NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Customer Segmentation for hotel H from chain C

Ana Teresa Maia, 20201562

Henrique Falcão, 20201519

Maria Benedita Elias, 20230491

Maria Leonor de Gusmão, 20230488

Pedro Carvalho, 20230487

Group H

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

In the highly competitive hospitality industry, understanding the diverse needs and preferences of guests has never been more critical. One of the key strategic priorities to thrive in such business is the ability to offer personalized experiences that resonates with each individual guest. In today's digital age, businesses are capable of gathering large volumes of customer data that through analytical tools and methodologies can be used to better understand and cater customers on a much more granular level.

The aim of this report is to aid Hotel Chain C to step away from the standard hospitality market segmentation that only focuses on the origin of the customer and instead develop a much more in-depth segmentation that takes into consideration more consumer characteristics. Moreover, this analysis will also help the marketing department to create more catered marketing strategies for each detailed customer segment, ensuring the ability to attract new customers while maintaining the loyalty of existing ones.

For the development of this project and in order to achieve the desired outcome, we followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. As a result, our main steps were the following: Business and Data Understanding; Data Preparation; Modelling; Evaluation and Deployment. These ensure a cohesive approach to the requests with a focus on the business needs that, in the end, is understandable to the parties involved with the deployment of the plans.

This Proof of Concept followed two approaches, meaning two clustering algorithms and the one that yielded the best performance according to our evaluation metrics and Hotel Chain C's business needs was chosen. The results obtained were four distinct segments designed based on their unique characteristics. After carefully analyzing each cluster, the appropriate marketing actions were determined with the aim of retaining current customers by addressing their desired benefits and increasing overall service satisfaction. Lastly, a plan for deployment and maintenance of this project is presented to ensure a smooth transition and implementation.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BUSINESS OBJECTIVES

The hospitality industry prioritizes its customers above all, focusing on fulfilling their requirements, enhancing, and solidifying relationships, and elevating their satisfaction levels. Due to its recent expansion, hotel chain C wants to change its customer segmentation approach to improve its knowledge about its current customers and derive catered marketing strategies to them while, at the same time, reach a bigger volume of new clients.

This objective can be divided into two sub-objectives:

(1) Create a new customer segmentation model that considers more detailed characteristics of the customer, such as demographic, behavioral, and financial. This way Hotel Chain C's marketing departments will have a better understanding of their customers and be able to visualize more specific client segments
(2) Develop marketing strategies catered to the customer segments outlined by the clustering model previously created to build more personalized experiences for Hotel Chain C's clients.

2.2. BUSINESS SUCCESS CRITERIA

The business success of this Proof of Concept will consist of various criteria. Each identified segment from the model should be able to be used by the marketing team to tailor and execute specific marketing strategies and the new segmentation accurately captures customer preferences and behaviours, verified by increased engagement, satisfaction, and loyalty from targeted marketing. Moreover, if there is an increase in the acquisition of new customers attributed to the marketing, based on the new segmentation.

2.3. SITUATION ASSESSMENT

The requirements and key problems raised by the new marketing manager, A arose after an expansion from the hotel chain C, which demanded a further investment in marketing and a more thorough description of its different customer segments. To achieve this, our team was given a dataset provided by hotel H in Lisbon which comprises a multitude of geographical, demographic, and behavioural customer characteristics.

Taking this into consideration, building reliable client profiles is a key element to target them more efficiently. To produce a good deliverable, we took time to gather the specific needs/concerns of A. Firstly and most straightforward the main driver we encounter in this proposal is the need to change the clustering direction. Instead of being only origin-oriented, our team understands the need to produce a diversified perspective to tailor each marketing campaign more adequately. Our initial idea is to preserve variables that will encapsulate sales levels but also give hints and reveal patterns in the realms of demographics (with age groups and behaviour of stays, for example), geography, through the region the clients are coming from, but also other relevant information such as the tendency to cancel the booking and distribution system used.

Regarding software interfaces that the team used, we developed our model by coding in Python, using Anaconda and Jupyter Notebooks.

2.4. DETERMINE DATA MINING GOALS

The first stage goes through Understanding of the Business where it covers all the topics of this initial part of the report. It lays the foundation for all the work that will be developed in the rest of the project.

Moving to the Data Understanding phase, our team spent a dedicated amount of time extracting information about what the data means, which inconsistencies must be dealt with and what are the overall state and quality of our database points.

The Data Preparation stage goes hand in hand with the previous one. At this point, we clean the data and transform it in order to have variables and information ready to be used. We deal with inconsistencies, missing values, outliers and create new variables, all to prepare the best set of characteristics to be employed in the clustering algorithm, free of irrelevant and redundant information. At this stage, we also did principal components analysis, a useful tool to go over the multidimensionality curse.

Then, the Modelling phase is where we used an algorithm for the clustering activity. With it, we group our objects of analysis based on the highest level of similarity between them constructing distinct groups, which will be the optimal goal for the development of tailored marketing campaigns.

Lastly, we must evaluate our model and our results, because a good solution also goes through a round of self-assessment until we believe the solution, we present is the most adequate one, considering all the criteria and the current situation. After that, we built the deployment strategy by writing this report and identifying maintenance plans to ensure the results keep being employed and updated. This section aims at connecting the object of analysis with the solutions developed to them.

Due to the time constraint, the data mining success criteria for this proof of concept will consist of creating the model that not only yields the best values in the performance metrics that will be used (Silhouette Score, the Davies-Bouldin Index, and the Calinski-Harabasz Index) but also the model that creates the best clusters for our business analysis.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

The data understanding phase is crucial to proceed with the development of the other steps when building a CRISP-DM model. There is a need to study the data to understand the representation of each variable presented, and in what way it can be useful when performing customer segmentation. The unit of analysis is the customers, not the bookings made by them, therefore the treatment applied is based essentially on the customers' characterization.

The data set provides a considerable number of records to be analyzed (111733 observations) with information regarding 29 columns (variables). There is a variety of data types ranging from numerical, categorical, and binary data. A brief inspection both on data types and cardinality was made to guarantee that the data was correctly identified.

By looking at the summary statistics it was possible to identify that some variables have a high dispersion of values, and even present some inconsistencies. For instance, some records in the *Age* variable were negative and others had unreasonable values above 100, which accentuated the high dispersion in values of this feature. Moreover, our dataset was also characterized by the presence of missing values, extreme values, duplicates (visualized through "NameHash", "DocIDHash" and "Nationality"), and an extensive variety of categories (the case of "Nationality" variable).

To finish the data understanding phase our team decided to compute some visualizations to check the current state of the data and examine possible relationships between variables. Variables such as "Age" and "AverageLeadTime" have negative values which again stresses the incoherences in our data. The binary variables regarding the customer's requirements show that customers usually have more specific preferences regarding the bed type required *(Check figure 1 in the annex).*

Lastly, some variables may be redundant for the segmentation and characterization of Hotel H customers i.e., some of them would represent the same information in the analysis. Therefore, a careful analysis of these topics is essential in the data exploration phase while also having some visualization for support, such as the correlation matrix.

For a more detailed explanation regarding the exploration and understanding of the dataset, we advise you to check the considerations and visualizations presented in the Jupyter notebook attached.

3.2. DATA PREPARATION

In this phase of the project, we put effort on making the dataset ready for the modelling phase and future customer segmentation.

Regarding the incorrect values that we identified in the variable "Age" our team decided to change the negative ones to their module and the ones more than 100 were all reset to 100. Then, since people under 18 cannot make hotel reservations we decided to not consider these people in our analysis, dropping them from the dataset.

Additionally, the dataset presented duplicated lines with the same conjugation of "Nationality", "DocIDHash" and "NameHash", meaning that there are people that appear more than 1 time in the data. Since we don't want to segment the same people twice, we decided to aggregate these lines, summing the numeric variables and keeping the most recent information of the other variables. All the assumptions regarding this step are present in the provided notebook.

Since "Nationality" has too many categories, customer segmentation becomes a challenge this high level of granularity. As a result, the records were aggregated considering only if they belonged to the European continent or not, and a new variable was created named "Europe".

On other variables such as "*DistributionChannel" and "MarketSegment"* we joined the categories with less than 5% of representation in the dataset into a new category named "Others" since they are not relevant enough to be presented on their own for the business.

Next, we move on to focus on the creation of new variables that we believe create value to understand the motivations regarding groups of people/clients in the Hospitality business. Thus, the new variables considered are the following: (1) "AlreadyClient"(2) "BookingsMadeNotRealised" (3)

“Total Revenue” (4) “TotalSR”, (5) “CancelationRate” (Check figure 2 in the annex for a detailed explanation).

In terms of outlier handling, as we are looking for a good customer segmentation, by identifying the presence of extreme values in some variables, we mitigate the problem of small observations influencing our results later on. The approach that our team proceeded with was to remove these few observations that represented unusual and unlikely behaviours in the context of our analysis. The removal was made through a manual limitation technique that can be found in *reference[1]*.

Furthermore, we scaled the dataset’s numeric variables using a common normalization method called Min-Max scaler that limits the variables values in [0,1] interval, thus the calculations of distance points for modelling techniques will not be influenced by the different scales in our data. To learn more about this method, *check the reference [2]*.

Subsequently, we were able to perform KNN imputation methods to treat missing values of metric features (they were only present in the “Age” variable). For more information regarding this method, *check the reference [3]*.

To further facilitate cluster segmentation analysis the “Age” variable was converted into bins, this discrimination can be seen in more detail in the notebook. *(Take a look at figure 3 and reference [4] to get a better overview of this topic)*.

Through the visualizations we saw before regarding the customer's requirements, the 2 variables related to bed type requirements are the ones with the highest relevance. With this in mind, the presence of the variables "SRKingSizeBed" and "SRTwinBed" was maintained and for the other requirements information we considered grouping them into a numerical single variable, “TotalSR” therefore, these individual variables were removed.

Looking at the overall representation of the variables we noticed that there is the need to evaluate the relevance and redundancy of each one. For that reason, we decided to drop “DocIDHash”, “NameHash", “Market Segment”, “Total Revenue”, “Lodging Revenue”, “PersonNights”, “BookingsCanceled”, “BookingsNoShowed”. Additional explanations regarding these decisions are available in the notebook.

To increase the interpretability of categorical features and make them computationally possible in our models, we applied One-hot encoding, turning all categories into binary 1s and 0s. To research more about this method *check the reference [5].*

As a final consideration, we dropped about 7% of the records presented in our data set. In our point of view and considering the current state of the data, this was the best approach to reach a coherent dataset and consequently facilitate the customer segmentation analysis.

3.3. MODELLING

In this phase of the project, we are going to explore 2 different widely known clustering techniques. At this step we are going to be able to understand patterns by grouping people with similar behaviors and characteristics together to achieve the goal of this project: the customer segmentation and respective marketing strategies for each cluster.

K-means

In our dataset, a significant number of variables were present, potentially leading to poor clustering solutions. This is because points (representing individuals) in a high-dimensional space tend to be more distant from each other, making it more difficult to form meaningful groups.

To address this, we applied Principal Component Analysis (PCA) to our dataset as a preliminary step. PCA is a dimensionality reduction technique used to transform high-dimensional datasets into lower-dimensional ones while retaining most of the variance in the data. Since our dataset was composed of metric and non-metric features, we had to encode the former ones to convert them into numerical representations to perform the PCA (which operates on numerical data).

We decided to retain 10 principal components using the elbow method and personal expertise, which represented about 92% of the variability of the data.

Following this step, we were ready to implement K-means, which is widely used to group data points based on their distance. Since this technique requires the specification of the number of clusters, we used two methods to help us decide. Firstly, we looked at the Inertia plot, which shows us how close the data points are from its centroid (the center of the cluster) for the different cluster solutions with different numbers of clusters. We want a low value of inertia, meaning that the points inside the clusters are similar to each other (are close to their centroid). Secondly, we looked at the silhouette score wanting a high value (1 being the maximum), meaning that the data points in one cluster are far away from the neighboring clusters. Based on these measures, 6 clusters were chosen. *For a more detailed explanation regarding K-Means application follow reference [6]*

After the clusters were formed, a careful analysis was made of each of them through bar plots and looking at their distributions, to better understand their characteristics and behaviors.

K-Prototypes

Finally, we performed K-prototypes which is a technique used for mixed data type datasets that uses Euclidian distance to measure the distance between numerical features and the number of matching categories to compute the distance between categorical variables. In this approach, we used the dataset with the categorical features in their natural format and the metric ones scaled. *For a more detailed explanation ~~regarding~~of this method check reference [7].*

To determine the optimal number of clusters, we utilized the Inertia plot and applied the elbow method together with our domain expertise, choosing 4 clusters.

To comprehensively analyze the cluster results, we examined the contribution of each categorical variable category to every cluster, alongside the mean values of the numerical variables.

3.4 EVALUATION

The model evaluation section aims at allowing a reasoning and understanding of how proper the solution is. Our team, as described in the last section, did two separate analyses, one considering K-means with PCA and one considering the algorithm K-Prototypes. The deliverable can only be one algorithm and consequently, our decision was made based on the model that had the best balance between performance metrics i.e., how well the clusters were formed, and our judgment of how well the division is. K-Means did not reveal much discrepancy since in one case age was better distinguishable and in K-Prototypes, some metric features were properly presented. However, regarding the measures used, we concluded that K-prototypes was the best alternative to follow. We used three measures to come up with our evaluation: the silhouette score, the Davies-Bouldin Index, and the Calinski-Harabasz Index, where K-prototypes was performing the best by far. *More details on these metrics can be found in the notebook and references [8] [9].*

From now onwards our final solution is the one provided by K-Prototypes and as a result, four different customer clusters were found. The following tables show the results obtained for each variable used, considering the values in the upper part to be in percentage.

labels_2	Age 0-20	Age 21-30	Age 31-40	Age 41-50	Age 51-60	Age 61-70	Age 71+	Europe 0	Europe 1	DC Direct	DC Other	DC Travel Agent/ Operator	SRKingSizeBed 0	SRKingSizeBed 1	SRTwin Bed 0	SRTwin Bed 1	Already Client 0	Already Client 1
0	0.01	0.10	0.19	0.28	0.20	0.14	0.09	0.13	0.87	0.16	0.05	0.79	1.00	0.00	0.91	0.09	1.00	0.00
1	0.11	0.14	0.18	0.15	0.21	0.12	0.08	0.16	0.84	0.21	0.04	0.75	1.00	0.00	0.77	0.23	0.99	0.01
2	0.01	0.10	0.20	0.17	0.29	0.14	0.09	0.16	0.84	0.06	0.01	0.93	0.23	0.77	0.81	0.19	0.99	0.01
3	0.10	0.12	0.17	0.27	0.18	0.11	0.06	0.23	0.77	0.11	0.01	0.88	0.13	0.87	0.89	0.11	0.99	0.01

Table 1 – Cluster Results: Non-Metric Variables

labels_2	Days Since Creation	Average Lead Time	Room Nights	Bookings CheckIn	Bookings MadeNot Realised	Total Revenue	TotalSR	Cancelation Rate
0	1018.47	80.86	2.95	1.00	0.00	423.12	0.13	0.00
1	342.53	44.91	1.76	0.60	0.00	290.72	0.29	0.00
2	991.91	104.25	3.22	1.04	0.00	493.17	1.24	0.00
3	337.00	43.73	1.80	0.57	0.00	302.49	1.28	0.00

Table 2 – Cluster Results: Metric Variables

Starting with **Cluster 0**, it has a cardinality of 20977 clients. This cluster was named *Family Vacation Cluster* as it represents a mix of ages (from 31 to 71+). Additionally, most of the customers are from Europe. They make no special requests and prefer to book in advance (average of 80 days). They book multiple rooms, which is expected of families, and tend to make multiple bookings. This cluster provides a higher amount of revenue (on average 423.12 euros) in comparison to others.

Moving to **Cluster 1**, this cluster is the one of highest cardinality (36843 clients) and the one with the highest potential. In terms of analysis, this cluster is the *Corporate Cluster*. It is characterized by having users mostly in the young working generation below their 40's, once again, mostly European. This is the cluster with the highest value in direct bookings made, this being a feature of the corporate nomenclature. These clients are practical, as expected of a corporate client. They never ask for a King Size Bed and overall, never made one special request. They are quick: they book close to the arrival date (generally one month) and do not come in big groups to the hotel (average of 1.76 Room/Nights).

Another distinction factor is that they tend to be one-time-booking clients, probably because they only come for meetings. The potential this cluster has is related to the fact that it is the one with the lower revenue amount, giving chain C the possibility to target them in multiple ways.

Cluster 2 is our *Silver Age European Vacation Cluster*. This is the lowest cardinality cluster but the most profitable one (average revenue amount of 493.17 euros). These are people of more advanced ages that ask for special requests, as almost everyone asks for a King Size Bed. They are mainly from Europe, booked their stay through an agent and often come in groups since they book the most rooms per night (3.22), which signals the vacation purpose. They like to plan as they have the highest time since the booking was made (104 days). These are loyal customers, which is an advantage to the hotel, they are the ones that tend to book multiple times and spend the most money, almost double as the previous cluster.

Finally, **Cluster 3**, with a cardinality of 28910. This is the *Mixed cluster* for us as it is a mix of some characteristics. However, starting with the most different one, this cluster has the most volume coming from outside Europe. These visitors are young, the youngest cluster we have (from 18 to 40), and as expected from young people they are the ones that book closest to the arrival date (43 days). They do not come in big groups, as the room nights variable is low. A curious characteristic is that these people might come from wealthier countries where the younger generations have a higher purchasing power as they tend to make special requests and provide a good amount of revenue.

Based on the description and interpretation made of our clusters we can see that they are clearly differentiated, and each one has their unique characteristics that can be taken advantage of for more effective marketing. Overall, we believe that our model is performing well and that moving forward with this option was a good choice. After confirming the proper distribution we have among clusters, we can move to the Marketing strategies in the next section.

4. RESULTS EVALUATION

As in any business, hotel chains need to raise the number of customers, as well as increase customer retention and engagement. Therefore, the following tailored marketing strategies for the identifies clusters are suggested:

Beginning with **Cluster 0**: *Family Vacation Cluster*, we propose to build on the family-oriented nature of this cluster by offering exclusive family vacation packages. These could include discounted fees for multiple room bookings and amenities for children, such as a kid's club with diverse activities. To encourage early bookings, offer exclusive perks for reservations made ahead of time, it is important to emphasize the value of planning ahead to secure preferred accommodation.

Next, in **Cluster 1**: *Corporate Cluster*, our team suggests specialized corporate packages fitted to the needs of business travelers, such as high-speed Wi-Fi, access to meeting rooms, and easy booking options for corporate events and conferences. Appeal to the practical nature of this cluster by offering last-minute deals and discounts for impulsive bookings. To change the norm of being one-time -bookers the chain could create a points system where after completing X-stays the hotel would offer discounts to these clients.

Moving to **Cluster 2**: *Silver Age European Vacation Cluster*, we propose to cater to the luxury preferences of this cluster by offering upscale experiences and amenities, for instance exclusive services such as personalized concierge assistance, gourmet dining experiences, and spa packages. Create faithfulness among this cluster by implementing tailored loyalty programs that recognize and reward repeated guests. Offer perks such as discounts on future stays, VIP access to special events, and personalized welcome amenities.

Finally, **Cluster 3**: *Mixed Cluster,* since our team recognized the diverse demographics within this cluster it would be important to offer a range of global experiences and services, highlighting the hotel's multicultural offerings and multilingual staff. Our suggestion would be to use digital marketing to reach this international cluster, and to offer incentives, such as complimentary upgrades. An example of a possible upgrade would be providing them with a direct shuttle to the airport or to the city and after they booked a room associated with a certain fee, we would double the offer. This way we would be aligned with their higher purchasing power whilst providing them with a more tranquilized welcome experience, in the case they are coming from the airport.

The hospitality market is increasingly competitive and that is the reason why it is of most importance to align marketing strategies with the needs and behaviors of each cluster to reach full customer satisfaction, revenue, and loyalty.

5. DEPLOYMENT AND MAINTENANCE PLANS

For the integration and resilience of this project the resulting deployment and maintenance plan, seen in *Figure 1*, was developed. By following this structured approach, Hotel Chain H can maximize the benefits of this customer segmentation project, leading to enhanced customer experiences and future business growth.

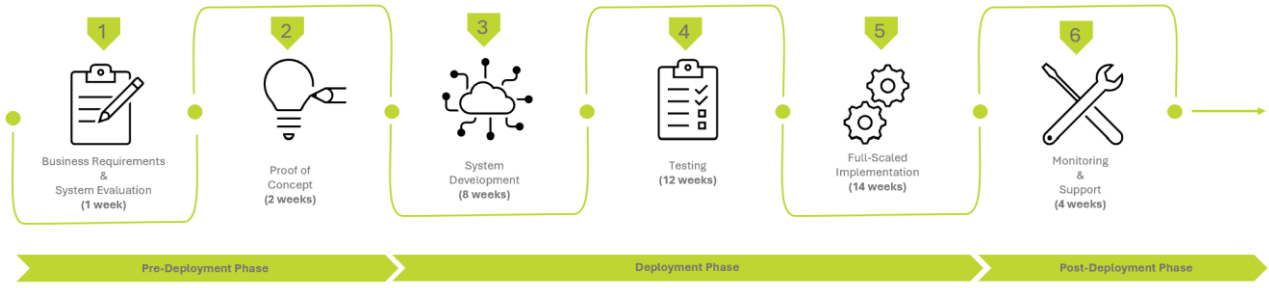

Figure 4 – Deployment and Maintenance Plan

1. **Business Requirements & System Evaluation** – Clearly outline and document the business goals and needs of the customer segmentation model as well as assess current IT infrastructures and data availability to identify any limitations or enhancements needed for the project. This stage involves stakeholders from relevant departments such as IT specialists, marketing personnel and customer service representatives in order to gather insights and ensure alignment with business needs. Such procedure entails a series of meetings between said stakeholders and the business analysts and data scientists that will develop and evaluate the customer segmentation model.

2. **Proof of Concept** – Our model, together with this report, presents an initial prototype using the selected customer data of Hotel H to demonstrate the segmentation capability. It shows the model's ability to segment customers and presents how these segments can be used to improve marketing strategies, customer service, and overall business outcomes. This prototype will be presented to a variety of stakeholders to collect feedback in order to assess business value and technical viability.
3. **System Development** – As Hotel Chain C has increased its number of hotels over time, it is proposed the use of the Azure Data Factory tool from Microsoft to handle the large volumes of data in regard to integration, transformation, and orchestration of data to create meaningful segments and also aid in developing accessible insights for actionable business use. Such a service consists of a robust, flexible, and scalable cloud data infrastructure that can better the data processing and deployment of our customer segmentation model. Our team will work alongside the IT department to create the necessary code to integrate our customer segmentation model with Azure Data Factory as well as transfer all the important customer data from the C hotel chain to this system.
4. **Testing** – Marked by a phased roll-out implementation to minimize disruption and allow for adjustments where first, a pilot program of customer segmentation in Hotel H is planned and executed. In this stage, the focus will be to put our improved customer segmentation model and derived marketing strategies into practice and gather feedback to gain insights on possible necessary adjustments. The evaluation of the pilot program outcomes will comprise the assessment of the model's performance against the initial business objective, through user acceptance testing for instance, and feedback from all stakeholders involved in the pilot (IT staff, marketing teams, customer service representatives, …), and customers themselves. Unit testing for new software to ensure all integrated components work together well is also crucial.
5. **Full-Scaled Implementation** – Upon successful pilot testing and subsequent improvement on the required fields, proceed with an extended integration of the model to more hotels in the chain, carefully monitoring performance and feedback at each stage. There is a need to ensure that adequate resources are allocated for the roll-out, namely IT, marketing, customer services, and data scientists personnel as well as additional hardware and software required. The integration of the model will be phased as it is implemented in additional locations or departments in stages, allowing for adjustments and troubleshooting. Additionally, comprehensive training for all users/collaborators on the updated model and processes will be planned.
6. **Monitoring and Support** – To ensure that our customer segmentation model continues to meet business objectives and operates efficiently a set of metrics needs to be established and tracked over time through a maintenance plan. Such a plan should reference measures regarding the accuracy of the model's segmentation, (metrics such as silhouette scores or indicators of customer behavior changes within each segment should be used), the tracking of key performance indicators relevant to the marketing objectives of the segmentation and establishment of feedback loops from customers and stakeholders. Additionally, tools and processes such as Azure Monitor provide the possibility for consistent monitoring of the data processing and model execution infrastructure and effectiveness. For support, the allocation of resources to regular training sessions for staff, through online materials for instance, ensures user competency and adoption.

6. CONCLUSIONS

To conclude our research, we must state that the results showed the need for a discrimination between types of customers that exist in the hotel chain. Forming heterogeneous clusters of customers, we saw that each of these groups need a particular treatment, it is important to adapt a specific marketing strategy capable of satisfying the needs of each group.

It is also important to highlight that the customer segmentation and the marketing strategies pointed to that need to be continuously updated in order to follow the customer preferences along the time.

We totally encourage Hotel Chain C to use this proof of concept as a guide to reach new customers, captivate the existing ones and as a positive consequence generate more profit.

6.1.1. Considerations For Model Improvement

Throughout the development of this Proof of Concept, we reached some limitations that should be considered in improved versions of this project.

More clustering algorithms should be explored, such as self-organizing maps and more algorithms that can work both with numerical and non-numerical data.

Moreover, the handling of outliers in this project was highly constrained due to the incapability of our computers to run algorithms such as DBSCAN to use in outlier detection. Thus, in future versions, a more thorough investigation into our outlier analysis is needed.

Other encoded methods such as the polynomial could be analysed as well to better fit our data and clustering needs.

Finally, more research and collecting of external and internal data can help to gather more information about consumers and the industry which in return can be crucial to derive more informative segments.

7. REFERENCES

[1]Detecção e tratamento de outliers: um guia para iniciantes. (n.d.). ICHI.PRO. Retrieved March 12, 2024, from https://ichi.pro/pt/deteccao-e-tratamento-de-outliers-um-guia-para-iniciantes-215826140075860

[2]scikit-learn. (2019). sklearn.preprocessing.MinMaxScaler — scikit-learn 0.22.1 documentation. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[3] Brownlee, J. (2020, June 23). kNN Imputation for Missing Values in Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

[4]How to bin age data into categories - issues with setting highs and lows. (n.d.). Stack Overflow. Retrieved March 12, 2024, from https://stackoverflow.com/questions/73655852/how-to-bin-age-data-into-categories-issues-with-setting-highs-and-lows

[5] Zach. (2021, September 28). How to Perform One-Hot Encoding in Python. Statology. https://www.statology.org/one-hot-encoding-in-python/

[6] Garbade, M. (2018, September 12). Understanding K-means Clustering in Machine Learning. Towards Data Science; Towards Data Science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[7] Ruberts, A. (2020, May 16). K-Prototypes - Customer Clustering with Mixed Data Types. Well Enough. https://antonsruberts.github.io/kproto-audience/

[8] sklearn.metrics.davies_bouldin_score — scikit-learn 0.22.2 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[9] sklearn.metrics.calinski_harabasz_score — scikit-learn 0.24.1 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

[10] Data Mining Labs from git hub repository. https://github.com/fpontejos/Data-Mining-23-24/tree/main/notebooks

8. APPENDIX

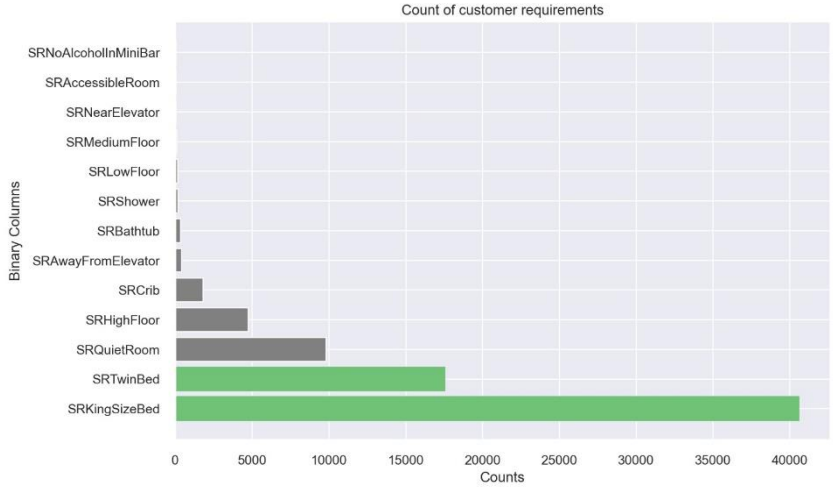

Figure 1 – Customers' requirements (Binary variables)

Variable name	Description
"AlreadyClient"	If the client had made more than 2 check-Ins we consider the person already as a client (1) if the client just made 1 or 0 bookings then new client (0) (Binary variable)
"BookingsMadeNotRealised"	Sum of "Canceled" and "Noshowed" bookings
"Total Revenue"	Sum of "Other" and "Lodging" revenues
"TotalSR"	Sum of the total requests made by the person excluding the ones related with bed types
"CancelationRate"	BookingsCanceld/Total Bookings rate

Figure 2 – New variables metadata

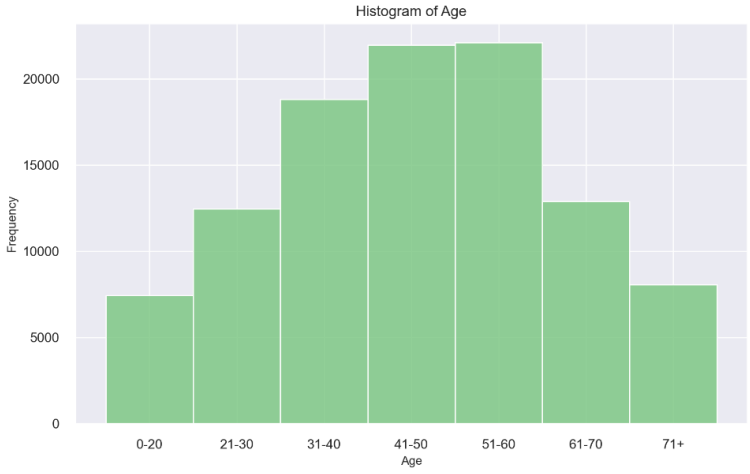

Figure 3 – Bins discrimination regarding "Age" variable