Master Degree Program in

**Data Science and Advanced Analytics**

# Business Cases with Data Science

*Case 1: Hotel Customer Segmentation*

Ana Carolina Ottavi, number: 20220541
Carolina Bezerra, number: 20220392
Duarte Girão, number: 20220670
João Pólvora, number: 20221037
Luca Loureiro, number: 20221750

Group Q: OptimaDataConsulting

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

Universidade Nova de Lisboa

March, 2023

# INDEX

# 1. EXECUTIVE SUMMARY

In an ever-competitive hospitality industry, understanding the customers is essential to creating a successful business strategy. By segmenting your customers based on their unique characteristics and behaviors, you can tailor your marketing and service offerings to meet their specific needs and preferences. In this report, we will provide an in-depth analysis of the hotel chain's customer base, identifying key segments and their distinct characteristics, behaviors, and preferences. Our analysis is based on a comprehensive dataset, including customer demographics, booking patterns, and other reservation details, providing valuable insights into the hotel chain's customer base. By understanding the customers' needs and preferences, we provided suggestions in order to enhance their overall experience, build brand loyalty and increase the bottom line.

# 2. BUSINESS NEEDS AND REQUIRED OUTCOME

## 2.1. Background

Hotel H belongs to an independent Hotel Chain C, located in Lisbon, Portugal. Hotel Chain C operated 4 hotels until 2015 and has been acquiring new ones since then. Furthermore, the Hotel Chain C created a marketing department and a new marketing manager position as part of its efforts to grow in the Hospitality market.

The Hotel Chain C knows that understanding current customers allows organizations to identify groups of customers that have different characteristics and behaviors. Understanding them is vital in every industry once the process of finding new customers begins by learning as much as possible from the existing ones.

The current Hotel Chain C actual segmentation is based on the origin of the sales and is done according to the hospitality standard market segmentation. The new marketing manager A considers that the segmentation does not fulfill the current needs and does not allow the organization to make better strategic choices about opportunities, product definition, positioning, promotions, pricing and target marketing, once it only reflects one of the customers' characteristics.

A new customer segmentation would allow the board and the marketing manager, A, to create and drive strategic options to improve the perceived value to the customers, market value proposal and profits.

## 2.2. Business Objectives

The main business objectives of our project are:

- Identify the characteristics and behaviors that should be used to segment the customers;
- Provide a segmentation solution and profile in terms of characteristics and behaviors;
- Suggest business insights to find and retain customers;
- Provide operational process suggestions to improve data quality;
- Provide updated and transparent data periodically to the marketing department.

### 2.3. Business Success criteria

We defined the following business success criterias:

- A descriptive and characterized segmentation solution approved by the marketing manager A;
- Develop a model that supports both our current and future need for customer segmentation.

### 2.4. Situation assessment

The project will be run by a team of five business/data scientists with access to internet connection and five laptops. A business solution according to the objectives must be delivered by 8th of March. We were able to extract an analytic based table from the operational systems with all the customer transactions and corresponding details are going to be used to develop the project.

As risks, we identified that the success of the project critically depends on the data's quality. On the other hand the quality of the data depends on the human precision in following the established processes to register and manage the clients on the Customer Relations Management tool.

As a strategy to deal with data inconsistencies we defined a clear customer definition based on data. We considered as our customers the clients that clearly engaged with our hotels performing the check-in.

In terms of benefits, our business is going to profit from a better data-driven segmentation. The business decisions will be done with bigger confidence, the business opportunities will be clearer and there will be more insights to increase operational efficiency through cost savings.

### 2.5. Determine Data Mining goals

- The main data mining goals of our project are:
- Explore the data and identify the variables that should be used to the customer segmentation;
- Use PCA to reduce dimensionality and speed-up model developments;
- Use K-means clustering to identify customer segments;
- Justify the number of clusters;
- Define and characterize each cluster.

## 3. METHODOLOGY

### 3.1. Data understanding

Before we could segment the data, we needed to understand the data we are dealing with, followed by a Quality Check of our variables. Initially, we started by analyzing a summary statistics table, which provided a general description of the variables to understand their behaviors, which revealed that we had a total of 29 variables and 111.733 rows to work with. Looking at the table statistics executed, it was possible to take away some initial insights for each feature.

| Features | Takeaways |
|---|---|
| ID | The unique identifier of a customer can't be clustered (uniform distribution). |
| Nationality | It was possible to disclose that there are 199 different possibilities for this feature. Most customers were French, German, Portuguese, British and Spanish. We then lowered the cardinality for Nationality. |
| Age | For this feature, the mean and quartiles have close values, which can tell us it follows a normal distribution. There were negative values and a maximum value of 123, which were not considered afterwards. |
| DaysSinceCreation | It was possible to observe that the minimum of elapsed days since a profile was created was 36 and the maximum was almost 4 years until 31/12/2018. We can also tell that at least 25% of our customers had their profile created for 288 days and 75% had their profile for more than 889 days. There were 1062 people that never checked in, yet had a profile in the database for more than 2 years, which were erased. There was a peak in 2016, revealing a maximum number of days since the creation of customers profiles. There was also a peak in August for this variable. |
| NameHash | There were repeated rows for NameHash, which we considered to be cases such as when people have the exact same name. |
| DocIDHash | We checked duplicates for this feature, which comprised around 10% of our dataset. The possible explanation for the repeated entries could be due to one individual using their own DocIDHash and others using a company related DocIDHash. However, we did disconsidered from our dataset the intersection of NameHash and DocIDHash duplicates. |
| AverageLeadTime | This variable was reported to have 13 negative values and we could argue that it contains outliers. From our analysis, at least 25% of the customers do their booking on the same day the booking is intended, and at least 75% books 95 days in advance. More so, the maximum registered was a booking done 588 days before arrival. We also decided to create a filter such that the values for this variable for each customer have to be bigger than DaysSinceCreation. In addition, those whose age is more than 60 years old had higher values for this variable, and those whose age is less than 20 years old had the lowest. |
| LodgingRevenue | We could immediately tell that there were customers registered that had a null revenue. Additionally, 25% of the customers had a null LodgingRevenue, so there is a big amount of customers who bring zero to little revenue to the hotel. 75% of customers had a revenue of 393,30 and our maximum value registered is 21.781,00. |
| OtherRevenue | At least 25% of the customers of the hotel chain brought 0 OtherRevenues, and 75% brought 84,0. We can also say that the maximum value obtained by a customer was 8.859,25. The reason behind so many clients with no revenue may be due to the fact that there were also at least 25% of customers that did not check in. From an initial standpoint, there may be outliers, which we will explore further on in this report. We realized that there were around 30% of records which have OtherRevenue and LodgingRevenue equal to 0, and no bookings as well. The majority of Revenue came from Travel Agent or Operators. |

| | |
|---|---|
| BookingsCanceled | At least 75% of the customers never canceled a booking, and the maximum value encountered was 15 canceled bookings. There were 125 records with one canceled booking. |
| BookingsNoShow | It was possible to observe that at least 75% of our population didn't have a record of no shows and the maximum value we encountered was a record of someone with 3 bookings with no shows. Those who failed their booking, didn't appear to have a big impact on revenue. |
| BookingsCheckedIn | From our summary statistics table, we can tell that at least 25% of the customers never checked in, and the maximum value of bookings we encountered were 76. From further analysis, we gathered that 16.884 customers never checked in, yet had a profile for over one year; 1.062 customers had their profile created more than three years ago from the data extraction and never checked in. We also found out a customer that brought revenue to the hotel chain, while never checking in, which we assumed to be an error. Further on, we discovered that French people had more bookings above all other nationalities. |
| PersonsNights | It was revealed that at least 25% of the population considered, had a null PersonsNights value, which can be related to the fact that at least 25% of our customers never checked in before. At least 75% of customers had PersonsNights equal to 0. Around 50% of customers had a value of 4.0 and the maximum value we unraveled was 116. The most common values for this feature were 0, 6, 4, 2 or 8. We also found out values that differ from 0 on RoomNights when PersonsNights is equal to 0. |
| RoomNights | It was possible to observe that 25% of the customers had RoomNights equal to zero, once again, most likely due to the fact that 25% of customers never ended up making a booking. 75% of the population considered had a value of 3 and the maximum registered is 185. This value could be possible in situations of weddings, work conferences, among others. The main values of this variable were 0, 3, 2, 4 and 1. There were no situations where RoomNights is equal to 0 and PersonsNights is different from 0. |
| DistributionChannel | There are 4 different types of Distribution Channel, which are Corporate, Travel Agent or Operator, Direct and GDS System. The most common channel used is through Travel Agent or Operator, which is also the channel that brought more Revenue, followed by Direct Channel. Additionally, it was found that for all of the Distribution Channels, most of the clients had ages between 41 and 50 years old. Both younger generations (<30y) and older generations (>60y) don't use Corporate or GDS Systems as Distribution Channels. |
| MarketSegment | The previous segmentation was based uniquely on this variable, however we found it is no longer relevant. More so, the majority of Distribution Channels were correctly associated with its corresponding Market Segment. There are 7 different possibilities, ***figure 1***, for this feature, which are Corporate, Travel Agent or Operator, Other, Direct, Complementary, Groups and Aviation. The category Other was revealed to be the one bringing the most revenue, followed by Direct. We proceeded to reduce the cardinality for this variable, which reduced the possible categories to Other, Travel Agent or Operator, Direct and Groups. |

| | |
|---|---|
| SR (all Special Request Variables) | At least 75% of customers asked to have a King Size Bed in their room. The requisition of other requests however was not very present. According to the mean, the most required Special Requests are SRKingSizeBed, followed by SRTwinBed and SRQuietRoom. Most customers had either 0 or 1 Special Requests. Most of these features were highly unbalanced, so we disregarded the majority of these Special Requests variables that provided no relevance further on. |

## 3.2. Data preparation

In this section the objective was to select the appropriate data to proceed to the modeling phase. We started out by removing unnecessary records, following previous insights gathered in the Data Understanding section of the project. Both Age and AverageLeadTime reported to have negative records, which were impossible in the context of these variables, so they were immediately discarded from our dataset. Existing rows in which both the NameHash and DocIDHash were duplicated were excluded. We also filtered out records where AverageLeadTime was higher than DaysSincecreation. Under our interpretation, DaysSinceCreation is either higher or equal to AverageLeadTime. This last situation is due to the fact that the time difference between the day a customer is registered in the hotel database and the day of extraction of this dataset, can never be lower than the time difference between the day the customer makes the reservation and does the *check in*. Another decision that was made was to delete all the customers that arrived yet never did their check-in, considering profiles of customers that have been registered for over two years from the date of the database extraction. This criteria drops 1062 records, whose clients, in our understanding, were not representative of the Hotel Chain C customers. Lastly, we decided to delete all the customers that never checked in, even though their profile is registered for over two years from the date of the database extraction. According to this criteria, we encountered 1062 rows, which ended up not being a part of the data used for segmentation.

Some modifications in variables were made. We dropped the column ID because it doesn't have a function in this project, since each observation corresponds to an individual customer. In addition, we also decided to remove the features NameHash and DocIDHash, which have a high cardinality and did not provide relevant information. In addition, we removed BookingsCanceled, BookingsnoShow, SRHighFloor, SRLowFloor, SRAccessibleRoom, SRMediumFloor, SRBathTub, SRShower, SRCrib, SRNearElevator, SRAwayFromElevator, SRNoAlcoholInMiniBar, TotalSR and MarketSegment.

We conducted a simple missing values analysis, which provided us both with the number of null entries across our dataset. The variable with the highest proportion of missing data, going as high as 3,74%, was Age. Except for Age, only DocIDHash is revealed to have null values. Regarding missing values, our decision was to use KNN Imputation. Using this popular technique, we input a new sample by closest samples and average these closest points and fill in the missing value[1] . We only used this method for the feature Age, which was considered in our final model. At this stage, we also modified the data type of the variable Age from float64 to int64.

[1] (Brownlee, Jason. (2008, August 13). Knn Imputation for Missing Values in Machine Learning

After defining our metric and categorical features and selecting the appropriate data, we plotted our variables. Regarding Histograms it's possible to identify the following:

- Age is unimodal and left skewed. There are fewer customer as age progresses;
- DaysSinceCreation is multimodal and right skewed. There are more customers with their profile created recently;
- AverageLeadTime is unimodal and right skewed. Few customers make their reservation with a lot of days in advance;
- PersonsNights is unimodal and right skewed.
- After lowering the cardinality for Nationality, the category more prevalent is Other.

The proposed algorithm for this customer segmentation project was K-means clustering, which is sensible to outliers. Therefore, verifying outliers is of extreme importance and determinant for the performance of the algorithm. The outlier removal process for this dataset required two different approaches - a manual approach, using boxplots for verification of outliers, and the IQR Approach. For the first method, we resorted to both counts plots and hist plots for each feature in the dataset, which clearly revealed most numeric features may still have a large number of outliers, demonstrated in the ***figure 2***. We made the following filtering for out variables, in order to exclude the presence of extreme and disproportionated values:

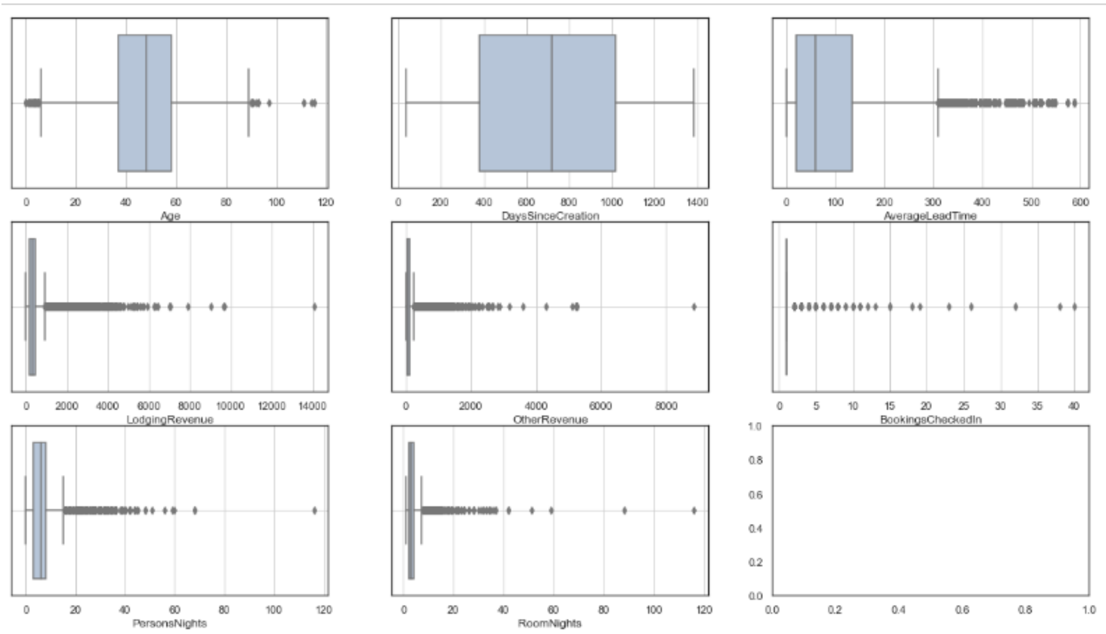


***Figure 2*** **-** "*Numeric features with large number of outliers*"

- We filtered out existing records of customers with more than 90 years old. A few of the rows containing people with this filter on, revealed to have BookingsCheckedIn equal to 0. Besides that, the expected living age in 2018 in Portugal was 80, which is in accordance with our threshold mentioned before;
- For AverageLeadTime, we defined one year gap for our threshold, going accordingly with the representative boxplot.

- We removed LodgingRevenue records that had a value higher than 6000. As for OtherRevenue, we limited our dataset once again by erasing customers whose OtherRevenue was higher than 2000, which accounts for one third of the range of this feature.
- Another decision was to remove rows that had more than 5 bookingsCheckedIn.
- By looking at our boxplots and histograms, we defined a maximum threshold of 40 for PersonsNights and a maximum threshold of 30 for RoomNights.
- We discarded rows that had BookingsCheckedIn equal to 0, which seem to be irrelevant to our customer segmentation;
- Almost all variables with exception of Age and DaysSinceCreation appear to still have outliers.

The other method used for outlier detection was the IQR method. With this method, initially we set up a fence outside of the first and third quartile. To build this fence we take the values that lie under the first quartile by more than 1.5 times the size of the interquartile range or lie above the third quartile by more than 1.5 times the size of the interquartile range. All the values outside the fence will be considered as outliers. After outlier removal with the IQR method, we determined the percentage of remaining data, which provided a value of 66.21%, when criterion is 3. In the end, we were able to confirm the existence of extreme values both from the Manual and IQR approach, however we decided to remove the most extreme values we encountered through a manual approach, which better selects data.

After manually removing outliers, we were able to keep 65,17% of our initial dataset. At this end of this stage, both histograms and boxplots were created once again, and we were able to confirm that they provided much more visual insights than before.

Plotting an initial Correlation Matrix makes it possible to identify and avoid redundancy in the dataset. Given the correlation matrix, we decided to remove variables that are highly correlated with each other and don't bring new information for our future conclusions.

PersonsNights and RoomNights: these variables have a 0.90 correlation, which can be explained by the fact that they are somehow connected. We then considered that PersonsNights was able to produce more interesting conclusions, and RoomNights was therefore removed.

OtherRevenue and LodgingRevenue: these variables have a 0.50 correlation. Both OtherRevenue and LodgingRevenue share high correlations with other 2 variables (RoomNights and PersonsNights). Although OtherRevenue is representative to our dataset, LodgingRevenue had bigger values, and consequently the feature OtherRevenue was dropped from our dataset.

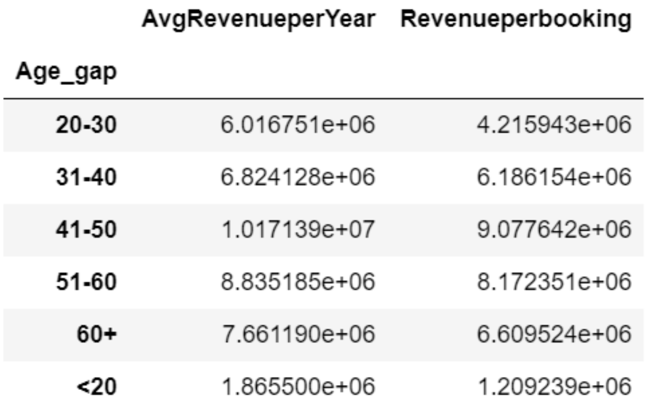
In this section we performed the Construct Data step from CRISP-DM methodology which is a task that includes a process to data preparation operations such as the production of derived attributes, complete new records or transformed values for existing attributes[2] in order to prepare the data for modeling. Therefore, in this section, at the step 3.3.1 of our jupyter notebook, we started by creating some new variables, namely:

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide. SPSS Inc.

- Total_revenue: the sum of the columns LodgingRevenue and OtherRevenue;
- AvgRevenueperYear: The average revenue per year of each customer by multiplying Total_revenue with 365 and dividing by the number of DaysSinceCreation;
- Revenueperbooking: Average of total revenue per room per book by dividing the column Total_revenue by the column BookingsCheckedIn.

Next, we performed a Outlier Manual Removal (First method) based on histograms, box plots and descriptive statistics resume analysis from the AvgRevenueperYear and Revenueperbooking variables. We mostly adopted a 'try and error approach' and the percentage of data kept after removing outliers was 65,17%.

Additionally, we created a function called age_gap to classify the age of customers into different groups. Our age_gap function takes the age of an individual as input and assigns them to an age group based on the age range. Then, the function is applied to the "Age" column in the dataset to create a new column Age_gap containing the age groups of each customer. As output there's a summary table that can be used to investigate the relationship between age groups and other variables in the dataset ***figure 3.***



| | AvgRevenueperYear | Revenueperbooking |
|---|---|---|
| Age_gap | | |
| 20-30 | 6.016751e+06 | 4.215943e+06 |
| 31-40 | 6.824128e+06 | 6.186154e+06 |
| 41-50 | 1.017139e+07 | 9.077642e+06 |
| 51-60 | 8.835185e+06 | 8.172351e+06 |
| 60+ | 7.661190e+06 | 6.609524e+06 |
| <20 | 1.865500e+06 | 1.209239e+06 |

***Figure 3 -*** "*Summary table that used to investigate the relationship between age groups and other variables*"

And then, at step 3.3.2 of our jupyter notebook, we decided to remove unnecessary/irrelevant features from the dataset, including all the features that have a Total_revenue equal to 0 since they do not represent a relevant target for the project. Also, some metric features such as Age, PersonsNights, Total_revenue, BookingsCheckedIn and DaysSinceCreation were removed.

Here, we rechecked the Correlation Matrix so we could verify that the way we developed the data preparation section so far changed the Spearman Correlation heatmap in a way that we can see less extreme colors representing strong positive or negative correlations (deep red and white from our range, respectively), which is demonstrated in ***figure 4.***

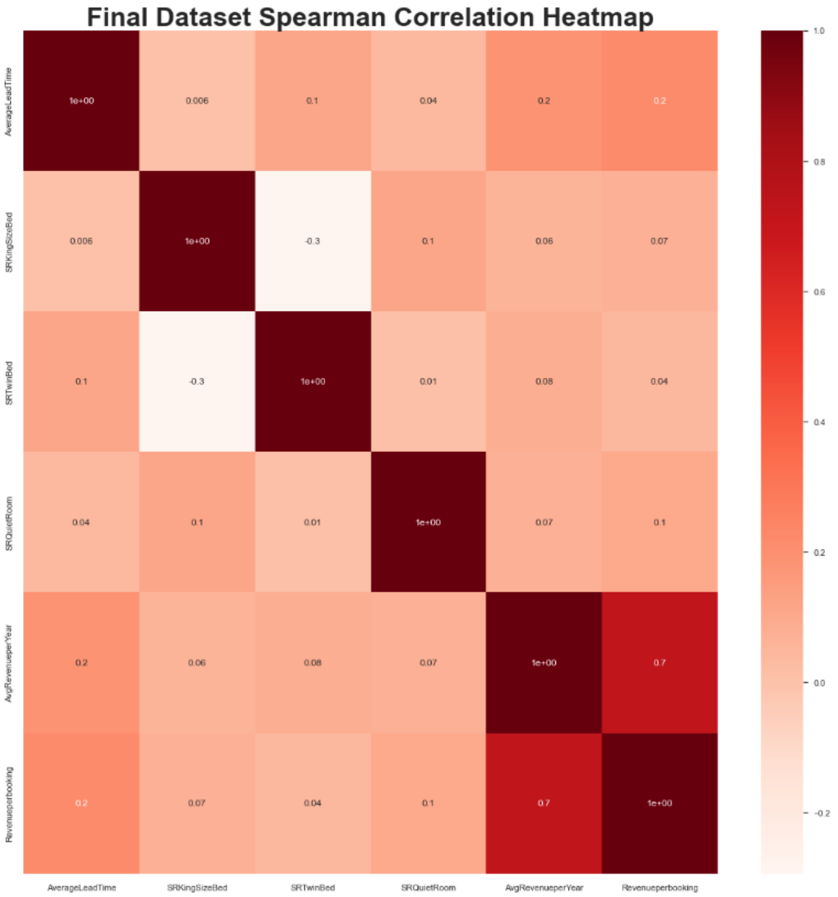


***Figure 4*** **-** *“Rechecked the Correlation Matrix”*

In this sense we could verify that only AvgRevenueYear/RevenuePerBooking has a higher correlation of 0.7, which is acceptable since these two variables have the same foundation.

To finish this section, we converted categorical features into a binary representation by applying One Hot Encoding at the 'Nationality', 'DistributionChannel' and 'Age_gap' features and then we saved the transformed data as 'df_ohe' variable.

Following the CRISP-DM methodology, we have the step 3.4 Integrate/Merge Data of our jupyter notebook, which task involves combining/aggregating/merging information from multiple tables or to create new records or values[3]. Therefore, here we integrated the One-Hot Encoded dataframe 'df_ohe', to the main dataframe.

In addition, we have a Data Formatting step, which is a task that involves making syntactic modifications to the data without changing its meaning to meet the requirements of the modeling tool[4]. The output of this task is reformatted data.

[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide. SPSS Inc.

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide. SPSS Inc.

In this sense, we started this section by Normalizing Data using the MinMaxScaler() method. It’s important to mention that we also tried 'StandardScaler' however we have obtained better results with 'MinMaxScaler'. We saved the normalized data as df_before_scaling as a security copy of the dataset before normalization. The method scales the data to the range [0,1].

Inside of Format Data step, we also performed the Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, which is a technique that transforms high-dimensional data into a lower-dimensional space by finding linear combinations of the original features that explain the maximum amount of variance in the data. In this sense, we created a df_before_pca dataframe as a security copy of the dataset before applying PCA.

And then we started by fitting the PCA algorithm to the dataset. The percentage explained by each component and the cumulative sum of the percentage explained are calculated and displayed in a table. The table shows that 11 components explain 94,48% of the variation in the dataset.

## 3.3 Modeling

In order to begin our modeling step based on CRISP-DM methodology, we performed the K-Means algorithm as a modeling technique, which is an algorithm that partitions a dataset into k clusters based on similarity between data points in order to predict the customer segmentations and provide this analysis to the marketing department work in the new strategy.

Then, we generated two tests designed to determine the best number of clusters for our analysis, which are Inertia metric and Silhouette Coefficient, as following below:

The first test was based on the Inertia metric of the dataset, which measures how well a dataset was clustered. In this sense, the lower the inertia values means better clustering results. We plotted the inertia result values per clusters and by analyzing the Inertia graph, under "Elbow Method", we could conclude that the ideal number of clusters is 6, as long as it presents the lower SSW ( ***figure 5***).

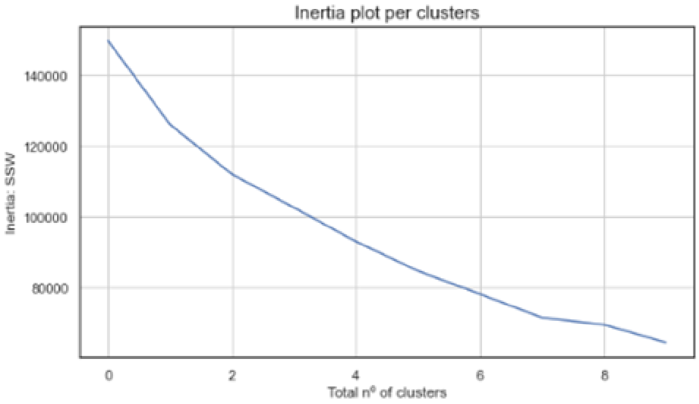


***Figure 5*** **-** *“Inertia plot per Clusters”*

Afterwards, the second test was based on the Silhouette Coefficient[5], which measures how similar an object is to its own cluster compared to other clusters. Initially, the Silhouette value increased as we increased the number of clusters from 3 to 4. However, there was a slight decrease in the Silhouette value as we increased the number of clusters from 4 to 5/6. The Silhouette value then increased again from 7 onwards until around 9, indicating that a higher number of clusters better represent our observations. Therefore, based solely on the Silhouette metric, we conclude that 8 clusters is the ideal number, as we can see at the ***figure 6***. The higher the Silhouette Coefficient means better results.

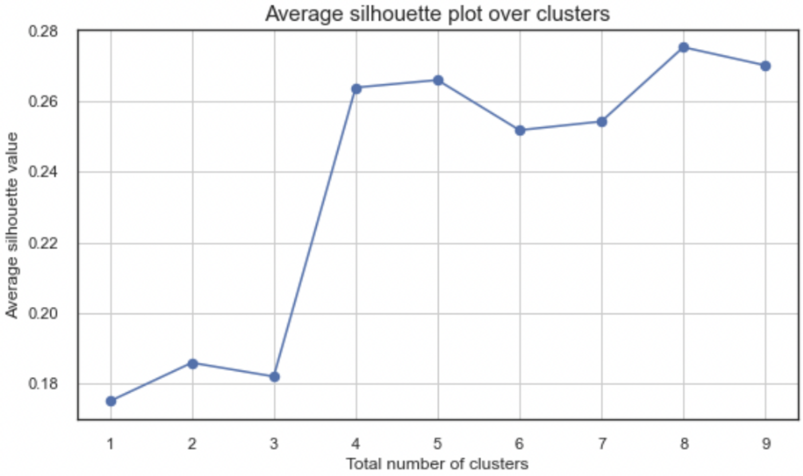


***Figure 6 -*** *"Average silhouette plot over clusters"*

In conclusion, we aimed to minimize Inertia and maximize Silhouette metrics. After careful consideration, we chose 6 clusters based on Inertia, as it is practical for marketing purposes to have fewer clusters. Too many customer segments may not be ideal for marketing.

At this modeling step, we used the K-Means algorithm to perform the clustering task and here we are going to present the final results of this algorithm, along with some important metrics to evaluate the quality of our clustering solution.

We decided to build the final clustering solution using 5 clusters to run the K-Means algorithm and we obtained the following metrics.

We evaluated the Silhouette Plot of K-Means Clustering in 5 Centers (see ***figure 7***). The Silhouette Plot[6] widget offers a graphical representation of consistency within clusters of data and provides the

[5] Scikit-learn. (n.d.). K-means clustering: Silhouette analysis. Retrieved March 7, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

[6] Yellowbrick. (n.d.). SilhouetteVisualizer. Retrieved March 8, 2023, from https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html

user with the means to visually assess cluster quality[7]. The higher the Silhouette score, the more similar the observation is to its own cluster.

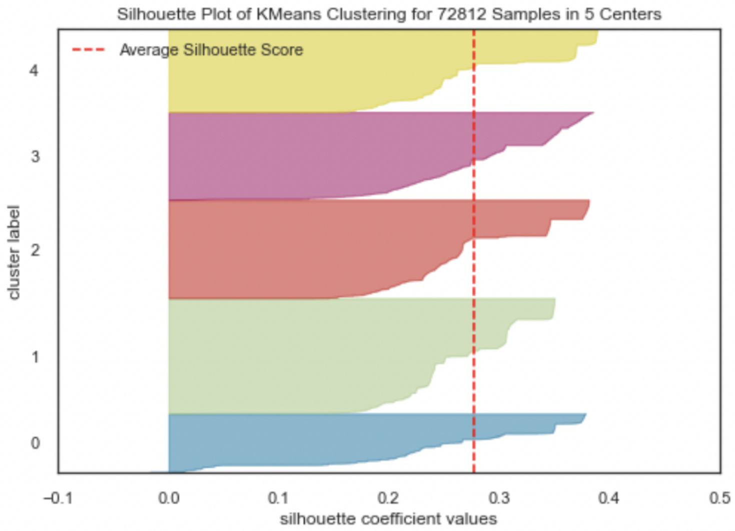


***Figure 7 -*** *"Silhouette plot of K-Means Clustering for 72812 samples in 5 centers"*

- Inertia: The within-cluster sum of squares (WSS) value for our final solution was 93632.06896990279. This value represents the distance between each data point and its centroid. We aimed to obtain the lowest value possible for this metric.
- Silhouette Score: The silhouette score (SS) for our final solution was 0.2774036571651134. This score represents how similar a data point is within-cluster (cohesion) compared to other clusters (separation). We aimed to obtain the highest value possible for this metric.

Regarding the cluster solution, we have five clusters with different characteristics. We observed that cluster 0 had the lowest cardinality, while cluster 1 had the highest. In terms of magnitude, we noticed some degree of similarity among all the clusters. Cluster 0 had the lowest magnitude, while cluster 1 had the highest.

The cardinality versus magnitude graph results in a positive correlation between these two metrics, which indicates that clusters with a higher number of observations correspond to clusters with a higher total sum of distances to their centroid.

Regarding cluster visualization, ***figure 8,*** we plotted the cluster sizes and distances in a two dimensional graph. Thus, it is possible to have an idea of the distance between the clusters and, according to the frequency, the size of them.

[7] Orange Visual Programming. (n.d.). Silhouette Plot. Retrieved March 8, 2023, from https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/visualize/silhouetteplot.html

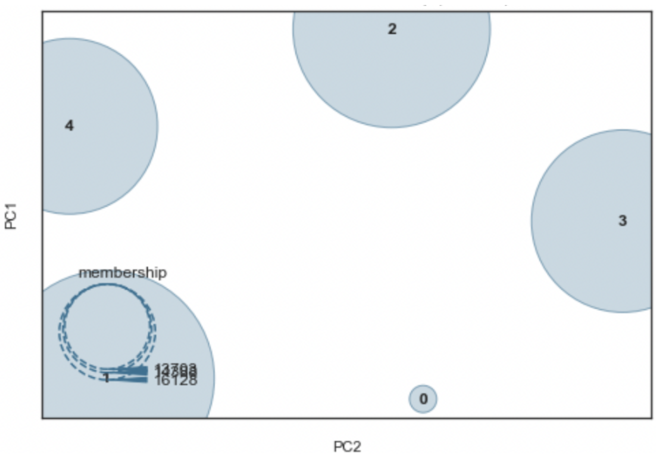


***Figure 8 -*** *"K-Means Inter Cluster Distance Map"*

The weighted values per feature for all 20 principal components was calculated, but the weighted importance was evaluated only for the 11 cumulative components, with a focus on component 11.

# 4. RESULTS EVALUATION

Finally, we performed the assessment of the model using the R2 score, which represents the proportion of variance in the cluster labels. The final R2 score for the K-Means clustering solution was 0.6413, indicating that the clustering solution explains about 64% of the variance in the data.

## 4.1 Assess Model

Also we performed the R2 per variable and it was possible to see that the variables with the highest R2 scores were age segments, explaining the majority of the variance in the clustering solution, contributing the most to the clustering solution. Therefore, running the model was a good solution for clustering the data based on the variables used.

## 4.2 Interpreting Clustering Algorithms

The clustering algorithm has successfully identified five groups of customers with distinct characteristics. By analyzing the final clusters, we can gain insight into the customers preferences, behavior, and booking patterns. This information can be used to improve marketing strategies and customer experience, ultimately leading to increased revenue for the hotel**.**

### 4.2.1 Evaluation

| Clusters | Main Characteristic |
|---|---|
| **Cluster 0:**<br>**Younger people with more wealth** | - Age_gap: 20-30<br>- The highest value of  French Nationality<br>- The lowest value of AverageLeadTime<br>- Distribution Channel: Travel Agent/Operator<br>- The second highest value of SRKingSizeBed and SRTwinBed |

| | |
|---|---|
| | - The lowest value of SRQuietRoom<br>- The highest value of AvgRevenueperYear<br>- The highest value of Revenueperbooking<br>- The lowest frequency and magnitude |
| **Cluster 1: Portuguese Clients** | - Age_gap: 41-50<br>- The highest value of  Portuguese Nationality<br>- Distribution Channel: Corporate  and Direct<br>- The highest frequency |
| **Cluster 2: Potencial and Mature Clients** | - Age_gap: 51-60<br>- Within the nationality classes, the one with the highest value of nationality is Others<br>- Distribution Channel: GDS Systems<br>- The highest value of SRQuietRoom<br>- The second highest value Revenueperbooking and AverageLeadTime<br>- The second highest frequency |
| **Cluster 3: Advanced Booking Customers** | - Age_gap: 60+<br>- The highest value of  German Nationality<br>- The highest value of AverageLeadTime<br>- Distribution Channel: Travel Agent/Operator<br>- The highest value of SRTwinBed<br>- The second highest value of AvgRevenueperYear |
| **Cluster 4: Low cost travel** | - Age_gap: 31-40<br>- Within the nationality classes, the one with the highest value of nationality is Others<br>- Distribution Channel: Direct (the second highest value)<br>- The highest value of SRKingSizeBed<br>- The lowest value of AvgRevenueperYear<br>- The lowest value of Revenueperbooking |

**<u>Cluster 0: Younger people with more wealth:</u>**

The main characteristic of this cluster is the age. These clients are young and have no troubles with the noise once they present the lowest value for SRQuietRoom. Usually they travel alone or in couples and don't book in advance, once the cluster presents the lowest value on AverageLeadTime. These customers are the most valuable once they spend more money per year and per booking than any other.

**Suggestions:**

- Through young social media influencers promote the hotel accommodation, services and experience to the younger people with acquisitive power worldwide and specifically on the French market as well.

- Create a specific and exclusive offer in terms of accommodation, service and experience to be promoted through the Travel agent/operator distribution channel worldwide, but focusing on the French market.

**Cluster 1: Portuguese Clients:**

The main characteristic of this cluster is to have a strong presence of Portuguese. Furthermore, these customers prefer to book directly with the hotel, without intermediaries, and are the most loyal ones to our hotels once they book frequently.

**Suggestions:**

- Create a media plan to the Portuguese market involving television and radios to promote the brand and increase notoriety.
- Recognize their frequent presence through some surprises and gentle offers during their stayings at the hotel.

**Cluster 2: Potencial and Mature Clients:**

The main characteristic of this cluster is the highest value of SRQuietRoom. So, these clients require a calm environment. The age of these customers is majorly between 51-60 and they present a high acquisitive power once they are the second highest value for revenue. They also make their reservations with time in advance.

**Suggestions:**

- Create a loyalty program with focused and tailored offers in order to promote loyalty and retain the customers;
- Design a tailored offer in terms of accommodation, service and experience in order to be promoted through the channels GDS Systems and Travel Agent/Operator.

**Cluster 3: Advanced Booking Customers:**

The main characteristic of this cluster is that the clients book with the larger antecedence compared with the other customers. German is the predominant nationality and Age_gap 60. Another characteristic is age, most likely they are retired customers. They may already have a financial reserve which contributes to early bookings.

**Suggestions:**

- Create a partnership with travel agencies in order to organize trips to aged customers with tailored experiences and activities;
- Design a campaign to aged customers promoting opportunities to travel with buddies.

**Cluster 4: Low cost travel:**

This cluster includes customers with ages between 31 and 40 from everywhere in the world, the ones with the lowest acquisition power.

**Suggestions:**

- Promote offers with discounts to middle age people in the less booked period of the year throughout the main distribution channels;
- Create a tailored offer for customers in middle age and sensible to prices.

## 5. DEPLOYMENT AND MAINTENANCE PLANS

After modeling and describing client segmentation, the next step consists in delivering the results and suggestions to the main decision makers as financial and marketing areas.

Firstly, the model should be executed quarterly to assess the changes, mainly to verify those related with features that use revenue in the composition. This part is also important to capture seasonal behaviors and external factors that influenced the period. The results can be accessed by Google Sheets and the analyzed data can be described by the consultants in a quarterly report delivered by e-mail to the decision makers. The report must explain the model in an appropriate business communication, so that the other areas can analyze and have a good understanding of the process.

As a suggestion for future implementation, an internal chatbot could be created for decision makers to consult the results. The chatbot can be developed using the Telegram bot and marketing coordinator, for instance, can write and ask for information related to the main characteristics already predefined. The chat offers quarterly updated information, but in a reduced form compared to the report. The advantage is that the implementation is fast, accessible and has low costs. The access to the key information can be consulted more quickly than in the report. Both, report and chatbot, should be complementary and not exclusive.

Furthermore, there is an annual report where all data information can be revisited and new results and insights are generated. Thus, decision makers utilize this report to maintain, improve or change completely merchandising, financial and operation topics.

## 6. CONCLUSIONS

Nowadays, more than ever it's important to take advantage of customer segmentation and clustering techniques, which are fundamental for businesses that are looking forward to better understanding their customers and developing more effective marketing strategies, and ultimately generating more revenue. By leveraging the power of data-driven analytics and decisions, businesses can gain a competitive edge in today's crowded hospitality industry.

From our final clustering solution, it was possible to conclude that the most profitable customers are from Cluster 0, 2 and 3, nonetheless Cluster 1 is of extreme importance, given their frequency. Cluster 4, out of all the clusters, were found to be somehow complicated - low profits and frequency.

Investing in marketing for this segment wouldn't be a wise decision, when there are segments that bring the hotel chain much more revenue and bookings. In the near future, it would be advisable to focus on marketing campaigns directed to Cluster 1, which is the most frequent segment.

### 6.1 Considerations for model improvement

After reviewing the data quality during the data exploration we noticed some inconsistencies that required attention. Therefore, in order to improve the quality of the model, we strongly recommend the implementation of processes for data input that are distributed among the hotel workers.

Additionally, it is recommended to establish rules within the system to prevent inconsistencies, such as the inability to input negative values for age, average lead time and duplicate rows for namehash. These steps will help ensure the accuracy and consistency of the data used for the model, ultimately leading to improved model performance.

## 7. REFERENCES

(Brownlee, Jason. (2008, August 13). Knn Imputation for Missing Values in Machine Learning

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide. SPSS Inc.

OpenAI. (2021). GPT-3: Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165

Scikit-learn. (n.d.). K-means clustering: Silhouette analysis. Retrieved March 7, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

Orange Visual Programming. (n.d.). Silhouette Plot. Retrieved March 8, 2023, from https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/visualize/silhouetteplot.html

Yellowbrick. (n.d.). SilhouetteVisualizer. Retrieved March 8, 2023, from https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html

## 8. APPENDIX

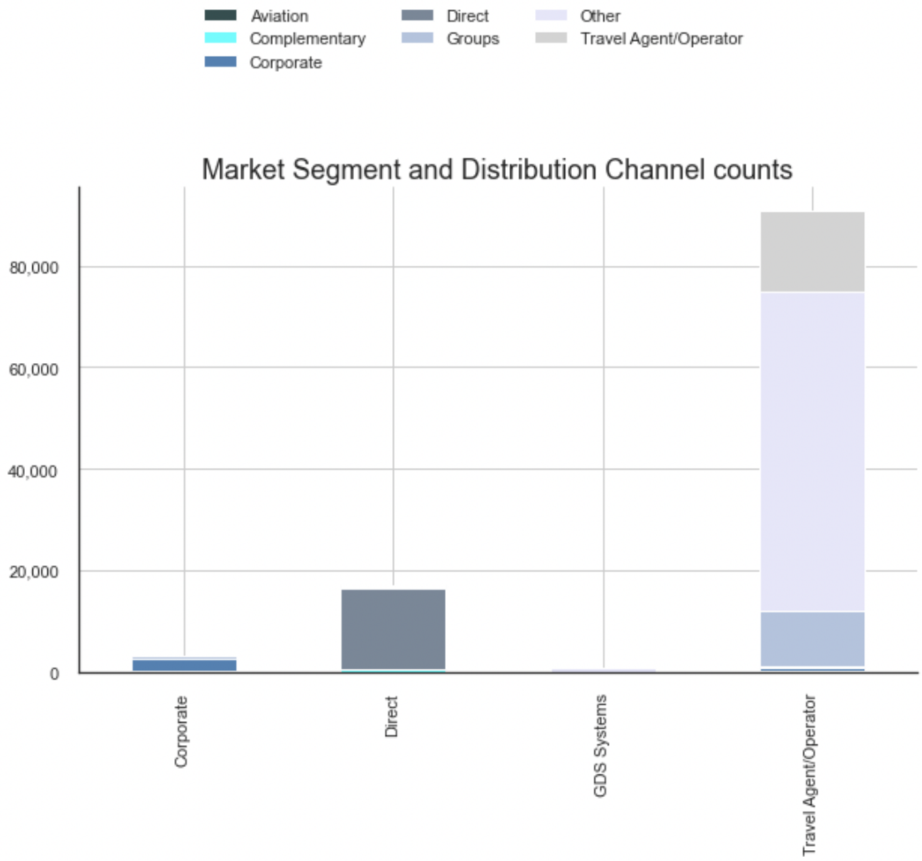


***Figure 1 -*** *"Market Segment vs Distribution Channel"*