MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel H Customer Segmentation

Duarte Mendes, number: 20230494

Dzmitry Nisht, number: 20230776

Inês Silva, number: r20201580

José Marçal, number: r20201581

Pedro Ricardo Sousa, number: r20201611

Group E

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

Hotel H, a Chain C hotel located in Lisbon, has a new manager, A, who acknowledges the importance of customer satisfaction and loyalty in the hospitality sector, and it is also aware that these goals can only be achieved by ensuring proper profiling of customer segments, using an adequate customer segmentation process . This initiative comes from verifying that the current customer segmentation in place has a great deal of limitations, since it is only profiling customers by sales origin. Looking to develop tailor-made marketing approaches and solutions for their customers' needs, Hotel H's main target is to perform an effective customer segmentation process that, besides sales origin, can also be based on geography, demographics, and behavioural characteristics of customers, and, in this way, capturing the diversity of their customer base. The ultimate goals of this projects are, then, ensuring customer satisfaction and loyalty, and revenue growth.

Chain C and Hotel H want to be able to indulge in the different customers' requests, needs, and preferences so that it can take on new, diverse and strategic opportunities related to marketing efforts, price definition strategy and concept and product creation, that will allow it to expand and become stronger against their peer competitors, which are well-known to be highly competitive.

Consequently, the CRISP-DM process was considered to be the most complete and accurate approach for achieving the desired results in this customer segmentation process. It involves firstly an extensive understanding of Hotel H's business, by analysing its background, defining clear business targets and success criteria for the project, assessing the current situation (resources and available data, risks/contingencies, and cost/benefits for this project) and determining the data mining goals.

After the business needs and required outcomes were defined, the methodology applied to the data encompassed the processes of data understanding, preparation, modelling, and evaluation. Through these phases, the available data will be pre-processed, and PCA (Principal components analysis) shall be used to assist in dimensionality reduction, which later will be applied to a K-Means model.

The results evaluation will allow us to analyse whether hotel H's customers were well grouped in terms of the resulting clusters, and these will be changed until a reasonable, as well as meaningful, final result is accomplished. Afterwards, once the clusters are well defined, these will be the ground basis to start developing suggestive marketing plans for each segment.

In conclusion, this project provides customer and business data-driven insights, as well as essential tools that can elevate Hotel H's marketing efforts and business strategy definition, by aiming at a diverse and insightful customer base that shall be explored in terms of their most diverse characteristics. Such process has the power to open the door for major opportunities that will allow hotel H to grow in the hospitality industry, as well as to elevate customer experience to higher levels.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Independently of the industry, all companies depend on customer satisfaction, and hotel H is no exception: thus, to be able to do so, it is necessary to understand their customers and their habits. In 2018, Chain C created a marketing department and hired a new marketing manager in 2018, as result of their expansion. The new marketing manager realised that the current customer segmentation approach of Chain C was not appropriate, as it only reflected only the sales origin of customers, and that it should also account for other factors, such as geographic (Nationality), demographic (Age) and behavioural characteristics (Number of Stays). Therefore, Chain C and Hotel H are urgently needing to improve their customer segmentation process to attract new customers, while continuing to captivate the current customers.

2.2. BUSINESS OBJECTIVES

The main purpose of this project is to understand and analyse the different needs and habits of hotel H's customers, which can be the foundation of an effective approach when applying marketing strategies. These objectives can be achieved by analysing and grouping clients according to similar characteristics, namely, geographically, demographically, and behavioural. The end goals include the increase of customer satisfaction and loyalty, but also enhanced and effective marketing strategies that can meet both current and target customer's desires and needs. Therefore, the marketing efforts should be able to retain its current customers and attract new ones. Financial and organizational objectives are also included, in terms of increase of revenues and overall operational efficiency.

2.3. BUSINESS SUCCESS CRITERIA

There are many ways to evaluate the success rate of the schemes provided. One of those ways relates to a customer segmentation process that can ensure proper definition of the different customer segments, while boosting significantly hotel H's marketing efforts. Moreover, other fundamental success criteria that will evaluate if the project respects the wishes of Chain C and Hotel H's board are the following: whether or not client satisfaction increased, which leads to the second metric - if customers come back to the hotel (customer loyalty). However, that would not be enough: the key outputs of this customer segmentation process are expected to shape hotel H's business and marketing strategies, which can ultimately lead to significant increase in revenue and occupancy rate numbers. Moreover, checking if the marketing campaigns were effective in capturing new clientele, being this measured through feedback made available by the customers, is also considered as one of the business success criteria.

2.4. SITUATION ASSESSMENT

The hospitality industry englobes people from all around the world, which means that it is vital that Chain C and Hotel H's marketing team keeps in mind that they are trying to reach people with a wide diversity of behaviours, characteristics, and needs. Nevertheless, the current marketing strategy of Hotel H is simply structured by sales origin, meaning they are losing critical data.

Secondly, the hospitality sector has suffered enormous alterations over the years, and, as so, business adaptation is a key factor for all the players involved in this market. Consequently, the adoption of technologies and understanding their impact on the business are two factors that should be kept in mind. Therefore, training Hotel H's employees to learn how to use those technologies, and also to be part of the Hotel H's marketing efforts, can be considered a step forward towards success.

Besides these factors, it is necessary to keep in mind that hospitality is one of the sectors with more competition. In this way, building unique brand and product concepts based on this project, with the power to enhance customer experience and attract new customers, can lead Chain C and Hotel H to the top players of the sector. Subsequently, it is a must that Hotel H should aim at effective and well-defined customer segments, that take more of the customer's characteristics in consideration.

The data resources that were made available to reach this goal consist of a dataset with 111733 entries and 28 columns, containing the reservations made until 2018, the year of dataset extraction. Such database should be the base of our work, as well as the beginning point for finding patterns within the customer data and construct appropriate customer segments. The data preparation was made using the most popular Python packages.

From a financial perspective, developing an extensive customer segmentation process can be expensive, both involving direct and indirect costs. However, it is thought that the benefits will outweigh the disadvantages, as an adequate customer segmentation can ultimately lead to increased revenues, reduction in operational costs, or higher levels of customer satisfaction and loyalty.

2.5. Determine Data Mining goals

The data mining goals consist of the technical objectives for this project, and defining customer segmentation through a thorough data analysis can be considered the general one. To reach it, first, it is necessary to carefully handle and prepare the data that was made available, by using a variety of techniques and Python packages.

Next, there is the selection of a clustering method that can appropriately conduct to a solution with a reasonable number of clusters in terms of business and marketing approaches, and also split customers into diverse and distinct customer segments. Other technical data mining goals are related to high distortion and silhouette scores, and an high correlation between cardinality and magnitude of clusters. For this project, the clustering method chosen was K-Means. The modelling step in this project requires dimensionality reduction, which was performed using the PCA (Principal Component Analysis) technique.

3. METHODOLOGY

The methodology should be a brief explanation of the main steps that compose business and data understanding, data treatment, model development, as well as results evaluation and deployment and maintenance plans. As previously defined, CRISP-DM methodology will be followed throughout the project, and each of its steps (Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment and Maintenance Plans) will be extensively explained, according to the work developed throughout the project.

3.1. Data Understanding

Data understanding, the next phase of this project, is crucial, and it depends on the data scientists' perspectives and approaches. With that in mind, when looking at the data each variable was analysed for potential problems and how its usefulness could be maximised, while having a closer perspective on the customer's data, and the relevant features.

Subsequently, the first problem that was encountered was the existence of duplicated rows, followed by a substantial amount of 0's included in the column *"BookingsCheckedIn"*, showing that clients who never used Hotel H's facilities were being considered.

Next, there was *Age*, which besides having missing values, also included unreasonable values for a customer of Hotel H (negative ages, considering underaged people and considering people with over 100 years). Another variable containing unreasonable values was *"AverageLeadTime",* which had negative values (Figure 1). Furthermore, "*DocIDHash*", which should be composed of unique values, had also missing values, meaning that filling them was probably going to be a problem. However, that was not the only problem with this variable: there were repeated "*DocIDHash*" values, which meant that the same person had more than one profile in the database (Figure 2).

Additionally, the data included customers of many nationalities but, of those, 5 main nationalities represented the majority, having the remaining really small percentages. In another perspective, looking at the variable "*DistributionChannel*", there was a clear distinction that the majority of people had the category "Travel Agent/Operator" (Figure 3).

Furthermore, a considerable number of outliers were detected, which were negatively influencing the data, since these observations were making the data skewed (Figure 4). In addition, binary variables ("*SRHighFloor*", "*SRLowFloor*", "*SRAccessibleRoom*", "*SRMediumFloor*", "*SRBathtub*", "*SRShower*", "*SRCrib*", "*SRNearElevator*", "*SRAwayFromElevator*", "*SRNoAlcoholInMiniBar*" and "*SRQuietRoom*") presented unreasonable distributions of 0's and 1's (Figure 5).

3.2. Data Preparation

With a deeper understanding of the data, it became time to prepare the data to be used for the clustering process. The first step was to remove any duplicate rows that were present, of which there were 111 rows. Next, the coherency of the data was checked by removing any clients who had never checked in, whose ages had been stored as negative values (which were considered as errors), lower than 16 years old (as it makes little sense to cater to children who are likely not the ones booking the stays) or higher than 95 years of age (there were no values of age between 95 and 110, therefore it was considered those above 95 to be typos). Other rows were also changed, for cases where

"*AverageLeadTime*" = -1, we assumed as a default value and changed it to be null. If "*PersonNights*" or "*RoomNights*" were inferior to the number of check-ins then those rows would be deleted. Any row where "*RoomNights*" was higher than "*PersonNights*" was also removed.

Finally, for "*DocIDHash"*, the value that repeats the most was established as default for missing values and set those values to be null, and subsequently removed any observation where "*DocIDHash*" was missing. To have unique clients per observation in the dataset, and since some rows had the same "*DocIDHash*", we considered a unique client by the combination of "*DocIDHash*", "*NameHash*", and "Nationality" to provide a more accurate list of clients. We grouped the dataset by these columns, joining information by the last available information, obtained with "*DaysSinceCreation*" (in case of age, or distribution channel), by sum (in case of revenues or rooms per night), or by max (in cases of special request).

After having a well-organized dataset with unique customers, new variables were created and modified to assist in analysing the data:

Table 1 – Creation of New Variables

New Variables	**Composition**
Special Request	Aggregation of all special requests with less variation in results
Requests on Bed Size	Two variables regarding bed size requests (King and Twin)
Total Revenue	The sum of Lodging Revenues and Other Revenues
Invalid Bookings	Aggregation of cancelled bookings and bookings where the client never showed up
Top Nationality	A set of columns categorizing whether a client belongs to one of the top 5 nationalities or falls into another category."
Distribution Channel	A set of columns categorizing whether a client comes from Travel Agencies or Operators, or Direct Booking

With all the new variables created, the missing values in "*Age*" and "*AverageLeadTime*" were filled with the use of logistic regression, since all others were treated during the previous steps. Afterwards, it was time to decide how to handle the numerical data. As stated in point 3.1, the data has a lot of outliers and a very skewed distribution. Since it was believed that these outliers had a lot of information, it was decided not to remove them. For "*TotalRevenue*", "*RoomNights*", "*DaysSinceCreation*" (which was renamed at this stage to "*CustomerTenure*") and "*AverageLeadTime"*, these were binned according to the previously defined criteria and interpretation, while maintaining a meaningful group of observations in each group. For *"BookingsCheckedIn"* (renamed to

"*OneTimeCustomers*"), a binary variable was created to determine if the customer had booked the hotel more than one time (Figure 6). For "Age" we divided into easily interpretable groups. "*PersonNights*" was removed due to the high correlation to other variables, especially "*RoomNights*" (Figure 7). Finally, after removing all variables irrelevant to the analysis and variables used during preprocessing, and renaming the columns, the data was normalized with MinMaxScaler to guarantee the magnitude of different variables does not skew the data.

A note about the removal of variables, we ended up also removing the "*MarketSegmentation*" since we want it to create a new one, and the previous one, besides being highly focused on the origin of sales, most customers are one unique class named "*Other*" (Figure 5).

3.3. Modelling

With our data prepared, it was time to start the modelling process. The first step is to perform a principal component analysis for dimensionality reduction purposes, where 18 principal components were kept, representing 94% of the variance of our data.

With the PCA finished, we moved on to the k-means clustering. After experimenting, we concluded that k-means++, and its ability to establish the initial centroid seeds according to the probability of being distant from other clusters, gave us the best results, and therefore that was the method to be employed in our project. Now, it was time to find the optimal number of clusters by analyzing the elbow graph and silhouette score. The elbow graph suggested more than 8 clusters (Figure 8), but that seems unreasonable to create a market plan. And, while there was some difference in the silhouette scores with relatively high scores for 3/5 clusters, we simply assumed them to not be significant enough (Figure 9). It was only after comparing the cluster results that we concluded that 5 clusters were the optimal choice since they provided the best interpretable results.

3.4. Evaluation

The five clusters that were formed presented some variance when it came to their cardinality, with the biggest cluster representing around 20000 members, and the smallest cluster containing just under 10000 members. However, this difference in cardinality (Figure 10) is very well explained by the difference in magnitude (Figure 11) of each cluster, as seen in our cardinality vs magnitude graph (Figure 12), which shows a linear correlation between the parameters, meaning that no major anomalies seem to exist in the clusters. Finally, we analyzed the 2D intercluster distance map, where we observed that, besides clusters 0 and 2, the clusters are quite far apart from each other, and that the points of cluster 4, are very close to each other, relative to other clusters (Figure 13). In tandem, these results give us confidence in the validity and robustness of our clusters and subsequent analysis.

3.5. Key Performance Indicators & Analysis of Customer Base

In a competitive area such as the hotel industry, the analysis of Key Performance Indicators (KPIs) has a fundamental role in establishing metrics that can evaluate success, identify certain areas with potential for improvement and optimization of processes, and also to support data-driven decisions to ensure a sustainable and thriving future for Hotel H.

As so, some KPIs that could help us to position Hotel H in terms of overall performance and organizational success were calculated: average revenue per client, average revenue per year and average revenue per booking checked-in (Figure 14). Additionally, using a complete dataset with the variables from the feature creation step, it was performed an analysis of Hotel H's customer base, to define an accurate view of the main descriptive characteristics of Hotel H's clients.

From the above analysis, some important insights were extracted. The number of checked-in bookings by the top 5 nationalities was also analysed, with France leading in this ranking, but followed closely by Germany and Portugal customers (Figure 15). In what concerns the distribution of customers by age, hotel H's clients belong mostly to the 50-59 age group, with age groups 30-39, 40-49 and >=60 showing also a high proportion of clients (Figure 16). In terms of cancelled bookings and no-shows, Portugal is shown distinctly as the customer's country with the most booking cancellations and no-shows (Figure 17). A total revenue distribution by the top 5 nationalities was performed, showing that customers from France, Germany and Great Britain contribute the most to the total revenue values of Hotel H, both in terms of lodging and other revenue (Figure 18). The analysis of customers per distribution channel also showed that travel agents/operators are the most common distribution channel used to make reservations at the hotel (Figure 19).

4. RESULTS EVALUATION

Limit analysis of the current customer segment is thought to be one of the major causes for the poor performance of Hotel H in keeping its current clients and acquiring new ones. Consequently, the reached solution will give a more in-depth view into their niche of the market and become a more viable solution. Following these clients were categorized into one out of the five clusters designed as the best description of the customers regarding demographic, behavioural, and geographic parameters.

Firstly, in the cluster "Mature Traveler Comfort" (cluster 0 – 30% of the total customers) 70% of the customers have ages above the 40 years old mark. In addition, 20% of these people are German while other top 5 nationalities represent 40% of the cluster. Moreover, most people in the cluster are not planners since 76% of them make their reservations less than 2 months in advance. Importantly, 32% of this group is included in the lower quartile of total revenue which means that it has a big portion of customers that poorly contributed to the business's revenue, and only 15% of the clients on the cluster are included in the upper quartile. Possibly, related to that, the cluster presents the lowest total number of rooms per night of all clusters, the highest rate of request for twin-size beds, and no customers in the cluster specially asked for a king-size bed. This cluster is full of people who book from Travel Agents or Operators.

Following this, "Diverse Direct Bookers" (cluster 1 – 15% of the total clients) is approximately 20% constituted by Portuguese and 8% by Spanish, where the remaining are distributed through a vast list of nationalities. 80% of this group of clients book their reservations directly in the hotel, while the rest use other types of booking. Most book their reservations 2 months before, with a few being a bit more prepared. Age-wise 30% of them are between 40 and 50 years old, while only 10% are below the 10% mark this group presents the least number of older customers above 60 years old. Around 34% of the customers are below the first quartile of revenue, higher than any other cluster, but otherwise present dispersed levels of revenue. Also, the segment presents the most loyal customers, even though low at the overall scale, about 12% of the customers are repeated ones, this being, with more than one check-in. Interestingly enough due to the higher predominance of Portuguese customers, this cluster also presents a higher number of customers with invalid bookings (cancelled or never showed up).

The following cluster, “Crown Comfort Seekers” (cluster 2 – 20% of the total customers), is predominantly composed of customers who book their reservations less than 2 months from the reservation, and all of them request king-size beds. Although very diverse in nationalities, 97% of them come from travel agencies or operator bookings (with no customers coming from direct booking). Regarding the age, 70% are above 40 years old. For other special requests, 21% of them have a unique request (besides bed size), but not many ask specifically for a twin-size bed. Regarding revenue coming from these clusters, there are no customers in the top quartile, being that 71% of them are between the first and third quartile of the total revenue.

Lastly, cluster “Top Spenders Collective” (cluster 4 – 18% of the total clients) as the name suggests is composed of the customers of the higher quartile of the total revenues, meaning that these are the customers that spend the most at the hotel. Most of them book between 2 and 6 months in advance, with a few that are more planned with more than 6 months. They usually come from travel agencies and operators (86%) while some come from direct contact with the hotel (12%). There is a high diversification, but as with the rest of the clientele, most customers are above 40 years old from multiple nationalities. They are on the line for the ones that request the most special requests, with 40% of them requiring a king-size bed and 17% a twin-size bed. Even though they are the higher paying customers, only 5% of them are repeat customers.

Although the analysis primarily emphasizes booking lead time to categorize customers, and some information regarding revenue, special requests, and distribution channels, in a general sense, it neglects other factors such as age and nationality.

5. DEPLOYMENT AND MAINTENANCE PLANS

Several steps have to be taken to have a successful deployment of a model.

Primarily, taking into consideration the low-quality market segmentation presented by Chain C and Hotel H, it was possible to redraw the assumption that this has a low investment in technology integration, marketing, and information management. Therefore, the first step is the model integration which will require specialised employees to apply it for a smooth assimilation likewise study what input the new model will accept and the output expected. Also, cooperation with the IT team is necessary to explore the models with the best compatibility to make the process as simple as possible. Moreover, mocked-up testing will allow us to verify if it behaves according to the expectations.

Next, a key point for its success is defining a data pipeline that will allow to use of real-time data, keeping everyone updated, and Chain C and Hotel H making use of all its features to create value. In another words, performing preprocessing and feature extraction allows efficient and accurate conclusions in a shorter period without having to hire external companies. Furthermore, this step is also important to define its flexibility as well as how it performs when confronted with errors, allowing it to take countermeasures in the eventuality of such a thing happening.

Furthermore, it is of the utmost importance to make an informed decision regarding the deployment platform, whether it should be cloud or sight while considering important factors, namely, its scalability and if it can handle any kind of setback it might come to face. Additionally, it is necessary to ensure that this is equipped with dependencies that allow all hierarchical levels to have access to the data they need. Nowadays, there is a constant technology change thus companies must keep updated with the latest cutting-edge technologies so when going through this process they must ensure the model is prepared for horizontal scaling.

Fourthly, it is testing where the product's performance is evaluated through tests to check if all the previous steps are being correctly executed so that the model can be successfully implemented.

Afterwards comes monitoring which has building KPIs as based lines to check the model’s functionality, increase the quality of the provided data as well and prevent anomalies and if any of these happen the stakeholders are informed of the problems.

Next is a feedback loop, which is an included feature that keeps a record of model performance, and KPIs reached and that includes feedback from its many users (stakeholders, employees, …) that when later observed can allow a tailor-made update of the model.

Lastly, documentation which is a throw description of all steps taken to not only the hotel knowing its model but also in case of any issues can be used as a guideline.

Nevertheless, the process is not over since it is necessary to do continuous maintenance of the model such as checking for deterioration, no business stays the same for long periods so this must be calibrated to the company’s needs. Moreover, using the feedback created by the model, it is possible to explore improvement areas. Finally, the model must follow privacy and security regulations for both customer and business protection.

6. CONCLUSIONS

To summarise the project, many factors showed how poorly the company is at keeping and establishing a well-founded customer base since the majority of the records showed one-time clients which illustrates that customers are not loyal or highly interested in Hotel H. Consequently, with this market analysis esteemed insights allowed to build profiles on the company's customers, making it possible to develop tailor-made marketing strategies for each group of individuals.

Although all groups have their marketing plan, it should be created a loyalty program to encourage returning to the hotel chain, since over 90% of the customers never went back to Hotel H but with this initiative, it is possible to create loyalty within Chain C. Furthermore, considering the high variety of nationalities, the company should also invest in multicultural marketing campaigns that tackle worldwide problems.

The group "Mature Travellers Comfort" shows low expenses when using the hotel's facilities so one approach to this problem is creating target advertisements and special offers (discounts, for example) so that they may feel more inclined to purchase further services. Moreover, promoting word-of-mouth publicity by incentivising clients to share their experience with their acquainted and with this receive a promoter discount with the C chain. Additionally, the company should develop local partnerships to offer personalised and exclusive packages with other local companies thus way expanding the company's reach. These initiatives fit these people since this group is mostly composed of people over 40 who come from generations where human interaction is more valuable than online advertisement. Another factor that could contribute to these customers' retention is investing in redecoration, specifically, in good quality and peaceful twin beds since this is the group that requested the most this type of room.

"Diverse Direct Bookers" seem to be spontaneous people since most bookings were performed face-to-face, at check-in. Moreover, this is the group of people with the highest percentage of return to the hotel which means that these people must be valued by the company. To do so an approach could be lower prices, having extra amenities available as well as performing room upgrades (these measures should be applied with constraints such as by bookings frequency). However, to avoid overloading employees (without warning, cleaners are not aware of how many rooms have to be available and sometimes the hotel may not have rooms available), it is necessary to change these people's habits and so promote earlier bookings with discount or with more flexible policies such as cancellation. Community marketing is the best option to reach these people since many of them look for recommendations from previous travellers and usually, these are inserted in travelling communities which can be reached through networking and events. In addition, clients' feedback can be used to make improvements.

The group "Crown Comfort Seekers" is composed of people with different backgrounds and ages which means that a diversity marketing plan should be applied to reach as many people as possible, being the best way through online channels. These people all have a preference for king-size beds showing that they are more inclined to have special requests and luxury. Consequently, renovation of these rooms as well as making VIP experiences is a good approach to personalise their services.

The "High-end travellers" cluster should be targeted with experimental marketing, in other words, it should be made available activities that would vitrine luxury and premium treats. Some examples could be wine tasting or food tastings with guest chefs. What is more, whenever people of this group make reservations there should be a premade list including their usual special requests to not only make this the minimum services made available to them but also add extra amenities that would captivate the client's attention. Consequently, to get all these perks available, the company should forge partnerships with elite influencers to accredit the hotel with credibility. Moreover, the hotel should invest in employing concierges to provide the optimum luxury experience.

To sum up, all these are marketing plans that allow to retain the current clients and people that fit into the description of these clusters but there is a clear opportunity in the younger market since these are an extremely low percentage of the customer base. Therefore, first, it is advised that this population is carefully analysed such as their spending patterns when they travel and try to have these available when they book with the company (after finding these customers' needs, those must be advertised to make the hotel more appealing to them). In addition, the under 30 years old group is greatly involved on social media so this should be the primary tool of advertisement (when using it as a marketing campaign, the content must be highly engaging since there are many ads on the internet and Hotel H wants to standout – TikTok's and Instagram posts). Young people will also expect the hotel to make available activities that explore their energy and needs (adventure activities, parties, etc.).

6.1. Considerations For Model Improvement

The main focus of the solution presented is enhancing customer retention. However, for model improvement, some considerations must be considered.

Primarily, the key to the success of this and any other model is data so the business must invest in robust data assemblage processes. Nevertheless, this alone will not work, it is of great importance that the features selected capture the diversity of behaviours of the customers.

Nowadays, technology and people's behaviour are constantly changing. Therefore, keeping up with the model's performance and adjusting it according to feedback will allow for the company to be ahead of trends and problems. What is more, constantly changing KPIs as well as updating marketing campaigns will enhance performance. Hence, it is required that the company invests in cutting-edge technologies such as a more robust CRM and apply tracing mechanisms which analyse customer trustworthiness and contentment.

Moreover, to optimise this process to the maximum, it should promote collaboration throughout all departments so that information is shared, being the best approach to the creation of a multi-specialised team that includes people from operations, marketing, data science and customer services.

In conclusion, if the recommendations made above are taken into consideration, Hotel H can increase exponentially its customer retention, building strong loyal relationships.

7. REFERENCES

How to count duplicate rows in the panda's data frame? (n.d.). Stack Overflow. Retrieved March 13, 2024, from https://stackoverflow.com/questions/35584085/how-to-count-duplicate-rows-in-pandas-dataframe

How to disable Python warnings? (n.d.). Stack Overflow. Retrieved March 13, 2024, from https://stackoverflow.com/questions/14463277/how-to-disable-python-warnings

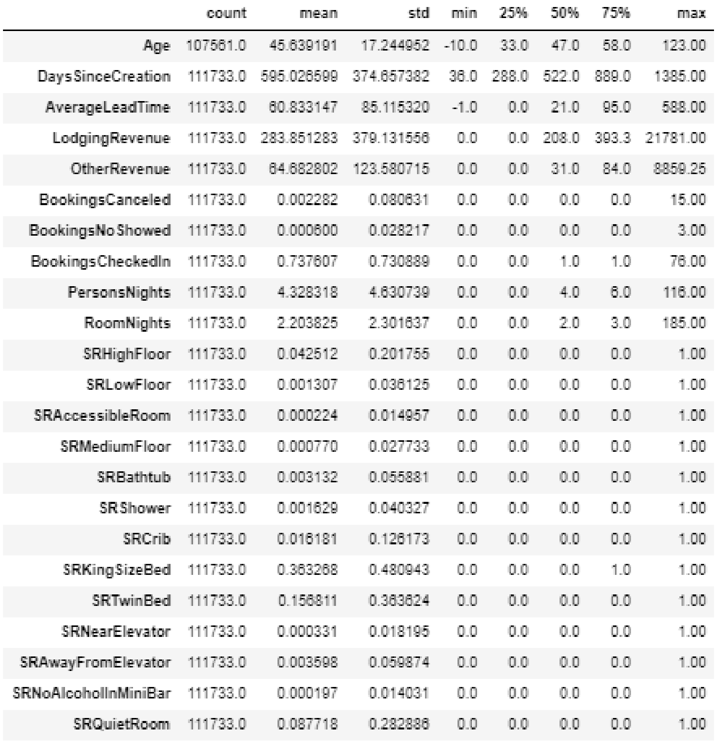
Marques, A. (2022, December 20). How to show all columns and rows in a Pandas DataFrame. Built-In. https://builtin.com/data-science/pandas-show-all-columns

Seaborn.Pairplot — seaborn 0.13.2 documentation. (n.d.). Pydata.org. Retrieved March 13, 2024, from https://seaborn.pydata.org/generated/seaborn.pairplot.html

Silhouette visualizer — yellowbrick v1.5 documentation. (n.d.). Scikit-yb.org. Retrieved March 13, 2024, from https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html

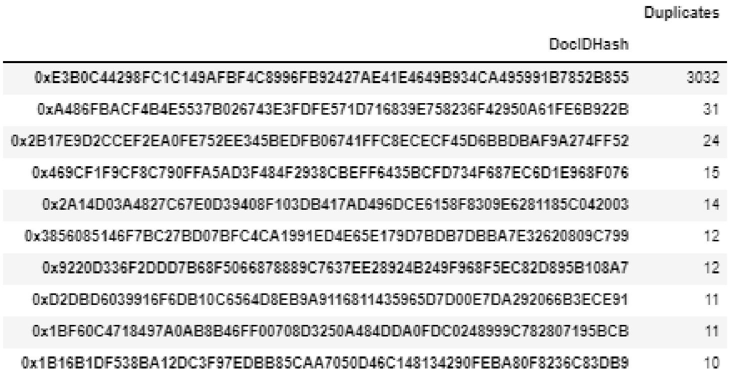
Colcol, S. (2017, September 18). How far do hotel guests book in advance? SiteMinder. https://www.siteminder.com/r/hotel-distribution/hotel-revenue-management/hotel-guests-book-advance/

8. APPENDIX



	count	mean	std	min	25%	50%	75%	max
Age	107561.0	45.639191	17.244952	-10.0	33.0	47.0	58.0	123.00
DaysSinceCreation	111733.0	595.026599	374.657382	36.0	288.0	522.0	889.0	1385.00
AverageLeadTime	111733.0	60.833147	85.115320	-1.0	0.0	21.0	95.0	588.00
LodgingRevenue	111733.0	283.851283	379.131556	0.0	0.0	208.0	393.3	21781.00
OtherRevenue	111733.0	64.682802	123.580715	0.0	0.0	31.0	84.0	8859.25
BookingsCanceled	111733.0	0.002282	0.080631	0.0	0.0	0.0	0.0	15.00
BookingsNoShowed	111733.0	0.000600	0.028217	0.0	0.0	0.0	0.0	3.00
BookingsCheckedIn	111733.0	0.737607	0.730889	0.0	0.0	1.0	1.0	76.00
PersonsNights	111733.0	4.328318	4.630739	0.0	0.0	4.0	6.0	116.00
RoomNights	111733.0	2.203825	2.301637	0.0	0.0	2.0	3.0	185.00
SRHighFloor	111733.0	0.042512	0.201755	0.0	0.0	0.0	0.0	1.00
SRLowFloor	111733.0	0.001307	0.036125	0.0	0.0	0.0	0.0	1.00
SRAccessibleRoom	111733.0	0.000224	0.014957	0.0	0.0	0.0	0.0	1.00
SRMediumFloor	111733.0	0.000770	0.027733	0.0	0.0	0.0	0.0	1.00
SRBathtub	111733.0	0.003132	0.055881	0.0	0.0	0.0	0.0	1.00
SRShower	111733.0	0.001629	0.040327	0.0	0.0	0.0	0.0	1.00
SRCrib	111733.0	0.016181	0.126173	0.0	0.0	0.0	0.0	1.00
SRKingSizeBed	111733.0	0.363268	0.480943	0.0	0.0	0.0	1.0	1.00
SRTwinBed	111733.0	0.156811	0.363624	0.0	0.0	0.0	0.0	1.00
SRNearElevator	111733.0	0.000331	0.018195	0.0	0.0	0.0	0.0	1.00
SRAwayFromElevator	111733.0	0.003598	0.059874	0.0	0.0	0.0	0.0	1.00
SRNoAlcoholInMiniBar	111733.0	0.000197	0.014031	0.0	0.0	0.0	0.0	1.00
SRQuietRoom	111733.0	0.087718	0.282886	0.0	0.0	0.0	0.0	1.00

Figure 1 – Summary Statistic of Numerical Data



	Duplicates
DocIDHash	
0xE3B0C44298FC1C149AFBF4C8996FB92427AE41E4649B934CA495991B7852B855	3032
0xA486FBACF4B4E5537B026743E3FDFE571D716839E758236F42950A61FE6B922B	31
0x2B17E9D2CCEF2EA0FE752EE345BEDFB06741FFC8ECECF45D6BBDBAF9A274FF52	24
0x469CF1F9CF8C790FFA5AD3F484F2938CBEFF6435BCFD734F687EC6D1E968F076	15
0x2A14D03A4827C67E0D39408F103DB417AD496DCE6158F8309E6281185C042003	14
0x3856085146F7BC27BD07BFC4CA1991ED4E65E179D7BDB7DBBA7E32620809C799	12
0x9220D336F2DDD7B68F5066878889C7637EE28924B249F968F5EC82D895B108A7	12
0xD2DBD6039916F6DB10C6564D8EB9A9116811435965D7D00E7DA292066B3ECE91	11
0x1BF60C4718497A0AB8B46FF00708D3250A484DDA0FDC0248999C782807195BCB	11
0x1B16B1DF538BA12DC3F97EDBB85CAA7050D46C148134290FEBA80F8236C83DB9	10

Figure 2 – Top 10 Document IDs with the Most Frequent Occurrences

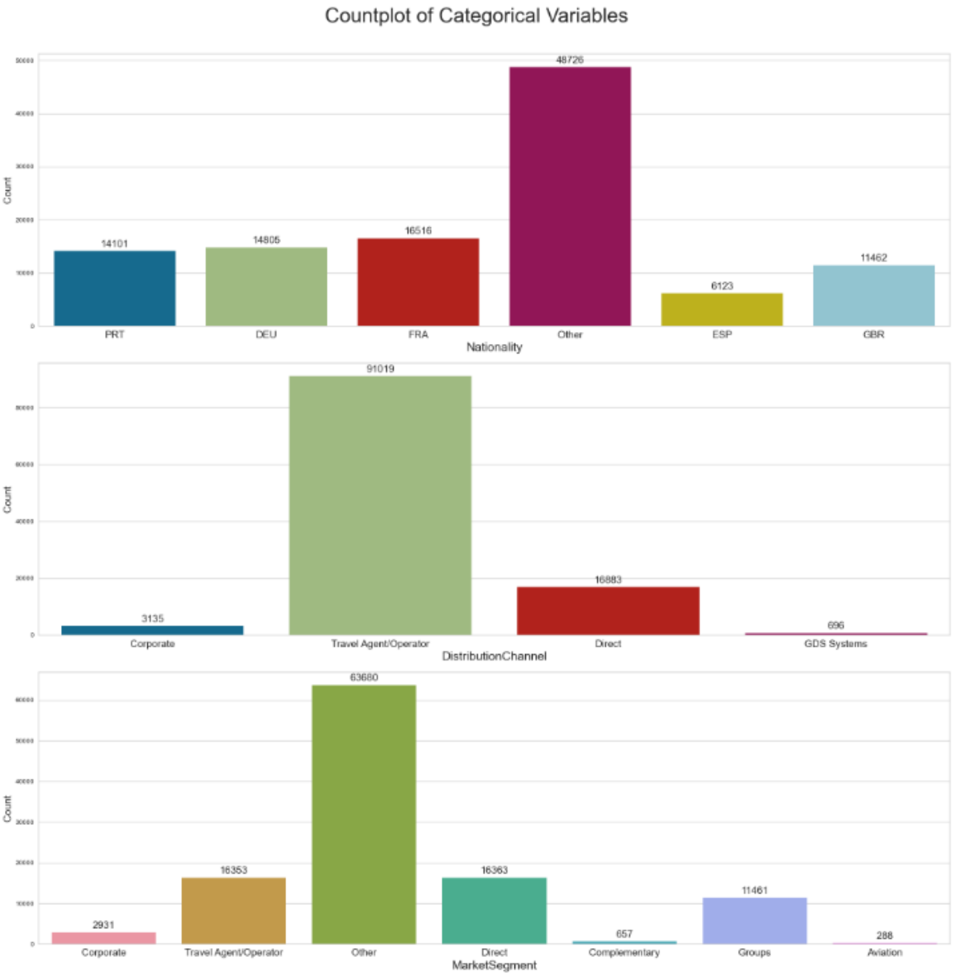


Figure 3 – Countplot of Categorical Variables

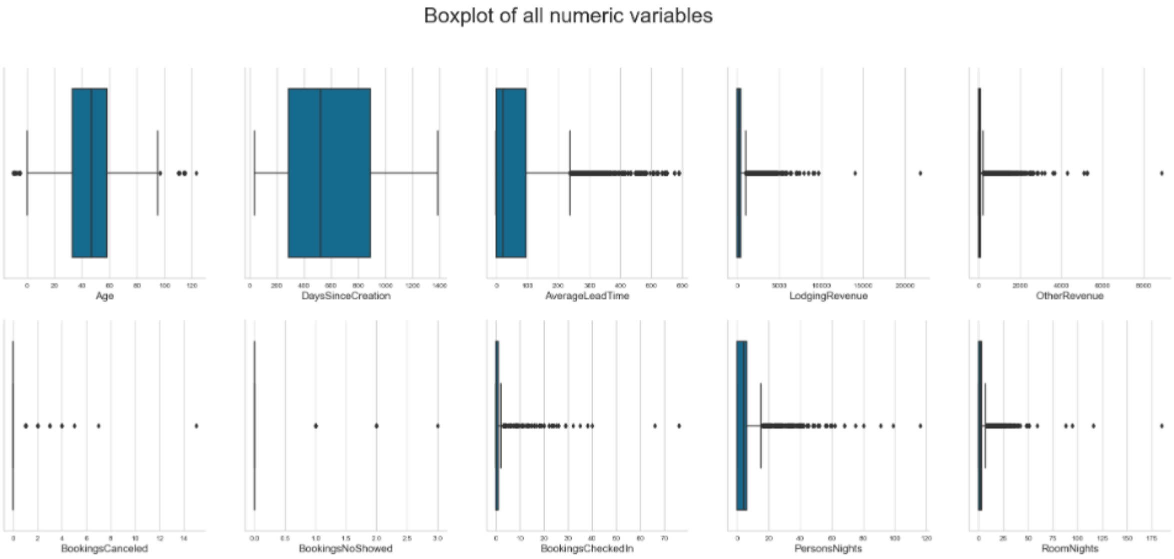


Figure 4 – Boxplot of Numerical Variables

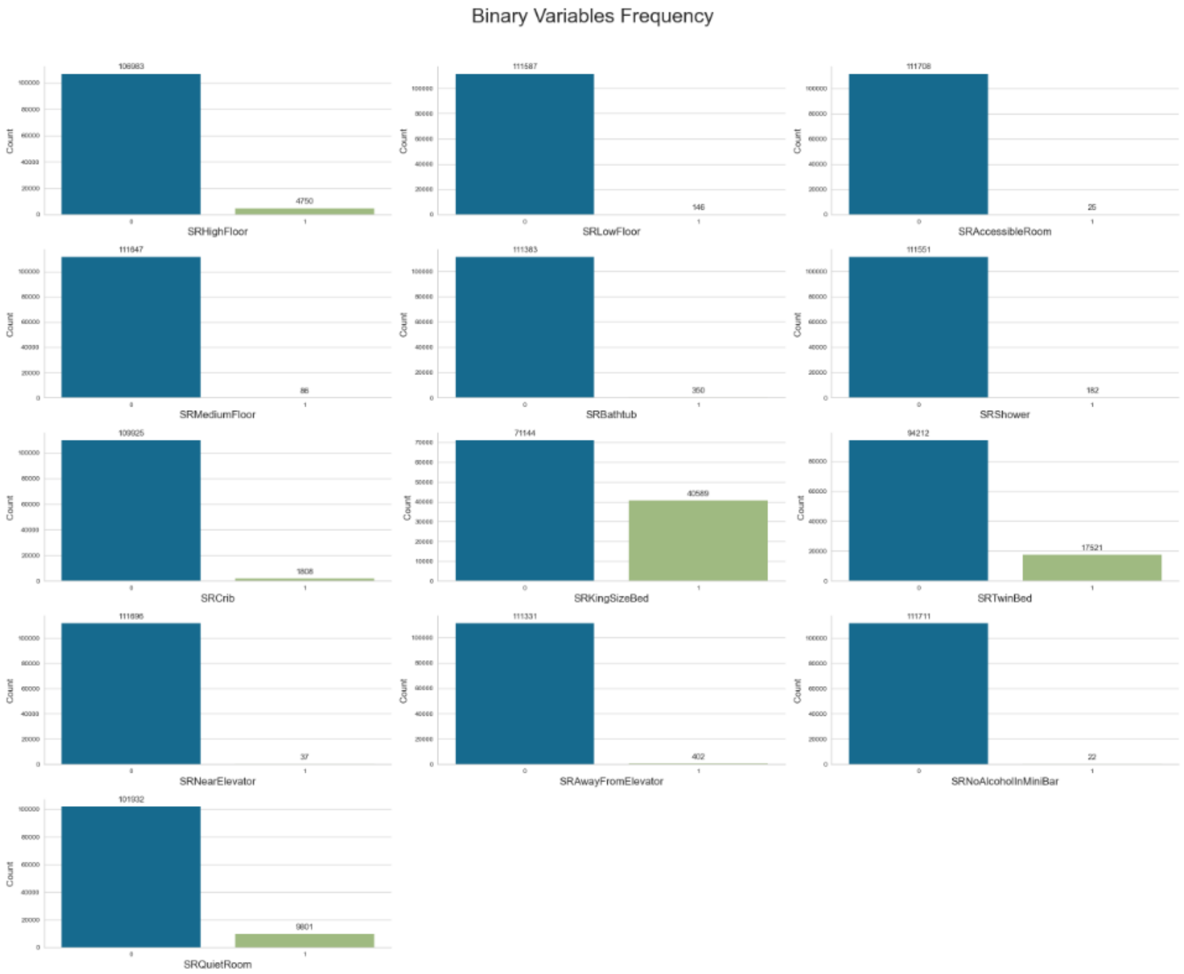


Figure 5 – Countplot of Binary Variables

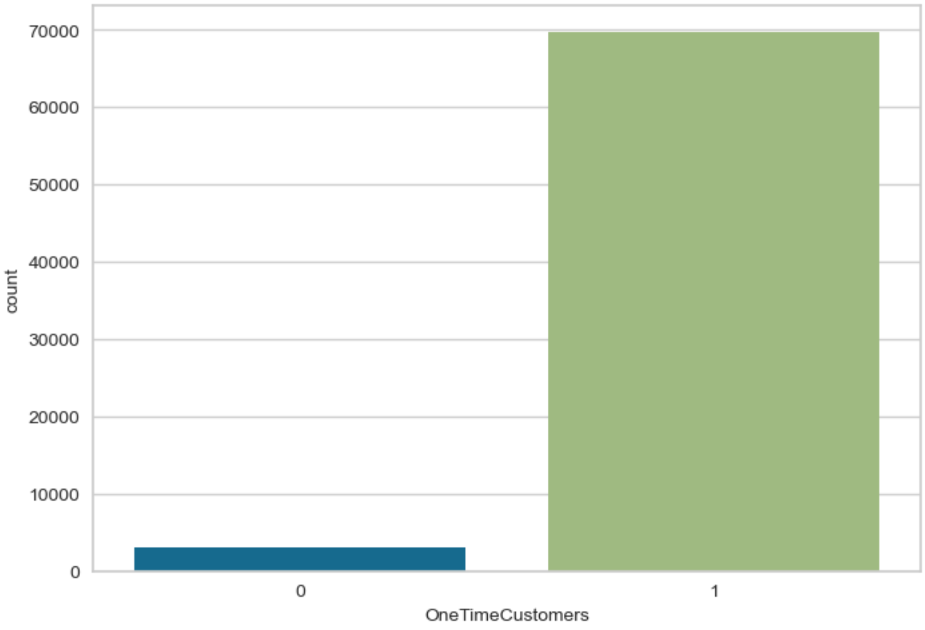


Figure 6 – Countplot of One-Time Customers

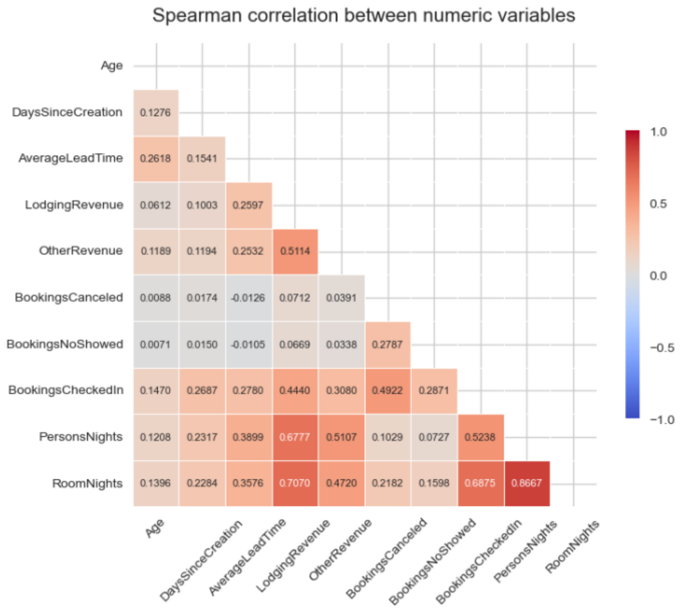


Figure 7 – Spearman Correlation Matrix

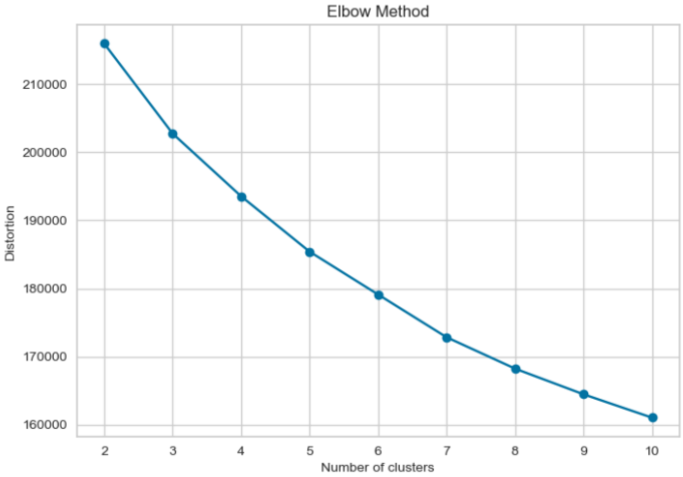


Figure 8 – Elbow Method Graph

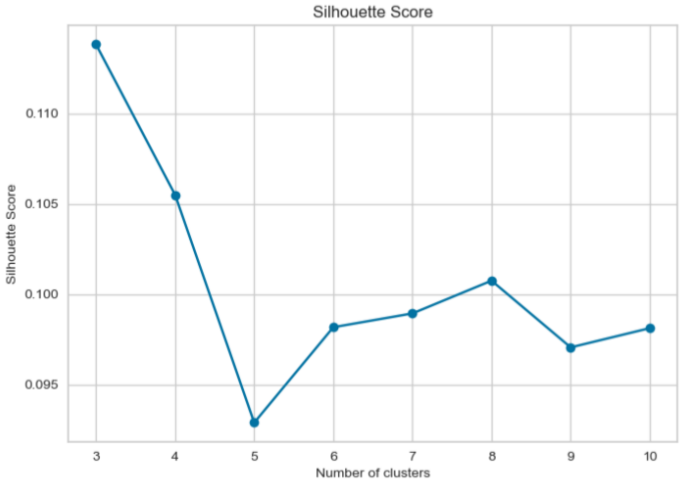


Figure 9 – Silhouette Score Graph

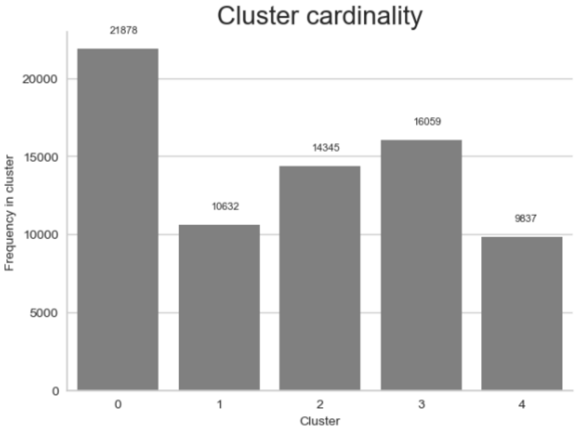


Figure 10 – Cluster Cardinality Graph

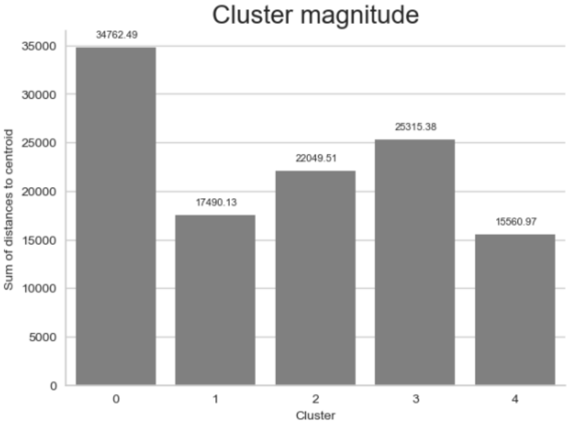


Figure 11 – Cluster Magnitude Graph

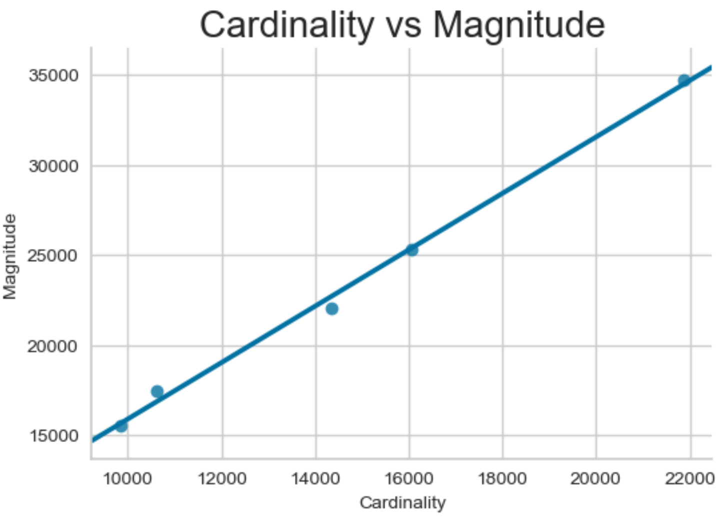


Figure 12 – Plot of Cardinality vs Magnitude

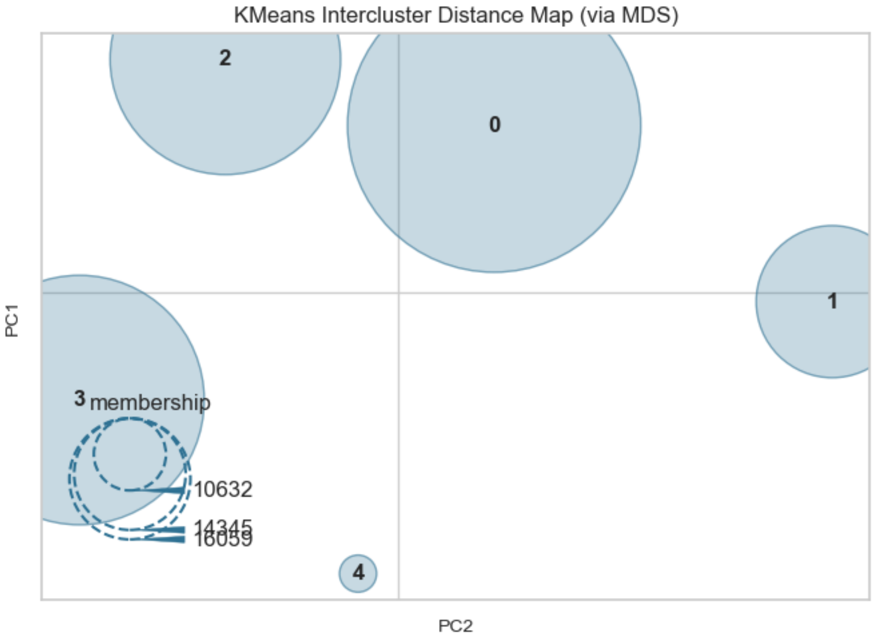
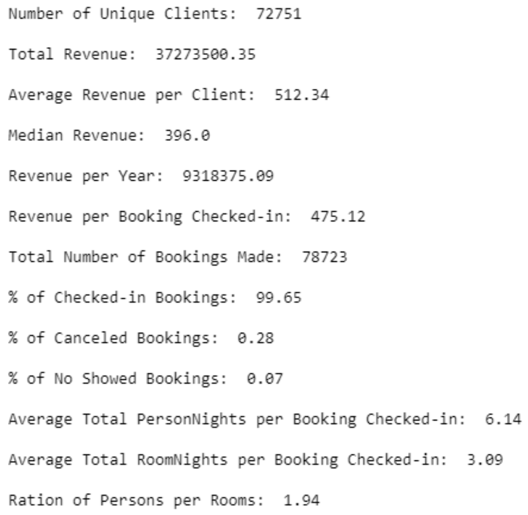


Figure 13 – Kmeans Intercluster Distance Map



```
Number of Unique Clients:  72751

Total Revenue:  37273500.35

Average Revenue per Client:  512.34

Median Revenue:  396.0

Revenue per Year:  9318375.09

Revenue per Booking Checked-in:  475.12

Total Number of Bookings Made:  78723

% of Checked-in Bookings:  99.65

% of Canceled Bookings:  0.28

% of No Showed Bookings:  0.07

Average Total PersonNights per Booking Checked-in:  6.14

Average Total RoomNights per Booking Checked-in:  3.09

Ration of Persons per Rooms:  1.94
```

Figure 14 – KPIs

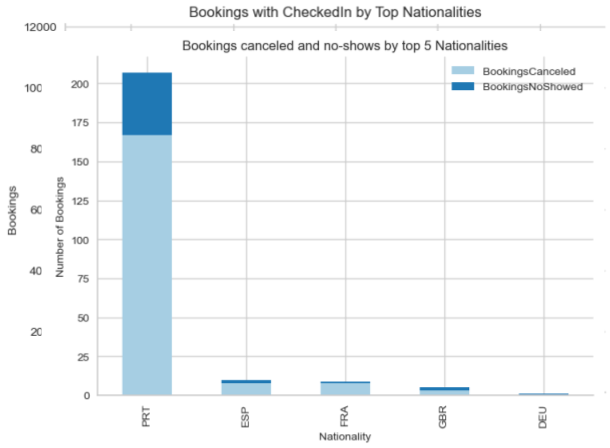


Figure 15 – Check-in per Top Nationality

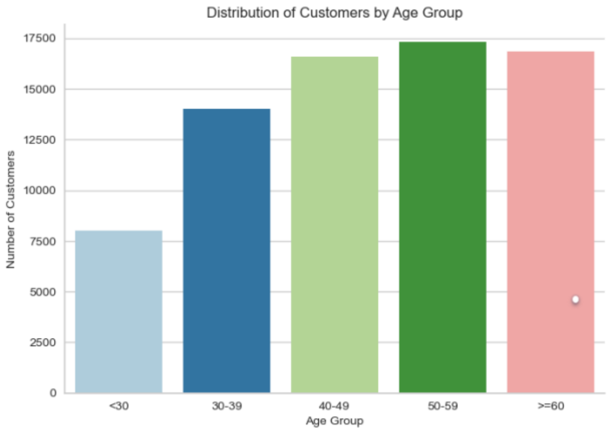


Figura 16 – Customer per Age Group

Figure 17 – Invalid Bookings per Top Nationality

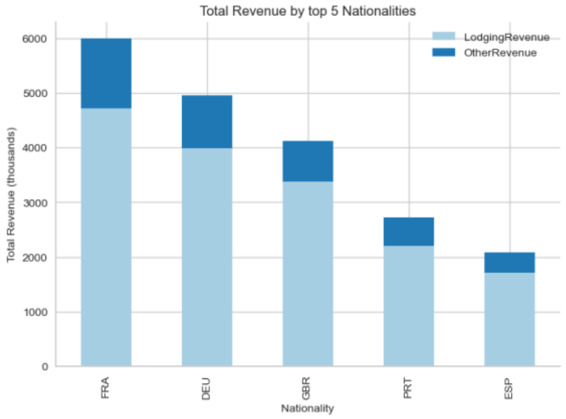


Figure 18 – Total Revenue per Top Nationality

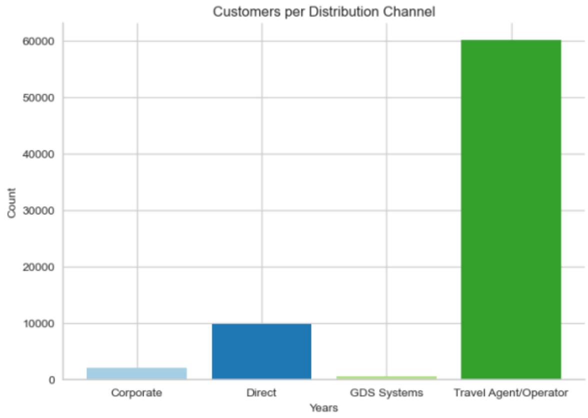


Figure 19 – Customer per Dist. Channel