more targeted marketing campaigns, improve customer satisfaction, and ultimately increase revenue. However, outliers can also distort statistical analyses and models if not properly handled. Therefore, it is important for Hotel H to carefully consider the role of outliers in their segmentation analysis.

The outliers on the data were analysed and brainstormed. It is important to note that the decision on how to handle them must be very carefully considered, as it might cause a big impact on the final clustering solution. Two different approaches were taken: DBSCAN and the use of manual approach.

According to Ester et al. (1996), DBSCAN is a density-based clustering algorithm that can be utilized for identifying outliers and is especially beneficial in detecting outliers in high-dimensional datasets that exhibit complex structures. Due to the high-dimensionality characteristic of the dataset this algorithm ends up being a great outlier detecting tool. To put it in practice, two parameters need to be defined: minimum points and the maximum distance between two points. According to Sander et al. (1998), the minimum points can be calculated following a formula where minimum points corresponded to twice the number of dimensions of a dataset. Since the dataset has ten metric features, the minimum points imputed was 20 and for the value of the maximum distance also known as epsilon, an elbow method was applied and then plotted to find the ideal value. The obtained plot showed that the ideal epsilon value was around 0,16.

Manually, customers with ages below 18 and above 89 were removed. There rows represent rooms that were booked by a family member, but the room name was assigned to one of the family kids for ages below than 18. For the other part, it was considered that these customers were exceptionally rare cases that could skew our data, so they were also filtered out. It was also filtered *Lodging Revenue* above 4000 euros, *Other Revenue* above 1400 euros, *Person Nights* above 29 nights and finally, *Room Night*s above 15 nights.

By removing those outliers, a small portion of the client's representation, who do not fall on the majority, is sacrificed, but on the other hand it is guaranteed a more stable performance on the output.

3.2.5. Feature Engineering

To promote useful and relevant information, the feature engineering step was performed where the variables related to the type of room booked were combined into a single variable named Room Type (*SRTwinBed*, *SRCrib*, and *SRKingSizeBed*). Similarly, the variables *SRHighFloor*, *SRLowFloor*, and *SRMediumFloor* were combined into a single variable named *FloorPreference*, and the remaining variables related to the specifics of the room were combined into a single variable named RoomPreferences (*SRBathtub*, *SRShower*, *SRNearElevator*, *SRAwayFromElevator*, *SRNoAlcoholInMiniBar*, and *SRQuietRoom*). Multiple features were extracted, such as Total Revenue (which is the sum of *LodgingRevenue* and *OtherRevenue*), Booking Rate (which represents the percentage of bookings that were not canceled or no-showed), Room Rate per customer (which is obtained by dividing the total revenue generated by the number of RoomNights to obtain the revenue per room), Revenue per Person (which is obtained by dividing the total revenue generated by the number of PersonsNights to obtain the revenue per person) and the Revenue Per Stay which is the average revenue per booking for each customer. This approach was adopted based on the guidelines provided by Kelleher and Tierney (2018) for feature engineering, which involves combining, selecting, and transforming raw data to derive new, more informative features that can enhance the performance of machine learning models.

According to Huang and Wan's study in 2019, binning is a method that can simplify data and reduce noise by converting continuous variables into categorical variables, which can facilitate the identification of relationships between variables. Therefore, this process was applied to the variables *Age* and *AverageLeadTime*. The Nationality variable was also grouped and converted to a new variable Region that groups countries by regions simplifying the interpretability side of the variable. From here, a map visualisation was created (Figure 1) with the purpose of facilitating understanding of the regions where customers originate from.

3.2.6. Feature Selection

In relation to the metric features, a correlation matrix analysis was done and can be seen in (Figure 2). Kendall's Tau Coefficient was the one chosen for the analysis due to the discrete nature of most of the metric features. It was verified the following insights. *BookingsCanceled*, *BookingsNoShowed* and *BookingRate* were eliminated due to their univariate nature. This was reinforced with boxplots projections to double verify the situation. *Age* does not present any relevant correlation with the other features and for that reason it was removed.

There were also two situations where variables were highly correlated. In the first case, *YearsSinceCreation* and *DaysSinceCreation* are perfectly correlated, showing evidence of data redundancy. It was decided to keep only one of the two variables, more precisely, *YearsSinceCreation*, due to its easiness of interpretability. In the second case, *TotalRevenue* and *RevenuePerStay* showed a high correlation value of 0.98. Furthermore, *TotalRevenue* was kept instead of *RevenuePerStay* since the first presented higher corelation scores with the remaining variables when compared with the other. The remaining variables don't present any strong correlation between each other, thus there is no evidence of redundancy.

Changing the focus to the categorical variables, in a first instance the following decision was made: Remove *NameID*, *DocIDHash*, and *MarketSegment* since they are not relevant for the new customer segmentation. From there it was important to understand if customer preferences when making a reservation could have an impact in our segmentation. To resolve this question, it was created a table showing the different proportions of the binary values presented in these variables. From there, it can be concluded that most of them do not seem to have major relevancy for the business objective, and because of that they were removed from the selection.

Regarding the Region*, DistributionChannel, RoomType, FloorPreference, LeadTimePreference* and *age_bins* variables, various violin plots were used to show the distribution of customers. A violin plot depicts distributions of numeric data for one or more groups using density curves (Yi, 2019). From the it can be concluded that the variables *FloorPreference*, *RoomType, Region/Nationality*, and *Age_bins* do not have discriminative power, therefore, might not be included for the clustering phase.

A distribution with *RoomPreferences* by *FloorPreferences* was also done. Knowing that the *RoomPreferences* variable represents an aggregation of binary variables already analysed individually and eliminated, plus having a distribution that does not add any valuable insight to the analysis due to lack of representation, it was decided to also exclude this variable.

On the other hand, *LeadTimePreference* showed some discriminative power. Nevertheless, *LeadTimePreference* is a derivative from *AverageLeadTime* and for clustering it was decided to only keep *AverageLeadTime* due to its numerical nature. In the end it was kept a total of six variables for the clustering analysis: *AverageLeadTime, RoomNights, RevenuePerson, RoomRate, RevenuePerStay, DistributionChannel*.

3.2.7. Data Encoding

One-hot encoding is a data encoding technique used to convert categorical or textual data into a numerical format that can be processed by the algorithms. The function 'get_dummies()' was used to one-hot encode the *DistributionChannel* feature in the dataset.

3.2.8. Data Normalization

The clustering algorithms that are going to be utilized in the following stages, are distance-based so it is important to normalize the given data before model utilization. Data normalization is a process of transforming numerical data so that it falls within a specific range or distribution, facilitating the process of algorithms. There are several normalization techniques used for this purpose, including Min-Max normalisation, Z-score, and log scaling. The MinMaxScaler method from the Scikit-Learn library was used to scale the data, which scales the data into a given range like, for example, between 0 and 1. This is particularly useful when the data does not follow a normal distribution, which is the case for the metric features in this dataset.

3.2.9. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique aiming to reduce the dimensionality of a high-dimensional datasets by identifying the directions with maximum variance in the data and projecting the data onto theses directions, resulting in a set of uncorrelated variables called principal components. PCA can help reduce the complexity of a dataset and make it more manageable, as well as identifying patterns and relationships in the data that may not be immediately apparent.

A table was created summarising the variance explained by each Principal Component. The first four PC capture a total of 95.1% of the total variance in the dataset, indicating that these are the most important component. The first principal component captures the largest amount of variance, with an eigenvalue of 0.2389 and a proportion of 0.7127. The second principal component has an eigenvalue of 0.0386 and a proportion of 0.1151. The difference between consecutive eigenvalues decreases as the principal component number increases, which suggests that the later principal components explain less variance in the dataset. This implies that retaining only the first principal components. can help simplify the dataset while still capturing most of the important information.

3.3. MODELING

With the purpose of performing a Market Segmentation, three different approaches were implemented: K-Means clustering using PCA, K-Means using the metric features, and K-Prototypes with both metric and categorical variables.

K-Means is an unsupervised clustering algorithm designed to partition unlabelled data into a certain number of distinct groupings (Jeffares, 2019), thus, being widely used in customer segmentation. It works by partitioning a given dataset into k clusters, where each cluster represents a group of similar data points based on their characteristics or attributes. One of the main advantages of K-means is its simplicity and efficiency in identifying patterns and grouping similar data points together. Another advantage of K-means is that it is scalable and can handle a large dataset with high-dimensional features.

K-Prototypes is an algorithm whose goal is to cluster large datasets with mixed numerical and categorical data. In this case, a Dissimilarity Coefficient and a Cost function are applied: for numerical variables, the dissimilarity measure is computed by the square root of the distance between the observation and the prototype, while for the categorical variables, a weighted categorical metric is used, which is defined by the number of mismatches of categories between two objects.

To determine the optimal number of clusters using K-Means, the Elbow Method and Silhouette Coefficient methods were applied. The Elbow Curve plots the value of the cost function delivered by different number of clusters, which depends on the Inertia (a measure of spread which quantifies the within-cluster variation). The elbow point represents the optimal solution. Moreover, the Silhouette Coefficient is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). By calculating the silhouette coefficient over a range of k and plotting the results, it is possible to identify the peak as the optimum.

As an alternative approach, a technique mixing both K-Means and Hierarchical Clustering algorithms was developed to find the number of clusters: compute K-Means with a large number of clusters, and use the centroids obtained to plot the dendrogram which illustrates the sequences of merges or splits of clusters. The greater the height, the larger the distance between different clusters.

Considering the first approach, K-Means with PCA, the Elbow method, the Silhouette method and Dendrogram suggest that three clusters are the optimal number for the dataset. Therefore, the K-Means with PCA algorithm was applied using a final number of clusters of three.

On the other hand, when using K-Means with the metric features, the Elbow Method and Silhouette indicates three clusters as the optimal value, while the Dendrogram indicate four clusters. Taking into consideration the business purpose, the K-Means algorithm was employed using three clusters.

Finally, due to time constraints, it was not possible to utilize the methods previously defined to find the optimal number of clusters when using K-Prototypes. Consequently, the final decision was to use three clusters, since it is a reasonable value considering the results previously obtained in the other algorithms and the business knowledge.

3.4. EVALUATION

Aiming to assess the performance of the three approaches used in Modelling, different evaluation metrics were used.

Firstly, the cluster cardinality and magnitude metrics were implemented, which shows the number of points per cluster, and the total point to centroid distance per cluster. Second, the cardinality versus magnitude plot was created, which illustrates the idea that normal clusters lie very close to the 45-degree line.

After thorough analysis, it is evident that the results using the second approach, K-Means with the metric features, are more interesting, since the clusters obtained are more balanced, with lower total intra-cluster distances.

Moreover, the Silhouette Coefficient, Calinski-Harabaz Score, and Davies-Bouldin Score were also employed. The Silhouette Coefficient ranges from -1 for incorrect clustering and +1 for distinguishable clusters. The Calinski-Harabaz Score is based on the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters. A higher score means improved cluster compactness. Finally, the Davies-Bouldin Score represents the average ‘similarity’ of clusters, where similarity is a measure that relates cluster distance to cluster size. The lower the score, the better the separation between the clusters.

Comparing the scores across different modelling solutions, it could be verified that in spite of having considerably higher results in the K-Means using PCA, these outcomes should be cautiously analysed. Since this approach results in one of the clusters with approximately 83% of the clients, the scores are inflated. Regarding K-Means and K-Prototypes, the final scores are rather similar. Although K-Prototypes presents slightly higher values, it is important to highlight that it is not a time efficient algorithm, thus not being ideal to apply in large datasets.

Considering the before mentioned conclusions, the K-Means with the metric features was considered as the final model.

4. RESULTS EVALUATION

After obtaining the final clusters, several analytical methods were applied to each label, including the comparison of the mean per cluster to the overall mean (as shown in Figure 3), the analysis of the distribution of data points in all clusters (as shown in Figure 4), and the examination of the cluster characteristics in general (as shown in Figure 5). Following a thorough analysis, a marketing strategy was devised for each segment with the objective of attracting new customers and targeting niche or unconventional customer segments, rather than using a standard segmentation method.

4.1. GEN X CUSTOMERS WITH POTENTIAL FOR FUTURE GROWTH

This segment represents the largest group of customers (48.8%) who exhibit the lowest average spending amount (321€) and primarily hail from Western Europe. Members of this group possess the highest value of the year since creation and revenue per person, indicating their potential for future growth. These customers belong to the gen x which is characterized by several behavioral attributes. They tend to prioritize value for money when seeking hospitality experiences, exhibit individualistic tendencies and may face time constraints due to family responsibilities, and they are also environmentally conscious.

In order to optimize revenue, it is efficient to implement a cross-selling strategy aimed at the identified customer cluster that values budget-friendly options. One such strategy involves offering bundled packages at a 15% discounted rate. These packages can include a hotel room, meals at the hotel restaurant, and bike rentals. This bundled package portrays the hotel as environmentally sustainable while providing guests with a fast and efficient hospitality experience through fast-casual dining to avoid time constraints, while also optimizing Other Revenue taking into to account that is below the average.

To increase the average lead time, which is currently below average, the hotel can promote these bundled packages through various channels such as its website, social media platforms, and e-mail marketing campaigns. Despite the relatively short lead time, the hotel can still gather information about customer preferences and customize the offer accordingly, as personalized experiences are valued by this customer segment.

In order to increase lodging revenue, we could use social media platforms to promote the benefits of more expensive hotel rooms as well as run a 25% per night sales promotion on the more expensive rooms during the hotel's low season and therefore increase the lodging revenue of this segment in the long term.

To strengthen customer engagement with the brand and promote an increase in check-ins and room nights in the future, it would be beneficial to offer activities that promote sustainability, such as beach clean-ups or tree planting initiatives. This is especially relevant since this particular customer segment is environmentally conscious and values sustainable practices. By providing such activities, the company can enhance its image as an environmentally responsible entity.

4.2. Baby Boomer Customers with High Average Lead Time

The second cluster, comprising primarily Western European customers, exhibits the highest average lead time of 227 days, constitutes 21.1% of the selected customers which have on average bookings check-in below the population. This segment is characterized by the Baby Boomer generation and is defined by financial stability, a tendency to prioritize personalized service, a strong emphasis on health and wellness and as not being digital native. This cluster contributes significantly to the majority of the company's overall revenue.

The hotels could collaborate closely with travel agents and other intermediaries to promote and market their supplementary premium services and to target Baby Boomer customers effectively as it is more efficient to use a traditional method to reach them.

To increase the number of RoomNights which is below the average, it could be advantageous to consider developing partnerships with events in order to promote the hotel, such as Lisbon Fashion Week or RockinRio. Through sponsorship of these events, the hotel can increase its brand awareness and attract new customers who may be attending the events. Additionally, by offering a 10% discount to those who book a stay for the total duration of the event, the hotel can incentivize more bookings and make itself a more attractive option for event attendees.

Considering that the target segment is more inclined to spend money on high-end experiences, it could be worthwhile to develop a premium service offering an additional cost for the hotel. This premium service could include welcome gifts, personalized recommendations for local activities and attractions from the hotel's concierge. Additionally, the hotel's restaurant could introduce fine dining options to cater to the discerning tastes of these customers. By offering such a premium service, the hotel can not only increase other revenue per customer but also create a distinctive experience for guests, potentially leading to positive word-of-mouth promotion and repeat bookings. It is important to refer to the fact that given the average booking lead time, it makes sense that the lodging revenue for this segment continues to be lower than average.

4.3. Baby Boomer Customers with High Revenue Contribution

According to the data, this cluster exhibits the lowest percentage of customers (30.1%). Additionally, this particular cluster is predominantly composed of individuals from the Baby Boomer generation who possess the highest Revenue per Stay, Room Night, and Bookings Check-in among all other customer segments. Taking into account the generation of this segment, it would be beneficial to establish partnerships with health and wellness brands. This would enable the company to offer guests additional products and services that promote healthy aging, thereby augmenting the customer experience.

Considering that this group holds the highest total revenue, it is imperative for the marketing plan to primarily concentrate on them. Hence, the marketing strategy should emphasize loyalty initiatives, such as offering a complimentary night at the hotel for customers who have been with the company for one year, coupled with a 10% discount on yoga and meditation classes or spa treatments. This is especially important since customers from the Baby Boomer generation place a significant emphasis on health and wellness.

To strengthen customer retention, the hotel could initiate a loyalty program that grants points for every stay and presents exclusive promotions to its members, thus encouraging sign-ups and participation. Additionally, this loyalty program can be promoted through various channels, such as the company website, social media platforms, and email marketing campaigns, to augment its visibility and reach.

5. DEPLOYMENT AND MAINTENANCE PLANS

Now that the analysis has been concluded, the focus goes in trying to understand how should Hotel H implement the new segmentation structure and insights discovered. On a general note, the following plan was created. Is important to note that the implementation of the plan depends on the CEO approval.

On a first instance, it is necessary to recruit staff with the technical ability to handle the inflow of data and have data maintenance and modelling skills. It is also recommended for Hotel H to train current staff accordingly so they gain awareness of the changes and, in the case, they have direct contact with the customer, they will be able to deploy them effectively. Assign members of different departments, such as marketing, sales, finance and operations to be part of the deployment plan. This phase is estimated to take about three weeks.

On a later stage, one month after training and recruitment, the Customer relationship management (CRM) needs to be adapted to avoid all of the data issues referred earlier in the project. This would be done in a two-month timeline. After the approval of the administrative department and the feedback of the users of this system the CRM should be tuned in a three-month schedule. Based on the clean data present in the CRM, the Analytics team could work together in order to define data-driven strategies and start on their implementation in the following business year.

Devising strategy and direction for your organization should never be a one-time event (Woods, 2021). The main goal of this project was to create a new customer segmentation that reflected customer characteristics. But customers are far away from static information. They are as dynamic as they can get, along with all external factors surrounding them and therefore a proper maintenance should be done.

The primary objective of the proposed maintenance plan is to monitor and calibrate the performance of each customer segment to assess the success of the strategies implemented. Another point of verification is going to pass through a verification check on market shifts. Even though these punctual market shifts can be minimal, it can have a big impact on how customers perceive a service. It is also proposed a maintenance schedule every business quarter so it can be in parallel with other normal business procedures occurring at the time. These checks can be made in a partial way instead of going through the intensive process of in-depth data gathering even though it is recommended to do it at least once a year. These partial data checks can be done by survey collection, for example.

Even though its advised to follow this maintenance schedule, in cases of major changes in the market, drastic events such natural disasters or internal position or strategy changes, it is recommended for Hotel H to revise its segmentation even if it is not the schedule time to do so.

6. CONCLUSIONS

In conclusion, the final objective of the project was successfully accomplished. By leveraging on the richness of the data and the skills of the Young Talent Consulting Group members, the team was able to draw fundamental conclusions for the business success.

In a world increasingly competitive, companies must be able to differentiate themselves in the market and take advantage of their capabilities. Hence, data is a key asset for the company, being essential to understand customers' needs behind the evident.

To gather value from raw data into the business, different steps were performed, from data understanding, through data treatment and manipulation, until data modelling. During this process, some challenges had to be overcome, namely regarding problems with data consistency, and availability of data.

Following the modelling phase, three different Market Segments with clearly different attributes were identified and described. This analysis enabled the development of strategies tailored to each market segment needs, which will have clearly positive impacts on the business, more specifically, on what regards being competitive in the market, increasing customer retention, and promote customer acquisition. It is important to highlight that the conclusions of the project have practical implications in the business in what concerns implementation and integration of the strategies.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Although the current model has potential, there are various enhancements that can be integrated to optimize its performance. Furthermore, it may be worthwhile to evaluate alternative models to determine if they yield superior outcomes.

First of all, the database modelling and management should be reviewed, as some key issues appeared during data manipulation. Some considerations for improvement could be not allowing negative values in variables such as Age and AverageLeadTime neither the introduction of customers with the same DocIDHash. Also, regarding night offers for staff, these should be identified separately from normal Check-In. Another important element would be to connect the children's rooms to the main room.

Moreover, to enhance the comprehensiveness of the data, it would be relevant to consider several additional features in the future. For instance, as suggested by Karpushin and Zaytsev (2018), collecting information about the hotel's location could facilitate the calculation of distances from the airport, city centre, and popular tourist attractions. Furthermore, recording data on different room types and their corresponding price ranges could help identify patterns in customer preferences and improve pricing strategies. Additionally, incorporating customer reviews, such as overall rating, number of positive and negative reviews, and average ratings for various aspects of the hotel such as cleanliness, staff and location, could provide valuable insights into customer satisfaction and areas for improvement. Finally, gathering information about the timing of bookings, check-in and check-out dates, and length of stay could assist in identifying seasonal booking patterns and potential opportunities for revenue optimization, as suggested by Cho and Jang (2019).

In order to enhance the scope of analysis, it is our belief that the incorporation of external datasets could be beneficial. For example, weather data could be utilized to investigate the potential impact of weather patterns on hotel bookings and revenue. Economic indicators, such as GDP, inflation rate, or unemployment rate, could also be incorporated to better comprehend how these factors may be influencing customer behaviour and revenue trends. Moreover, data on competitor hotels in the same region could be leveraged to gain insights into market share and competition, and to compare indicators such as ADR, revenue, and room night.

In the beginning of data processing, part of the dataset was separated from the mainframe and considered as potential clients. For future consideration, this group should also be analysed to try to understand these customer characteristics are and how these can be possibly retained.

Finally, it is important to consider that K-Means and K-Protypes have some limitations, namely, inaccurately identifying convex regions, and performing poorly when outliers and noise are present in the data. Hence, further algorithms should be tested, for instance, Density-based clustering algorithms (DBSCAN, Mean-shift) which are able to identify clusters in data without assuming a particular shape.

7. REFERENCES

A Step-by-Step Explanation of Principal Component Analysis (PCA). (2022). Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Alipour, H., Ahani, A., & Ebrahimi, A. (2015). Hotel no-show prediction: A hybrid approach based on fuzzy logic, artificial neural network and support vector machine. Journal of Hospitality and Tourism Technology, 6(1), 1-21. https://doi.org/10.1108/JHTT-07-2014-0017

Cho, Y. J., & Jang, S. S. (2019). Investigating seasonal patterns of hotel room prices and occupancy rates. *Journal of Travel Research*, 58(2), 274-288.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.*

Hayasaka, S. (2022, February 11). How Many Clusters? - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/how-many-clusters-6b3f220f0ef

Huang, Y., & Wan, J. (2019). *A survey on data discretization techniques.* Big Data Mining and Analytics, 2(1), 1-22.

Instituto Nacional de Estatística, I. P. [INE]. (2019). *Estatísticas do Turismo 2018*.

Jeffares, A. (2019, November 19). *K-means: A Complete Introduction - Towards Data Science.* Medium; Towards Data Science. https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

Karpushin, M., & Zaytsev, A. (2018). *The study of the impact of hotel location on customer satisfaction: A case study of hotels in Moscow.* Journal of Tourism and Hospitality Management, 6(2), 31-43.

Kelleher, J.D., Tierney, B. (2018). *Data Science: An Introduction*. Chapman and Hall/CRC

Kurgan, L. A., & Musilek, P. (2006). *A survey of knowledge discovery and data mining process models.* The Knowledge Engineering Review, 21(1), 1-24.

Lisboa OFFICIAL Site. (2020). Turismo de Lisboa. https://www.visitlisboa.com/

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data.* John Wiley & Sons.

Ramezani, M., & Farsijani, H. (2013). *Data Mining Approach for Hotel Customer Relationship Management.* International Journal of Business and Management, 8(3), 1-10. https://doi.org/10.5539/ijbm.v8n3p1

Sander, J., Ester, M., Kriegel, H., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194. https://doi.org/10.1023/a:1009745219419

Soria, J., Chen, Y., & Stathopoulos, A. (2020). K-Prototypes Segmentation Analysis on Large-Scale Ridesourcing Trip Data. Transportation Research Record: Journal of the Transportation Research Board, 2674(9), 383–394. https://doi.org/10.1177/0361198120929338

Travel and tourism account as a % of GDP. (2022). Pordata.pt. https://www.pordata.pt/en/Portugal/Travel+and+tourism+account+as+a+percentage+of+GDP-2632

Yi, M. (2019). *A Complete Guide to Violin Plots*. Chartio; Chartio. https://chartio.com/learn/charts/violin-plot complete-guide/

Das, A. (2022, January 9). Deciding number of Clusters using Gap Statistics, Davies-Bouldin Index, & Calinski-Harabasz Index for K-Means and Hierarchical Clustering using Python. Medium; MLearning.ai. https://medium.com/mlearning-ai/deciding-number-of-clusters-using-gap-statistics-davies-bouldin-index-calinski-harabasz-index-2ce9acfb611

Wei, H. (2020, January 2). *How to measure clustering performances when there are no ground truth? Medium*; Medium. https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c#:~:text=The%20Calinski

Woods, S. (2021, September 14). *Customer Segments Change: You Need a Segmentation Maintenance Plan - Sterling Woods Group*. Sterling Woods Group. https://sterlingwoods.com/blog/customer-segments-change-over-time/

Yi, M. (2019). *A Complete Guide to Violin Plots*. Chartio; Chartio. https://chartio.com/learn/charts/violin-plot-complete-guide/

8. APPENDIX

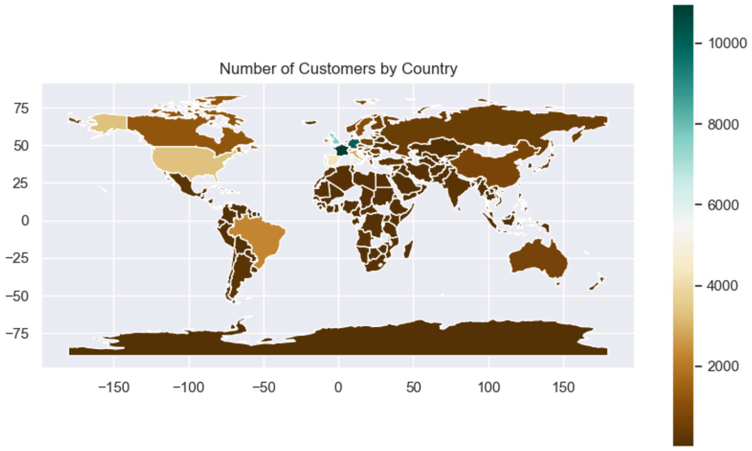

Figure 1: Number of Customers by Country.

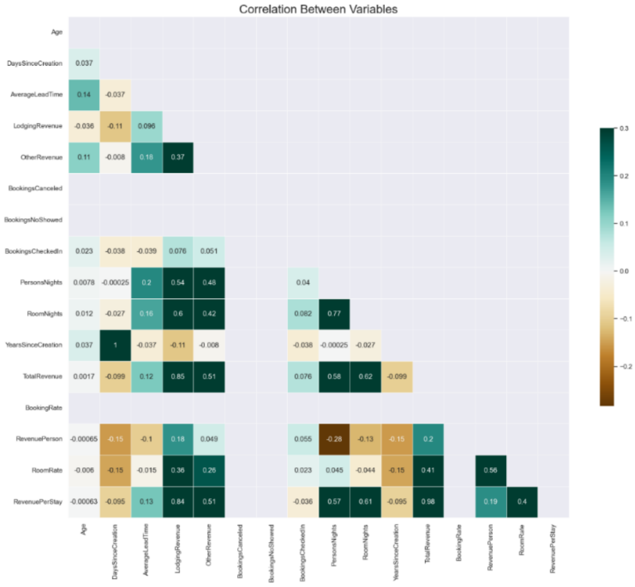

Figure 2: Correlation Matrix after Data Preparation.

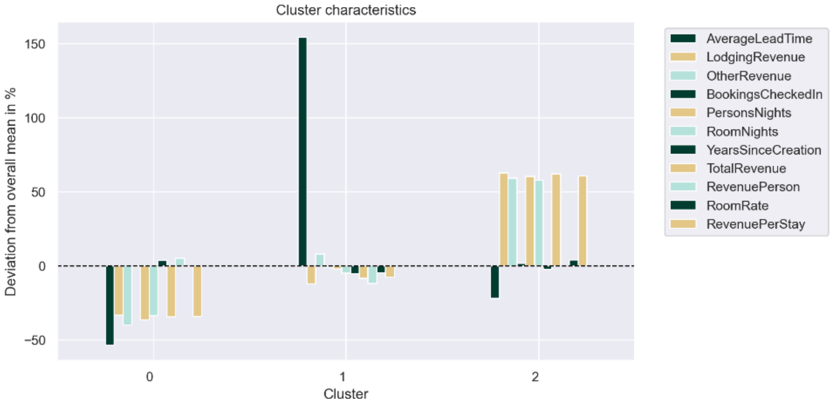

Figure 3: Comparison of each cluster feature mean with the overall population mean.

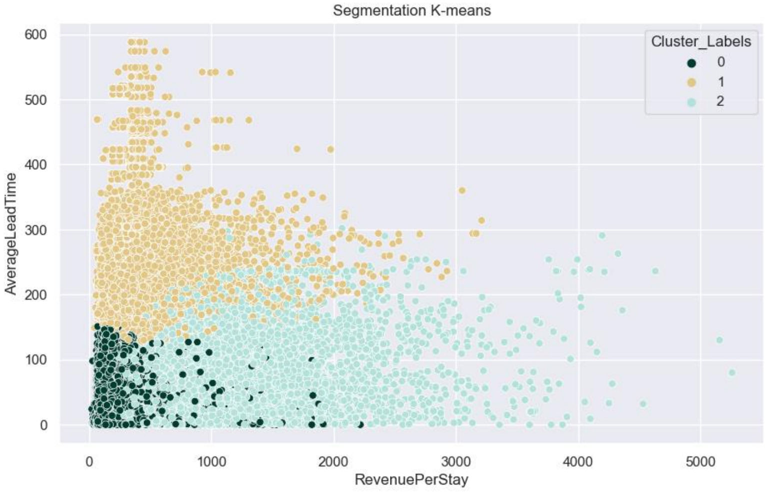

Figure 4: K-Means algorithm Results - RevenuePerStay VS AverageLeadTIme.

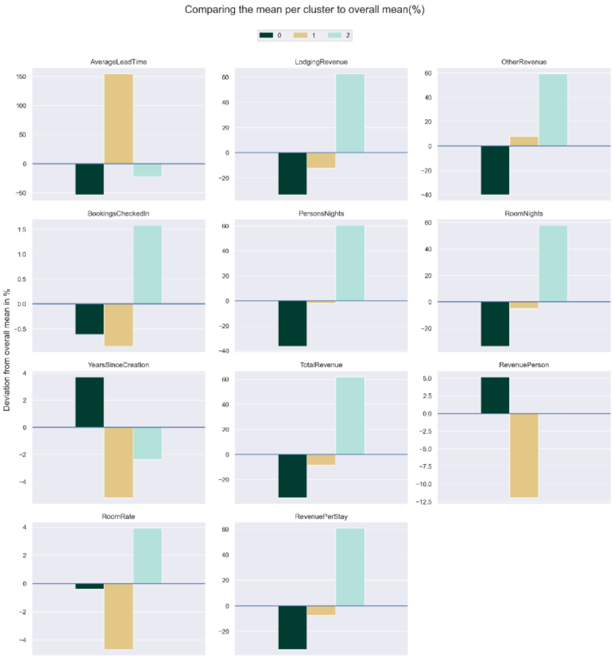

Figure 5: Comparison of each cluster feature mean with the overall population mean.