NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel Customer Segmentation

Bernardo, Pinto Leite, number: 20230978

Emília, Santos, number: 20230446

Nicolás, Zerené, number: 20230779

Ricardo, Kayseller, number: 20230450

Stephan, Kuznetsov, number: 20231002

Group F

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

Nowadays, the hotel industry faces intense competition in attracting new customers, as tourists increasingly seek out budget-friendly accommodation options. With the steep rise of online marketplaces such as Airbnb, Booking.com, and many others, including Local Accommodation Establishments (AL), it's essential for each and every hotel to develop its own business model, relying on a thorough analysis of customer preferences and offering customized products based on segmentation. According to Philip Kotler, “market segmentation is the sub-dividing of a market into homogeneous sub-sections of customers, where any sub-section may conceivably be selected as a market target to be reached with a distinct marketing mix.”

Our project introduced customer segmentation for an urban Hotel H in Lisbon, which was expanding its business through the acquisition of new hotels.

It was important to analyze and understand the actual customers, to gain a better understanding of which kind of clients were choosing the hotel. In the research, data was used based on information from bookings and reservations for the last years. A more advanced segmentation strategy for the Hotel H chains was being developed, focusing on revenue contribution, demographics, geography, and consumer behavior.

The first checkpoint was to prepare this data, to create different kinds of clusters which could provide the company with an idea of different architypes for every client who stayed in Hotel H.

For the analysis, PCA was applied to achieve a different visualization due to the reduced dimensionality. The separation of the perspectives was accomplished by a process of K-means, which eventually highlighted the types of customers with bookings in the hotel. In the final, we applied a K-prototypes model to categorize our dataset into three distinct customer perspectives: Demographic, Behavioural, and Geographic. This segmentation allows for targeted insights and strategies tailored to each group's unique characteristics and preferences, enhancing customer engagement and service customization.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BUSINESS OBJECTIVES

Given the competitive landscape, the aim is to leverage advanced data science techniques to understand customer behavior and preferences deeply, thereby enhancing customer satisfaction, loyalty, and help defining the marketing strategies to achieve new and old customers on a overall business profitability perspective. The following objectives are designed to guide the implementation of customer segmentation strategies:

2.1.1. Enhance Customer Satisfaction Through Personalized Services: The primary goal is to increase customer satisfaction by offering personalized experiences based on customer segments. Analyzing customer data enables the identification of unique preferences and behaviors, allowing for the tailoring of services and communications to meet individual needs.

2.1.2. Optimize Marketing Efforts and Increase Revenue: By segmenting the customer base, marketing strategies can be more effectively targeted, leading to higher conversion rates and increased revenue. Segmenting customers based on characteristics such as demographics, purchase history, or engagement patterns allows for more precise and impactful marketing campaigns.

2.1.3. Improve Product and Service Offerings: Customer segmentation provides valuable insights into the needs and preferences of different customer groups. This knowledge can be used to inform the development of new products and services or to enhance existing offerings, ensuring they are more closely align with customer expectations and improve market competitiveness.

2.1.4. Drive Operational Efficiency: Understanding customer segments allows for more efficient allocation of resources, including staff, inventory, and marketing budgets. Tailoring operations to meet the specific needs of different customer segments can reduce waste, lower costs, and improve overall operational efficiency.

2.1.5. Define Strategic Decisions: Data-driven insights obtained from customer segmentation analysis can support strategic decision-making regarding business expansion, investment, and resource allocation. Identifying high-value customer segments and understanding their behavior and preferences can guide the strategic direction of the business towards areas of growth and opportunity.

2.1.6. Foster Customer Loyalty and Retention: Personalized experiences and targeted marketing efforts contribute to increased customer satisfaction, which is crucial for building loyalty and encouraging repeat business. By engaging customers in a manner that resonates with their specific needs and preferences, businesses can significantly improve customer retention rates.

2.1.7. Leverage Competitive Advantage: In the highly competitive hospitality industry, the ability to effectively segment and target customers can provide a significant competitive advantage. Businesses that excel in understanding and meeting the unique needs of their customer segments are better positioned to attract and retain customers, ultimately leading to greater market share and profitability.

2.1.8. Streamline Hotel Operations by Synchronizing Resources with Customer Needs: Sharpen operational agility by meticulously aligning staff, amenities, and services with the nuanced demands of identified customer segments. This synchronization ensures that each aspect of the hotel's functioning is responsive to the real-time needs and expectations of guests, thereby optimizing resource utilization, reducing redundancy, and increasing guest satisfaction through tailored service delivery.

2.2. BUSINESS SUCCESS CRITERIA

A hotel is a business entity that offers lodging, meals, and other services in exchange for payment, aiming to generate high revenue. For a hotel chain to thrive, it must focus on several key areas:

1. Implementing an efficient booking system;
2. Ensuring top-notch accommodations;
3. Prioritizing safety and security;

4. Effective management;
5. Providing excellent customer service;
6. Adhering to sustainability practices;
7. Engaging in effective marketing and branding.

In the competitive hospitality sector, hotels must continuously adapt to the changing demands of customers and market conditions. Achieving success hinges on prioritizing guest satisfaction while maintaining operational efficiency and profitability. It's vital for hotels to align their services with guest expectations and market trends.

Customers are the lifeblood of the hotel industry, as their patronage is essential for the industry's success. Without customers, hotels cannot operate, regardless of the quality of staff or accommodations. Therefore, it's crucial to retain existing customers and attract new ones. Understanding customer preferences, expectations, and behaviors is key to this. Conducting thorough customer analysis can reveal what customers value in a hotel, their expectations, reasons for returning, and factors influencing their loyalty.

Customer segmentation is a powerful strategy to boost hotel revenue by targeting specific customer groups with tailored marketing campaigns. However, segmentation based solely on one characteristic is insufficient. The goal of this project is to enhance customer segmentation by incorporating more distinctive customer characteristics. This will lead to more personalized marketing campaigns, resulting in higher customer satisfaction and repeat business.

2.3. SITUATION ASSESSMENT

Regarding the inventory resources needed to achieve this analysis:

- **Personnel**: hotel chain board, marketing manager and marketing department, BI team, data mining experts, and the database administrator, if necessary;
- **Data**: dataset (CSV file: Hotel Customer Segmentation) provided by the hotel chain;
- **Computing/software resources:** Python or other computing language capable of doing a clustering analysis and data mining tools.

Requirements, assumptions and constraints should also be considered when assessing a situation:

- **Requirements:** comprehend, prepare and preprocess the dataset, do the modelling and deployment. Having done this alongside a report, gather the insights and apply the strategies.
- **Assumptions:** people under 16 years old are not suitable for the segmentation, given that cannot be targeted in terms of business and marketing;
- **Constraints:** treatment of some variables and quality of devices and software used to develop the clustering segmentation (not time efficient).

In terms of risks and contingency actions, some were denoted:

- **Risks**: data not well prepared/pre-processed, not well-defined clusters, and, consequently, marketing campaigns/strategies that don’t correspond to the actual clients and their needs;
- **Contingencies:** re-prepare the data, as well as develop a new clustering segmentation (obtaining new insights on each cluster) and, afterwards, change the strategies.

Considering costs and benefits related to developing and implementing this project:

- **Costs**: software related costs, in order to have the most helpful, precise and well driven campaign the software should be the best. Operational costs also must be taken into account, as well, as costs for the marketing strategies.
- **Benefits**: precise strategy and increase in revenue, as well as having new clients and maintaining and improving the satisfaction of the ones that already book with this chain.

2.4. DETERMINE DATA MINING GOALS

Data mining is the process of determining important information from a large dataset. It's also known as knowledge discovery in data (KDD). Data mining has improved organisational decision-making through insightful data analyses, using several algorithms for the appropriate objective (IBM, 2023).

In the case of this project, we wanted to look at customer segmentation. To help create better groups (or segmentations), we used a partitioning algorithm.

This partitioning algorithm was K-means, which is an iterative method and used to cluster points into groups represented by a centroid. Every data point is assigned to the closest centre (Ahuja et al., 2022). This algorithm creates clusters by separating a number of samples and then it tries to minimize the inertia (sum of squares).

K-means will always converge, sometimes to a local minimum. The convergence depends on the initialization. The way to solve this is by initializing the clustering at random centroids. We use the initialization method 'k-means++' that initializes the centroids to be distant from each other, leading to better results than random initialisation (2.3. Clustering — Scikit-Learn 0.20.3 Documentation, 2010).

3. METHODOLOGY

3.1. DATA UNDERSTANDING

3.1.1. DATA COLLECTION

The dataset provided an overview of hotel customer behavior and preferences, consisting of 111,733 observations across 29 diverse variables. These variables ranged from text and numeric to categorical types, creating an overview of customer profiles. The dataset had a high data integrity for analysis with 0.2% of missing values and a very low duplication rate, providing a strong basis for our further analysis (annex 1).

3.1.2. DATA EXPLORATION

After some initial data preprocessing and setting the "ID" as the index, it was discovered that the dataset contained 111 duplicates, which were subsequently removed. To expand the customer base and to maximize revenue, the clients were sorted by their minimum age criterion to optimize the strategies effectively to demographic patterns. Moreover, through the analysis of booking patterns and

special requests, strategies were being developed to increase revenue via personalized services and promotions.

ln order to make future steps of the analysis, “DocIDHash” feature was dealt with: there were a lot of rows with same “DocIDHash”, which should be an unique value for each customer. Regarding duplicates, the dataset was cleaned based on a specific combination of fields, leaving only unique entries according to the criteria of “Age”, “NameHash”, “DocIDHash”, and their occurrences. This process revealed the true number of customers.

Some further analysis showed that certain preferences, such as low floors, accessibility features, and specific locations accounted for less than 1% of total requests. They were removed from the analysis, due to the significant imbalance in these preferences, to focus on more impactful attributes. King-size and twin beds, along with quiet rooms, were in higher demand. Features with less demand were overlooked and attention was concentrated on those that had a greater chance of increasing customer satisfaction and making business more successful.

In the review of the hotel's booking sources, it was found that a significant majority of our guests were referred to by travel agencies and tour operators. Comparatively, a much smaller number of guests made their reservations through corporate channels or Global Distribution Systems (GDS).

3.2. DATA PREPARATION

3.2.1. AGE VARIABLE

The data preparation phase began by removing all observations from the dataset for individuals under 16 years of age and in a next step, those who are above 90 years old. The rationale behind excluding individuals under 16 is based on the understanding that they are not the primary decision-makers for hotel bookings and thus may not significantly influence the choice of hotel or room amenities. Similarly, individuals who are 90 years old or older were removed due to their outlier status in the context of hotel segmentation, which allows the analysis to concentrate on the primary market demographic that makes booking decisions. This focused approach is expected to yield more accurate insights into customer preferences and behavior patterns, which are directly relevant to business outcomes.

In a subsequent step, the “Age” variable was encoded into 6 categories, following an analysis of its distribution (annex 2). This categorization offers an analytical advantage by simplifying the complexity of the age variable, making it easier to identify patterns and trends within distinct age groups. For predictive models, this approach enhances interpretability and could potentially improve model performance by reducing noise and concentrating on age-related segments that hold more significance for our analysis.

3.2.2. MISSING VALUES

Additionally, the dataset contained 4092 missing values (NaN) in the age variable. These missing values were filled with the median age to maintain the integrity of our data analysis. Choosing the median as the fill value helps mitigate the influence of outliers and maintains the distribution's central tendency, which is particularly advantageous in preserving the robustness of our dataset. The variable

"DocIDHash", with 932 missing values, was initially treated by replacing the (NaN) with the mode. However, it was later dropped from the dataset as it was deemed irrelevant for the analysis.

3.2.3. SERVICE REQUEST VARIABLES

In our hotel segmentation data analysis, several highly imbalanced variables were identified, which do not offer meaningful insights into our model preferences and behaviors. Specifically, variables indicating preferences for low floors, accessible rooms, bathtubs, and others were analyzed based on the percentages of their binary responses (0s and 1s), revealing that over 99% of responses fell into the single category (0).

This analysis helped identify the most imbalanced variables, leading to their exclusion from further analysis to ensure a focus on more impactful and balanced data for strategic decision-making. The following variables were dropped from our dataset as they were identified to be highly imbalanced and not providing significant insights: "SRLowFloor", "SRMediumFloor", "SRBathtub", "SRShower", "SRNearElevator", "SRAwayFromElevator", "SRNoAlcoholInMiniBar", and "SRAccessibleRoom", removing these variables enhanced the precision of our analysis and the effectiveness of our operational strategies, ensuring the focus on the preferences and needs that truly matter to our guests and to our business performance (annex 3).

3.2.4. OUTLIERS REMOVING

The strategy for removing outliers from key variables, "Age", "AverageLeadTime", "LodgingRevenue", "OtherRevenue", "BookingsCheckedIn", and "RoomNights", involved using boxplot visualization and calculations based on the interquartile range (IQR) to establish specific thresholds for outlier exclusion. This method effectively filtered out anomalous data points, ensuring that the remaining dataset accurately represented common booking behaviors and revenue patterns, which are essential for a comprehensive analysis. The process retained 99% of the original data, striking a balance between cleaning the dataset and preserving its integrity for meaningful insights.

3.2.5. CORRELATED VARIABLES

During the analysis, it was observed that the metric variables "RoomsNights" and "PersonsNights" were highly correlated with a correlation of (0.87). Since the number of nights in a hotel is an essential metric for the analysis and segmentation of the business case, we decided to drop the "PersonsNights" variable. Continuing our analysis of the correlation between variables, regarding now the non-metric variables, it was observed that with a (0.70) correlation, the variables "MarketSegment" and "DistributionChannel" are moderately correlated, and this was confirmed by the chi-square test with a result of zero, this suggests there is a very strong statistical significance in the association between the two variables. Therefore, since market segments may change in the future, unlike distribution channels, which rarely change, the "MarketSegment" variable was dropped from the analysis (annex 4).

3.2.6. REDUCING NUMBER OF CATEGORIES

During the feature engineering phase, to reduce some of the less important categories in the analysis, we decided to consolidate the categories of certain variables after evaluating their distribution;

- “Nationality”: This variable was reduced to the top 6 nationalities, assigned with codes ranging from 1 to 6, and grouped the remaining countries into a single category (0). During the analysis, it was found that adding one more category to the model, in this case, due to the highest frequency in the segment, category (USA), substantially improved the model in terms of cluster distribution (frequency);
- “DistributionChannel”: This was reduced to the 2 main categories, with the remaining categories included in a catch-all category (3).

By doing this, there was a significantly decrease the number of categories in both variables, allowing for a more detailed analysis focused on relevant and objective results. This approach simplifies the model, potentially improving its interpretability and performance by concentrating on the most influential factors.

3.2.7. NEW VARIABLES

Variables such as “TotalRevenue”, “CancellationRate”, and “TotalSpecialRequests” were developed to refine our analysis. “TotalRevenue” combines “LodgingRevenue” and “OtherRevenue”, offering a comprehensive measure of a customer's value and aiding in segmentation, reducing those variables to one. “CancellationRate”, the ratio of canceled bookings to total bookings, provides insights into booking reliability, with “BookingsCanceled” being dropped for efficiency. “TotalSpecialRequests” aggregates specific preferences, enhancing service personalization and satisfaction. These adjustments streamline the model, focusing on key insights for targeted strategies and customer understanding.

3.2.8. DUPLICATED VALUES

There were 111 duplicate values within the dataset, which were subsequently removed to ensure data integrity. Additionally, within the “DocIDHash” variable, some instances where the same code was associated with observations were identified, having different nationalities and “NameHash” values. These inconsistencies were also removed from the dataset to maintain accuracy and reliability in this analysis.

3.3. MODELING

The process began with ensuring the data was prepared, setting the stage for the clustering analysis. To enhance visualization and reduce the complexity of the data, Principal Component Analysis (PCA) was utilized prior to applying the k-means clustering algorithm. The selection of the optimal number of principal components was guided by two plots: a scree plot and one illustrating the explained variance. These were analyzed using the elbow method, a technique that helps identify the best number of clusters for the k-means algorithm. To further understand the algorithm's performance, additional graphs were created, including a cluster magnitude and a cluster cardinality plot, considering one to contrast both kind of metrics.

The labels generated from the clustering process were integrated into the original data frame to preserve the insights gained. Following the application of the k-means algorithm, the resulting views were combined to form a comprehensive clustering solution that encompassed all the dataset's information.

The distortion score plot, a metric measuring the average distance between data points and their cluster centroids, indicated that five clusters were optimal for the data. The goal of clustering is to minimize this distance, aiming for a more accurate and meaningful grouping of the data. But also, using the Sillhouette method, with a reduced dimensionality, said it should be 4, so in the case of the model K-means clustering was then performed using 4 clusters. This given, that the distortion score just express the distance between the point and the centroid, considering that the Silhouette score looks for the closest points and also the nearest closest center.

3.4. EVALUATION

In terms of the first method applied for dimension reduction – **PCA** (Principal Components Analysis) - instead of having 36 dimensions, the analysis was continued with only 18, which represents 97.63% of the cumulative variance explained (annex 5). Although it does not correspond to 100%, it is still a representative amount of variance and with a much lower number of dimensions, easing the process that comes afterwards.

Regarding **K-means** algorithm, using 4 clusters and 18 dimensions, the Silhouette was used method to validate the consistency of the clusters. The score was approximately 0.4, which is optimal for this metric (annex 6). Cardinality and Magnitude of the clusters seem to be correlated and, so this leads to the conclusion that there are no major anomalies on the 4 clusters (annex 7). Although the Silhouette method leading to a solution of 4 clusters, which was the option that was chosen (silhouette is in this case the best and most accurate method), according to the elbow method it could be 5 clusters (annex 8).

Concerning the other algorithm developed - **K-Prototypes** - to a segmentation by 3 perspectives (Demographic, Behavioral and Geographic), which gave a total of 5 clusters (see annex 9 – t-SNE, annex 10 – elbow method). The Silhouette score was approximately 0.2, lower than the one obtained from K-means, concluding that the latter is the best method in this context. Although the Silhouette Score was not the best, considering that for this same algorithm a Calinski-Harabasz Index of 6554.87 (the higher, the better) and a Davies-Bouldin Score of 2.65 (the lower, the better), it is indicative that a good clustering definition was made. Lastly, to complement these other scores, an Adjusted Rand Score (scale between -1 and 1) was implemented, its score being 0.66 (considered good).

4. RESULTS EVALUATION

After cleaning and modifying the dataframe, and applying all the methods explained above, it was possible to place the existent customers of the Hotel chain H into 4 distinct groups (annex 11).

Ou of the 100049 customers, 36 174 people went to Cluster 0, 32763 people went to Cluster 1, 12 124 went to Cluster 2 and the rest 18 988 were placed in Cluster 3 (see annex 12). Below are some important features of each cluster (annex 13).

Cluster 0: This cluster shows people from other parts of the world (nationality code 0). The customers of this group tend to be between the ages of 25 and 35. This group shows the most Total special

Requests (2 and 3), however, no requirements for twin beds, Higher floor, quiet room or cribs, so it can be assumed that this group represents young workers that stay in this hotel for business reasons.

Cluster 1: This group is predominantly composed of people with French nationality, followed by German nationality. It characterises the people from central Europe. Here it can be seen that the age distribution is well spread. However, this group is shown to spend the least amount of money (Total revenue is the lowest of all groups ~342,88€) and tend to ask for quieter rooms compared to the other clusters. Customer from this cluster don't ask for twin beds. These are also the people that tend to book more rooms per night through Travel Agencies/Operators.

Cluster 2: The customers from this group are mainly of Portuguese nationality and are people between the ages 56 and 65. These people tend to book directly with the Hotel or by other method. These customers usually spend more money (Total revenue is the highest ~372,76€). This group looks for cribs the most, however, twin beds are not required at all. They also, usually, book the rooms in very short notice (on average only 41 days before arrival) and the highest number of check-ins.

Cluster 3: This cluster describes the customers of the extreme age groups. This is the majority of people in this group are 66 and over and are younger than 25 years old. This group doesn't spend a lot (second least highest total revenue) but it usually books the rooms in advance the most (74 days before arrival). However, shows the least number of bookings checked-in compared to the other clusters. People from this cluster are more likely to ask for a twin bed and rooms at a higher floor, and one total special request.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. Plan Deployment

Deployment and maintenance phases were important stages in the customer segmentation project that assured the effectiveness, relevance, and impact of the marketing strategies based on segmented customer data.

The deployment plan to benefit from the customer analysis involved strategically targeted marketing and service packages to meet the needs, observed in the four unique customer groups.

Targeted Marketing: offer business deals for the young workers in cluster 0, create budget-friendly options for Europeans in cluster 1, promote extras for the well-spending Portuguese seniors in cluster 2 and have special early deals and rooms with easy access of booking for the mix of young and elderly clients in cluster 3.

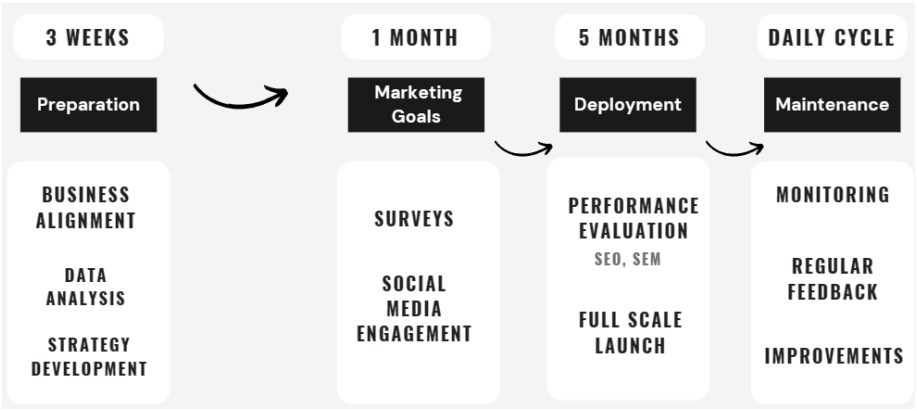


Figure 1 – Deployment Plan

5.2. Maintenance Plan

Cluster 0: Since clients in this cluster are mostly international business travellers aged 25-35 with high instance of special requests, priority should be to fulfil the unique needs. Offer them special amenities and discounts such as extended room service hours to match the business traveller's schedule, ensure optimal internet connection and availability of business amenities.

Cluster 1: Customers within this cluster tended to spend less and often book through travel agencies, targeting them with budget-friendly packages through these channels could bring positive outcomes. Given their preference for quieter accommodations, marketing this feature was essential. Highlighting existing quiet zones or floors within the hotel could serve as a compelling selling point for this demographic.

Cluster 2: To improve the customer experience and favour loyalty within this group, providing direct booking incentives could lead to increased profits, given their preferences for booking directly. Given their tendency for higher spending, premium services or exclusive offers should be offered.

Cluster 3: Future maintenance strategies should focus on enhancing room safety, comfort, and accommodating specific preferences like twin beds and higher-floor rooms. Comprehensive staff training and proactive room preparation are recommended to ensure these guests receive customized services that will definitely improve their experiences.

Regular updates: The hotel should periodically update their customer segmentation to keep up with evolving preferences and patterns. Data from past reservations could improve guest profiles, securing the maintenance of high-quality services.

Guest Feedback: The hotel should seek out opinions from guests regarding their room preferences and experiences. Guests might not have been fully aware of certain room features they could request. With this feedback, the hotels should refine their service offerings to meet guest requirements more precisely.

6. CONCLUSIONS

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

6.1.1 Collection of New Types of Information

Customer Feedback: Incorporating data on customer satisfaction, including ratings and comments about their stay. This can help identify factors influencing loyalty and customer preferences.

Online Behavioral Data: Tracking customer interactions with the website and online booking platforms, including clicks, visited pages, and time spent. This data can reveal preferences and booking intentions, enriching segmentation.

Socioeconomic Data: Additional information about the socioeconomic background of customers, such as profession, family income, and educational level, can help refine market segments.

Travel Preferences: Details about travel preferences, including the desired type of accommodation, activities of interest during the stay, and reasons for the trip (business, leisure, etc.) can provide a basis for more personalized segmentation.

6.1.2 Improvements in Data Quality

Data Standardization: Ensuring that all collected data follow a consistent format to facilitate analysis. This includes standardizing country names, income categories, and other categorical variables.

Real-time Data Validation: Implementing tools or processes to validate data at the time of entry, especially for data collected online or through direct interactions with the customer. This can help reduce errors and inconsistencies.

Regular Data Audits: Establishing a process for periodic reviews of data quality, including checking for duplicates, correcting entry errors, and assessing data completeness.

Collaborative Data Sharing Initiatives: Engaging in or establishing data-sharing partnerships with other organizations to enrich the dataset further. This could involve industry consortia or public-private partnerships that allow access to a broader set of anonymized data, improving the model's predictive power while maintaining privacy standards.

7. REFERENCES

- 2.3. Clustering — scikit-learn 0.22.1 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/clustering.html#clustering
- Ahuja, K., Nayyar, A., & Sharma, K. (2022). Comprehensive Guide to Heterogeneous Networks. Academic Press.
- Interpretar resultados e ajustar o clustering. (n.d.). Google for Developers. https://developers.google.com/machine-learning/clustering/interpret?hl=pt-br
- IBM. (2023). What is Data Mining? | IBM. Www.ibm.com. https://www.ibm.com/topics/data-mining

8. APPENDIX

Annex 1 - Dataset Overview

Overview | Alerts 32 | Reproduction

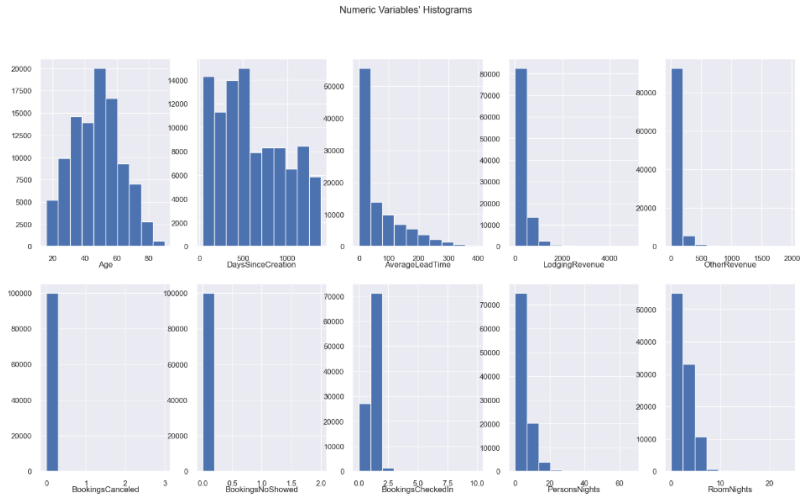
Dataset statistics

Number of variables	28
Number of observations	111733
Missing cells	5173
Missing cells (%)	0.2%
Duplicate rows	89
Duplicate rows (%)	0.1%
Total size in memory	24.7 MiB
Average record size in memory	232.0 B

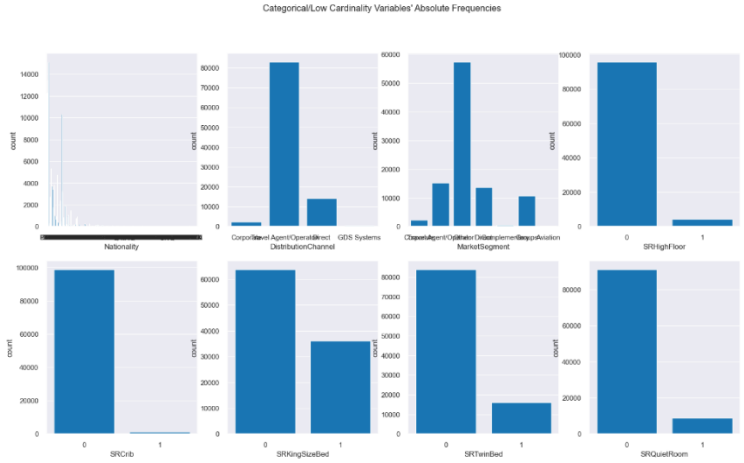
Variable types

Text	3
Numeric	9
Categorical	16

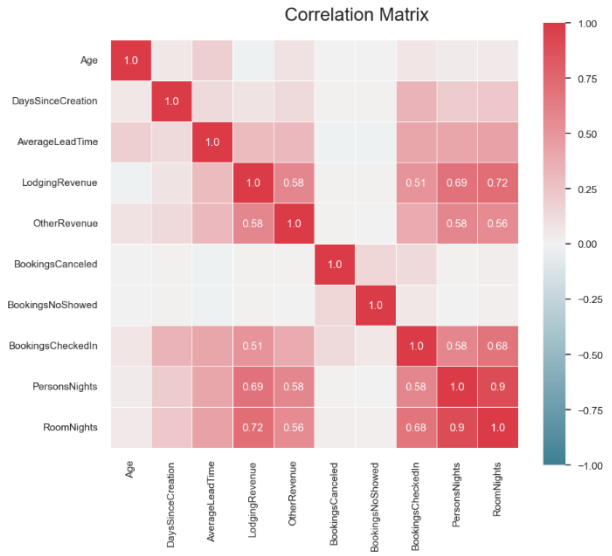
Annex 2 – Metric Variables



Annex 3 – Non-Metric Variables



Annex 4 – Correlation matrix

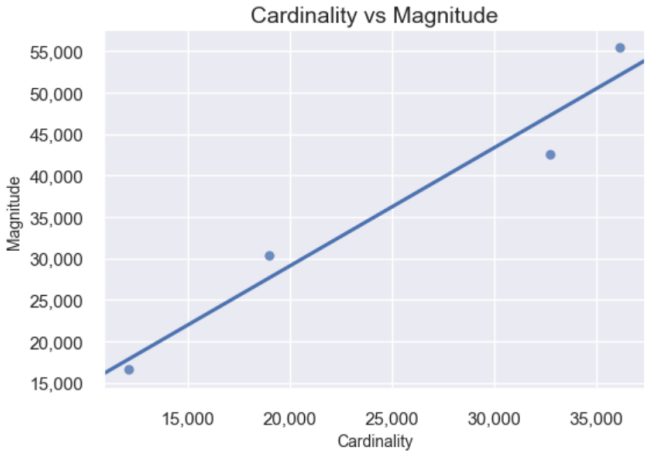


Annex 5 – PCA (Principal Components Analysis)

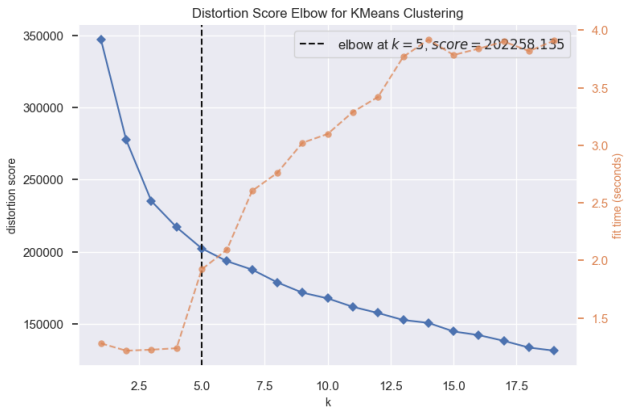
	Component	Variance explained	Cumulative variance explained
0	1	0.211390	0.211390
1	2	0.116002	0.327392
2	3	0.080684	0.408076
3	4	0.070959	0.479036
4	5	0.067968	0.547004
5	6	0.063019	0.610023
6	7	0.051970	0.661993
7	8	0.048577	0.710570
8	9	0.041303	0.751873
9	10	0.036993	0.788865
10	11	0.032177	0.821042
11	12	0.029760	0.850802
12	13	0.027354	0.878156
13	14	0.022481	0.900636
14	15	0.018105	0.918741
15	16	0.017253	0.935994
16	17	0.015692	0.951687
17	18	0.013283	0.964970
18	19	0.011320	0.976290
19	20	0.009984	0.986274
20	21	0.007300	0.993574
21	22	0.003441	0.997015
22	23	0.002290	0.999306
23	24	0.000412	0.999718
24	25	0.000268	0.999986
25	26	0.000008	0.999994
26	27	0.000006	1.000000
27	28	0.000000	1.000000
28	29	0.000000	1.000000
29	30	0.000000	1.000000
30	31	0.000000	1.000000
31	32	0.000000	1.000000
32	33	0.000000	1.000000
33	34	0.000000	1.000000
34	35	0.000000	1.000000
35	36	0.000000	1.000000
36	37	0.000000	1.000000

Annex 6 - K-means (Silhouette Score)

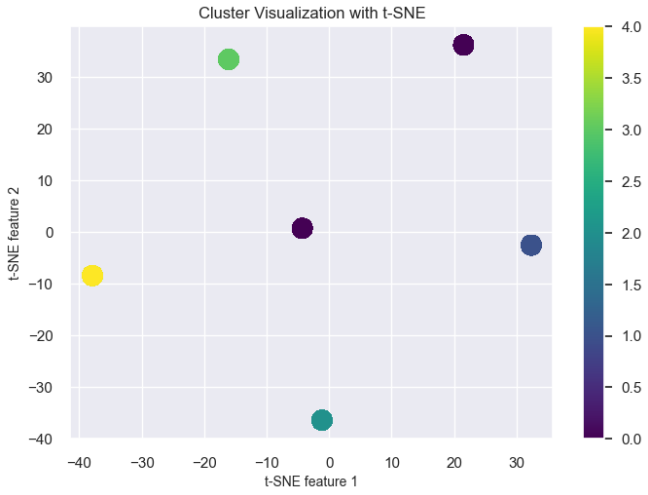
Annex 7 - K-means (Cardinality Vs Magnitude)



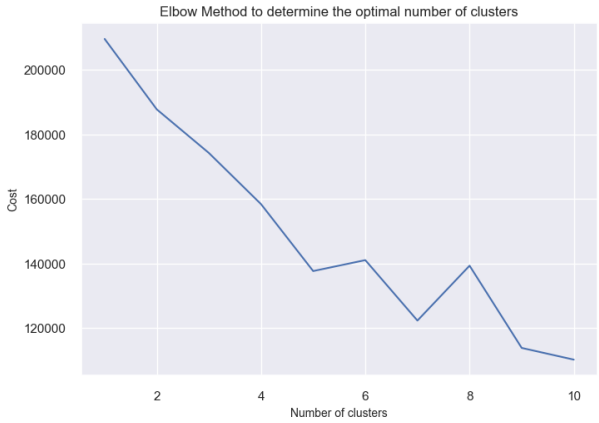
Annex 8 – K-means (elbow method)



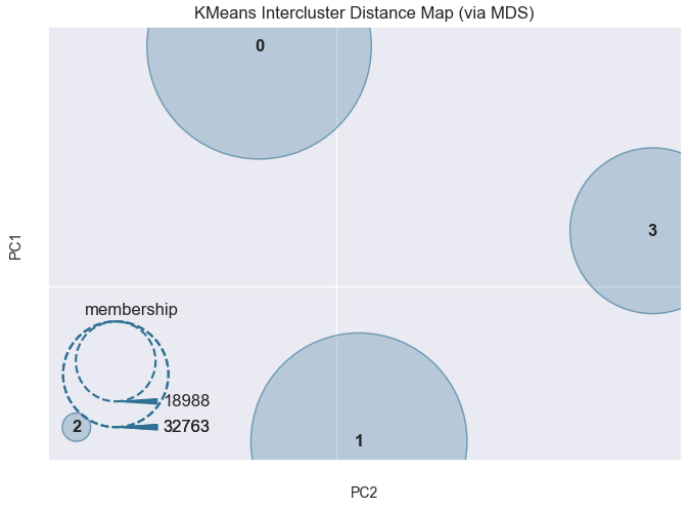
Annex 9 – K-prototypes (4 cluster t-SNE visualization)



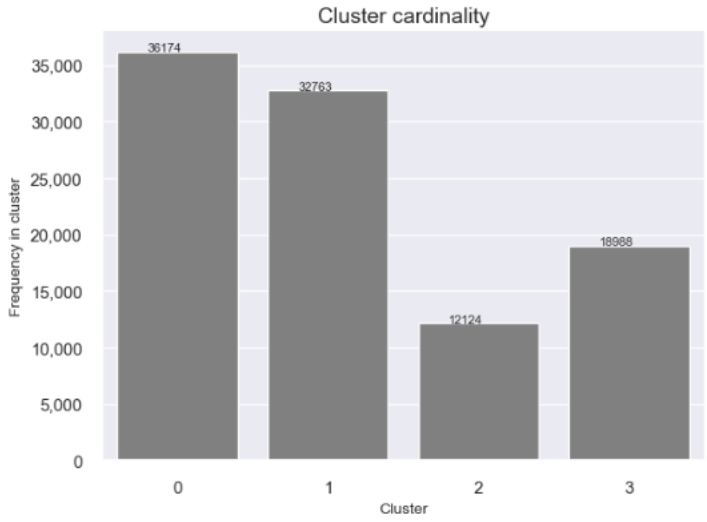
Annex 10 – K-prototypes (elbow method)



Annex 11 – Distance between clusters



Annex 12 - Cluster Distribution



Annex 13 – Cluster results (the darker the blue, the higher the value is per feature)

	0	1	2	3
AverageLeadTime	59.784845	64.718585	41.311283	72.902517
DaysSinceCreation	584.593437	658.473034	614.065985	553.255108
ChannelCode_3.0	0.009952	0.000000	0.195398	0.011639
RoomNights	2.345828	2.387999	2.123062	2.100116
TotalRevenue	369.419638	342.882611	372.761217	360.102591
ChannelCode_2.0	0.085559	0.000000	0.804602	0.066674
ChannelCode_1.0	0.904489	1.000000	0.000000	0.921687
BookingsCheckedIn	0.738624	0.772335	0.802375	0.692121
Nationality_Code_5.0	0.049041	0.055764	0.073161	0.045292
Nationality_Code_2.0	0.102477	0.160211	0.084213	0.187013
age_bins_>=66	0.116271	0.117389	0.075553	0.196545
age_bins_<25	0.056919	0.056405	0.051303	0.085949
Nationality_Code_3.0	0.102145	0.133779	0.227318	0.078207
Nationality_Code_4.0	0.111378	0.092452	0.078192	0.121603
age_bins_25-35	0.196550	0.173977	0.171231	0.176006
SRHighFloor_1.0	0.067258	0.005219	0.029528	0.070413
SRHighFloor_0.0	0.932742	0.994781	0.970472	0.929587
Nationality_Code_6.0	0.057666	0.029637	0.063923	0.048294
TotalSpecialRequests_2.0	0.190551	0.006074	0.008001	0.104066
Nationality_Code_1.0	0.159645	0.176480	0.118360	0.112861
SRTwinBed_0.0	0.994278	1.000000	1.000000	0.161997
SRTwinBed_1.0	0.005722	0.000000	0.000000	0.838003
age_bins_46-55	0.262426	0.289198	0.317964	0.224826
TotalSpecialRequests_3.0	0.014043	0.000061	0.000165	0.006478
age_bins_56-65	0.171256	0.176113	0.146404	0.152886
Nationality_Code_0.0	0.417648	0.351677	0.354833	0.406731
TotalSpecialRequests_1.0	0.795406	0.000000	0.058232	0.889404
SRCrib_1.0	0.017222	0.001343	0.017568	0.017537
SRCrib_0.0	0.982778	0.998657	0.982432	0.982463
SRQuietRoom_0.0	0.871565	0.994231	0.972369	0.809722