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

Jaime Duarte

Shanjida Roman

Yousef Ebrahimi

Diogo Martins

David Martins

Group D – Data Vision Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

INDEX

1. EXECUTIVE SUMMARY

Data Vision Analytics was founded by a group of five students from NOVA IMS, Data Science in Business Analytics program, aspiring to mark the footsteps to the modern era of data science in business. Five enthusiastic members from four different nationalities, each equipped with their own unique abilities and experiences, now working simultaneously on this hotel customer segmentation project.

As we are all aware, there is harsh competition in the hotel and tourism industry. Businesses are now more in need of understanding their customers to sustain and grow their businesses.

The focus of this project is to understand current customers characteristics in terms of revenue brought to the company, geography, demography, psychography and consumer behavior. Data Vision analyses the given data using the latest methods and technological tools. Our plan is to find patterns in the features of the clients, allowing us to assign each client to a different group. Customer segmentation identifies the discrete group of customers with a high level of accuracy based on current customer data. This solution will allow the business to enable its marketing department and product developers to improve their business strategy by addressing customers individually in a more effective manner.

Following the creation of these clusters, we will provide some insights of our own regarding the results obtained, to help make the hotel make strategic decisions to retain and attract customers. This will include our recommendations for the deployment of our solution, as well as any monitoring and maintenance measures necessary.

2. INTRODUCTION

Hotel H is looking to grow by acquiring new customers and optimizing their catering to the current ones. Since its bigger than before hotel chain board decided to invest more in marketing and they have created a marketing department with new marketing manager. The current system already has customer segmentation; however, new management feels that the current model is not productive enough to meet clients' needs and attract new customers.

Management can set up new plans and processes in a very effective manner with new segmentation. Data vision is aware that segmentation must be adaptable, durable, scalable, and flexible in order to be used for future requirements.

Data Vision has chosen to use the CRISP-DM methodology to complete this project. In recent years, the most widely used methodology for data science has been the Cross Industry Standard Process for Data Mining (CRISP-DM), which is a process model that has become standardized across industries. This methodology helps with project planning and management, encourages best practices, and yields the best results. It can also record framework and replicate projects.

Figure 1 illustrates the six sequential phases of CRISM DM, starting with business understanding, followed by data preparation, modeling, evaluation, and deployment.

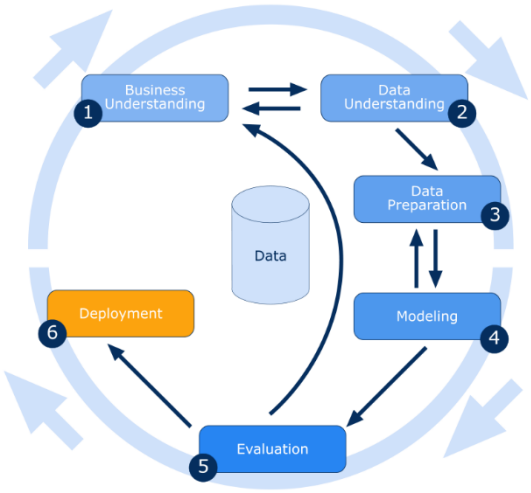

Figure 1: CRISP-DM methodology

To achieve the goal in the best possible way, Data Vision strictly adheres to these steps. In the following pages of this report, we explicitly explain how we conducted the analysis and segmented the target market for hotel H's best marketing development process at each stage.

3. CUSTOMER SEGMENTATION WITH CRISP-DM METHODOLODY

3.1. BUSINESS UNDERSTANDING

From our understanding, Hotel H receives customers from a variety of distribution channels, including travel agencies, tour operators, online travel agencies, and company websites. Not to mention, the hotel hosts visitors from many different nationalities who come from all over the world. Some visitors book far in advance than others, while others book a stay right away. It is impossible for the marketing manager to meet individual needs without better segmentation.

3.1.1. Business Objetives

According to Data Vision, the primary goals of the company are to reach every customer and meet their specific accommodation needs. For this, developing targeted marketing strategies to draw

in new and returning customers is a key goal. In addition, Hotel H wants to concentrate on offering top-notch customer service, and maintaining sustainable growth based on the analysis of available data.

3.1.2. Business Success Criteria

Data Vision is confident in this project's success and, by extension, in the success of the business deliverables throughout the project. In order to accomplish this, we ensured that the data was accurate, that pertinent insights were offered, and that the data would be used going forward.

We analyzed the data and created the model to the best of our abilities using our experiences, knowledge, and resources. A successful endeavor would be one in which a high quality and reliable model is produced. This model should result in the creation of clearly heterogeneous clusters, that will allow for the extraction of useful insights to be used at will by the Hotel.

The model should be able to handle any increase in data volume and be as efficient as possible. The project must also comply with all applicable ethical and legal requirements for data protection.

3.1.3. Situation Assessment

The currently available data is somewhat flawed but will be of great use. The company keeps satisfactory records, which will allow for quality analysis. However, a good data preparation process will be needed in order to extract meaningful insights. Advice regarding future data collection practices will need to be provided, as it will improve future operations. Hardware and software assessments or recommendations will not be conducted, as these fall outside this project's scope.

The only meaningful constraint for the fulfillment of this project is time. We don't believe any restraints at the resource; legal or ethical level will arise.

3.1.4. Determine Data Mining goals

As we've stated, our goal with this project is to deliver a quality customer segmentation model and insights into the results of the said model and its future applications. More specifically, the goal of this segmentation is to facilitate the creation of targeted marketing campaigns, which we at Data Vision Analytics and Hotel H's management, believe will bring about a positive change in customer acquisition and retention.

3.2. DATA UNDERSTANDING

This phase begins with data collection and proceeds with activities that will enable us to become acquainted with the data. This is done to identify any obvious problems and insights.

3.2.1 Collect Initial Data – The data we received is in csv file that contains 111,733 rows and 29 columns in total. Figure 3 shows the overview of collected data. At first sight, it seems to be quality data, with very dew

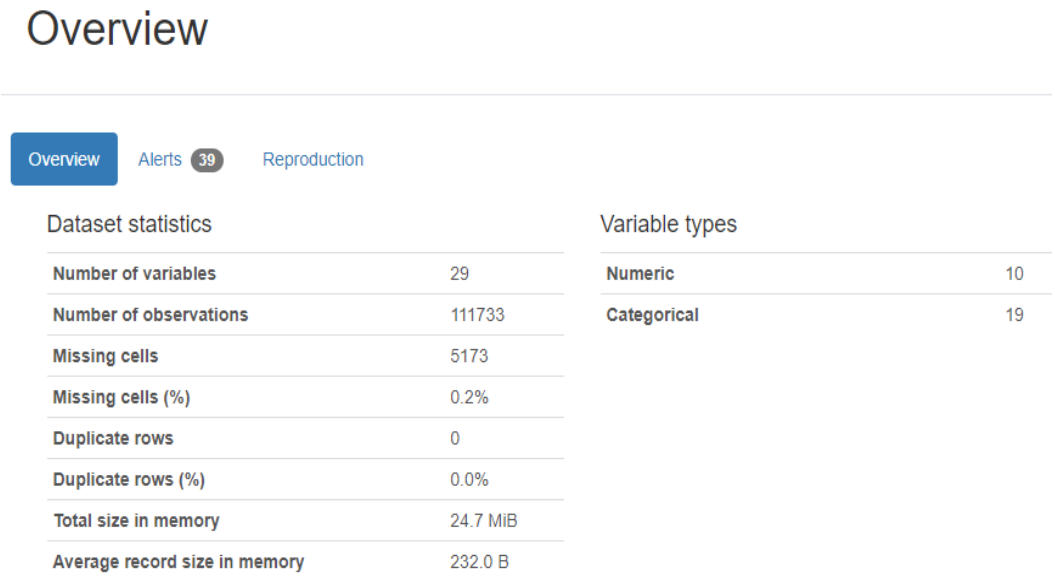

Figure 2: Overview of whole database

3.2.2 Describe Data – The dataset has numerical and categorical values including different datatypes such as int64, object, float64. 29 columns describe the nature of data that has been collected from existing customers. Dataset came with meta data which describes each field and what it means for the business. Some variables have duplicate values and null values.

```
RangeIndex: 111733 entries, 0 to 111732
Data columns (total 29 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   ID                    111733 non-null  int64
 1   Nationality           111733 non-null  object
 2   Age                   107561 non-null  float64
 3   DaysSinceCreation     111733 non-null  int64
 4   NameHash              111733 non-null  object
 5   DocIDHash             110732 non-null  object
 6   AverageLeadTime       111733 non-null  int64
 7   LodgingRevenue        111733 non-null  float64
 8   OtherRevenue          111733 non-null  float64
 9   BookingsCanceled      111733 non-null  int64
 10  BookingsNoShowed      111733 non-null  int64
 11  BookingsCheckedIn     111733 non-null  int64
 12  PersonsNights         111733 non-null  int64
 13  RoomNights            111733 non-null  int64
 14  DistributionChannel   111733 non-null  object
 15  MarketSegment         111733 non-null  object
 16  SRHighFloor           111733 non-null  int64
 17  SRLowFloor            111733 non-null  int64
 18  SRAccessibleRoom      111733 non-null  int64
 19  SRMediumFloor         111733 non-null  int64
 20  SRBathtub             111733 non-null  int64
 21  SRShower              111733 non-null  int64
 22  SRCrib                111733 non-null  int64
 23  SRKingSizeBed         111733 non-null  int64
 24  SRTwinBed             111733 non-null  int64
 25  SRNearElevator        111733 non-null  int64
 26  SRAwayFromElevator    111733 non-null  int64
 27  SRNoAlcoholInMiniBar  111733 non-null  int64
 28  SRQuietRoom           111733 non-null  int64
dtypes: float64(3), int64(21), object(5)
```

Figure 3: Data Information

3.3.3 Explore Data – We explored the dataset from hotel H looking at both the raw data and some visualizations. After setting "ID" as an index, we were able to see that there were 111 duplicated rows in the dataset and proceeded to eliminate them. We then noticed there many rows with the same "DocIDHash", which is supposed to be a unique value, as it represents an official identification of the costumer. This meant that different bookings by the same customer were being stored in different rows. This poses a problem, as these entries should be merged into one, revealing the true value of the client. We removed these duplicated "DocIDHash" rows, got the sum or the mode for each feature, depending on its characteristics, and merged the rows back into the dataset. This change reduced the number of rows by over 10000, revealing the true number of customers on record.

This allowed us to proceed with more confidence in our exploration of the data. We noticed some incoherent values in “Age” and “AverageLeadTime”. Boxplots and histograms were created to better visualize the distribution of values for each of the features, these are also important to address future removal of outliers with more accuracy. Some notable findings were a very small number of people who had a value of 1 in “BookingsCanceled” and “BookingsNoShowed”. By counting the values of “AverageLeadTime” we were able to confirm that about a third of the customers don’t do any bookings before arriving at the Hotel. On this topic, there were over 30000 values of 0 on the “BookingsCheckedIn” feature. This is very relevant since this number vastly exceeds the supposed number of people that didn’t show up/canceled given by the “BookingsNoShowed” and “BookingsCanceled” features. In our interpretation, these are people that might have attempted to make a booking but didn’t actually stay at the Hotel, we believe this hypothesis is being corroborated by our findings during customer segmentation, as we will show later on.

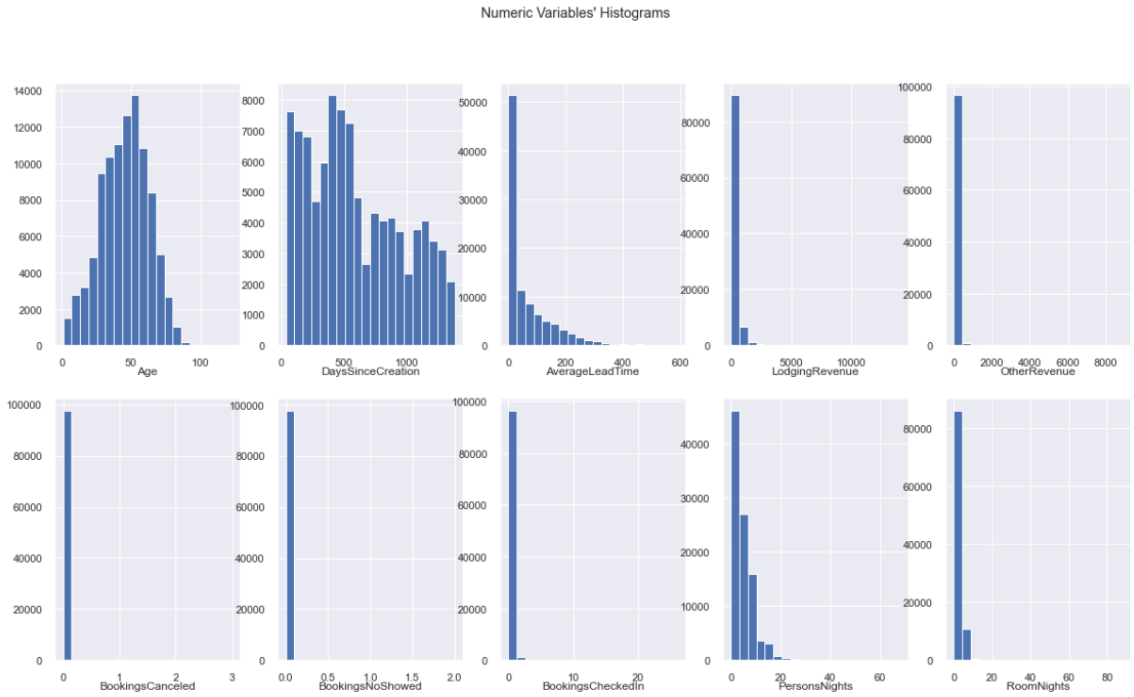

Figure 4: Metric Variables Histograms

This process was repeated for the categorical variables as well, with several value counts being performed for the different features. Almost all of Hotel H’s clients are European, with the vast majority being Portuguese.

Figure 5: Map of customer origin

We noticed a lot of features relative to client's requirements about the quietness, floor, bathroom equipment and more characteristics of the room had close to no values of 1. This could mean clients are not aware of the option to ask for these specifications, which would demand a change on the Hotel's behavior.

The most striking finding was on the "DistributionChannel" feature, where we were able to discern that the vast majority of clients are being directed to the Hotel trough a travel agency/operator. Only a very small minority of clients are coming through their corporation or GDS Systems.

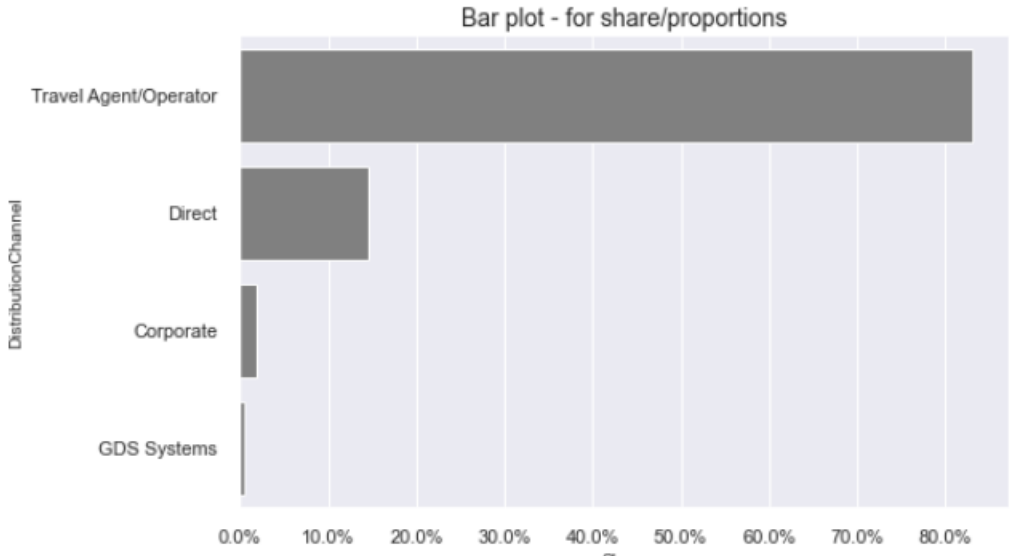

Figure 6: Share of clients by Distribution Channel

3.3.4 Verify Data Quality – Having explored the dataset, we felt that a simple data preparation phase was necessary. The quality of data provided by the Hotel was of good enough quality to perform our analysis.

3.3. DATA PREPARATION

3.3.1 Clean Data – Data was cleaned before visualizations. This was done by deleting duplicate values, missing values, strange or incoherent values and outliers. To help identify outliers, boxplots were used. This allowed for a manual process of handling outliers. Other methods like using the Inter-Quartile Range were performed, but were not used do to removing more data than desired.

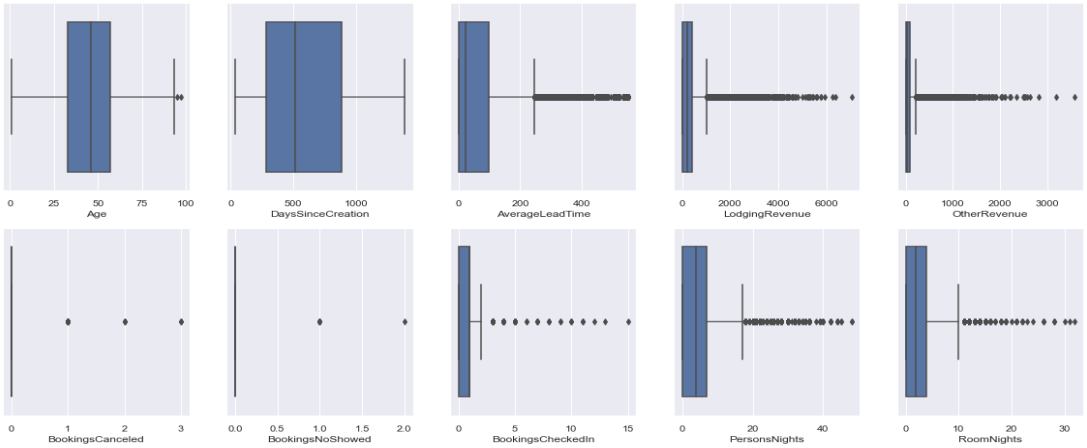

Figure 7: Boxplots of metric features

3.3.2 Construct Data – We replaced missing values, which were only present in the "Age" variable, using the KNN method. This method fills in the missing values by replacing them with the age of clients that are most similar to them.

3.3.3 Integrate Data – Since we felt that the "Nationality" feature would be problematic after encoding, due to the number of features it would create, we decided to modify it somewhat. We wanted to replace nationality with continent, and for that it was necessary to import a file that would allow us to convert the ISO codes available into the region. This was the only external data integrated into our workspace.

3.3.4 Format data/Feature engineering – In this stage, we created 3 new features, "Region", "Age_bins" and "TotalRevenue". "Age_bins" replaced the "Age" feature allowing for a better understanding of the data.

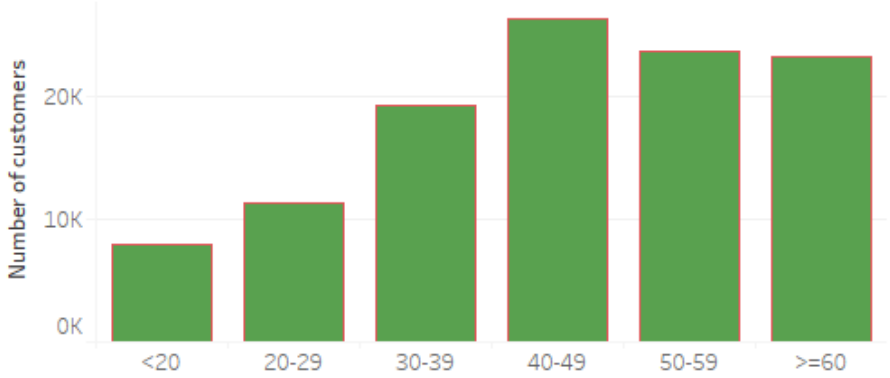

Figure 8: Age Bins

Total Revenue replaced the “OtherRevenue” and “LodgingRevenue” features, this allowed for a more straightforward analysis of the revenue brought to the Hotel by each client. Finally, “Region” was created to replace “Nationality”, for the reasons previously stated.

3.4. MODELING

3.4.1 Select Modeling Techniques - On the modeling select techniques we decide to use K means algorithm because is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. Principle Component Analysis (PCA) was also used to reduce dimensionality. The silhouette score and distortion score were used to find an appropriate number of clusters (K). We opted for the silhouette score’s result, since we felt 7 clusters allowed for the design of better targeted marketing strategies compared to the 4 clusters the distortion score recommended.

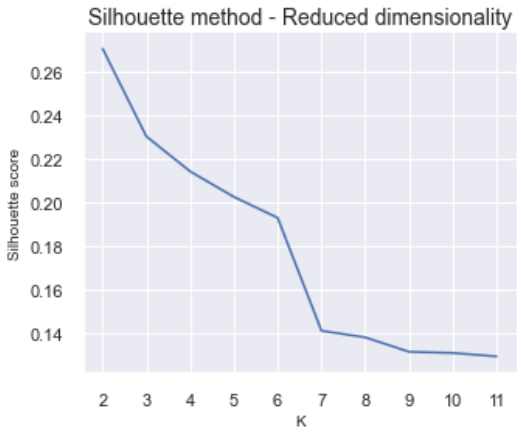

Figure 9: illustrates K-means chart, 7 clusters

3.4.2 Assess Model – To assess our K-Means model, we analyzed its Silhouette score as well as how useful the clustering solution was. Clusters had distinct characteristics and were quite homogeneous regarding cardinality, with cluster 6 being the exception. We believe a good result was achieved.

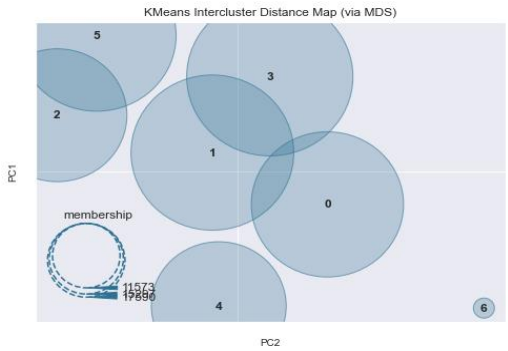

Figure 10: Visualizing the clusters sizes and distances in 2D

3.5. EVALUATION

3.5.1 Evaluate Results - Evaluation allows us to check the original business objectives and all other findings. The result shows that the objectives of the business meet, and some good, targeted marketing strategies developed.

To analyze the clustering, we need to first identify which variables are important for each cluster. Looking at the clusters, we can see that some variables have a higher range of values for certain clusters. To compare the different clusters, we can look at the maximum and minimum values for each variable across all clusters. Also, we can also calculate the mean value and standard deviation for each variable across all clusters. In summary:

All customers are mainly from Europe, and they use preferentially the Travel Agent/Operator Channel.

Cluster 0 "MV":

Middle Value Customers consists of customers that usually request a crib. They are relatively older (>= 50), with an average age of over 50 years old. They tend to stay for an average of 6 person/nights and have stayed in the hotel for a total of 3 room/nights. These customers usually request a quiet room and a room with a shower, away from the elevator. They generate a significant amount of total revenue for the hotel. Interestingly, these customers have age of creation in the middle (between the oldest and youngest) and they are the most booking earliest bookers customers.

Cluster 1 "NV" (A)":

No Value Customers consists of clients on the two sides of the ages lower than 20, and higher than 60, (children and seniors). Due to the age of this cluster, their hotels charges may be covered by a working-age adult hotel customer, that's why this group is the youngest customers on the Age of customers since creation. This group has room preferences, but seemingly no stays at the hotel (bringing in no revenue to the hotel). This can be either due to and error in data collection or because these people are shopping around and trying to find the best hotel for them.

Cluster 5 "VLV":

Very Low Value Customers, consists on customers between 30 and 59, they are the cluster with highest number of customers (qtd) , oldest Age since creation, and one of the lowest average lead time, so it means that they book very earliest to the date, it means they are looking always for the lowest prices (promotions) , mostly no checked-in and usually ask for a crib, a room with a king-size bed, quiet room, bathtub, and no alcohol in the mini-bar, and prefer a high floor.

Cluster 3 "HV":

High Value, in this Cluster Customers has the highest Total revenue, between the age of 40-59. They usually book the room directly. The number of adults and children in each booking, multiplied by the number of nights length-of-stay is 10. The room/nights the customer has stayed at the hotel so far is 4. They usually book the room many days before arrival date the customer

makes bookings. In addition, they ask for quiet room. They are the smallest group, and they are regular customer (Middle age since creation).

Cluster 6 "VHV":

Very High Value, in this Cluster Customers has the highest AVG revenue, between the age of 40-59. They usually book the room directly. The number of adults and children in each booking, multiplied by the number of nights length-of-stay is 16. The room/nights the customer has stayed at the hotel so far is 7. They usually book the room many days before arrival date the customer makes bookings. In addition, they ask for quiet room. They are the smallest group and they are regular customer (Middle age since creation).

Cluster 4 "LV":

Low Value Customers, consists on customers between 30 and 59, they are the cluster with second highest number of customers (qtd) , youngest Age since creation (newest) , and one of the lowest average lead time, so it means that they book very earliest to the date, it means they are looking always for the lowest prices (promotions) , mostly no checked-in and usually ask for a crib, a room with a king-size bed, quiet room, bathtub, and no alcohol in the mini-bar, and prefer a high floor.

Cluster 2 “NV” (B):

No Value Customers B, consists clients on the two sides of the ages lower than 20, and 50-59 (between 50 and 59) , (children and working age adults). This group has room preferences, but seemingly no stays at the hotel (bringing in no revenue). This can be either due to and error in data collection or because these people are shopping around and trying to find the best hotel for them.

3.5.2 Review Process - Data mining is an effective process that allows us to analyze raw data and extract valuable insights. Through this approach, we can identify important customers who contribute significantly to our revenue. By developing targeted marketing strategies for each specific group of customers, hotels can achieve their goals more efficiently while keeping their customers satisfied.

Specifically, the hotel's most valuable customers are those who spend the most money. Additionally, by utilizing segmentation techniques, the hotel can gain valuable information to develop targeted campaigns that will effectively reach these high-value customers.

3.5.3 Determine Next Step - To optimize the process, there are certain guidelines that can be followed. These include ensuring accurate data entry and implementing a universal booking software. Additionally, it is important to carefully review customer names and documents to avoid creating duplicate records.

To maintain data integrity, it is recommended to conduct this process regularly at set intervals, such as every season or every six months. This approach allows for the collection of data over time, enabling more informed decision-making for the long term.

3.6 DEPLOYMENT AND MAINTENANCE PLANS

3.6.1 Plan Deployment – Deployment should be accompanied by development of business strategies that take advantage of the customer segmentation:

- Targeted Marketing: Based on the clusters, the hotel can develop targeted marketing campaigns to attract customers from different regions, age groups and booking preferences. For instance, for customers in Europe who book by corporate, the hotel can offer loyalty programs and promotions to encourage them to stay longer and generate more revenue.
- Room Preferences: The hotel can use the information on customers' room preferences to tailor their room offerings. For example, they can ensure that customers in Cluster 3 are offered rooms quiet, while customers in Cluster 2 is offered king size beds.
- Travel Agents: For customers in Cluster 5, the hotel can establish partnerships with travel agents/corporate and operators to attract more customers from the Europe.
- Address issue present in clusters 2 and 5. These clusters have costumers revealing their room preferences, but not actually staying at the Hotel. If this is due to these costumers browsing around many Hotels, and choosing the best for them, it might be wise to offer them special discounts according to their preferences, since they seem to be choosing other hotels over Hotel H. It also might be due to an error in the data collection, which would also need to be addressed.

3.6.2 Plan Monitoring and maintenance

- Regular Updates: The hotel should continually update the customer clusters to reflect changing preferences and trends. They can use data from previous bookings to improve their understanding of their customers and ensure they continue to provide excellent customer service.
- Customer Feedback: The hotel can request customer feedback on their room preferences and overall experience. There are some requests regarding room specifications that, based on their values, costumers are possibly not aware of their ability to make. They can use this information to improve their services and tailor their offerings to suit their customers' needs.
- Revenue Management: The hotel can use revenue management techniques to optimize their revenue based on demand and supply. They can also use this information to tailor their offerings and promotions to encourage customers to book directly or stay longer.
- **Cluster 0:** Given that Cluster 0 customers tend to book through travel agencies the hotel should enhance travel agency partnerships. Offer them incentives to promote the hotel to their clients, such as exclusive deals or packages. Since these customers also tend to stay for longer periods, the hotel should consider creating packages that cater to their needs. Offer them discounts or special amenities such as free breakfast, airport shuttle, or spa services. Finally, Cluster 0 customers are valuable to the hotel, so it is essential to maintain their loyalty. The hotel should consider creating a loyalty program that rewards them with points or exclusive perks, such as free room upgrades, late check-out, or complimentary drinks.
- **Cluster 1:** Customers in this cluster primarily use corporate bookings and represent the latest addition to the hotel's customer base. The hotel can tailor its services to meet the specific needs and preferences of this group, such as offering corporate discounts, upgraded amenities, or personalized services for business travelers.

- Investigate why **Cluster 2** is not adding much value to the hotel: Couples without children are a small percentage of customers who do not add much value to the hotel. The hotel can investigate why this is the case and develop strategies to attract this group of customers, such as offering romantic packages or targeted marketing campaigns.
- Offer specialized services to **Cluster 4:** Given that this cluster consists of older customers, it may be beneficial to offer specialized amenities and services that cater to their needs and preferences. This can include offering mobility assistance, providing comfortable seating areas, offering a variety of dining options, and aiding with luggage. The hotel can also consider simplifying the booking process and make it more user-friendly. This can include offering booking assistance over the phone or online chat, providing detailed instructions on how to book, and offering support for any technical difficulties.
- Develop strategies for attracting younger customers: **Cluster 5** represents a younger demographic that makes the second-highest number of direct bookings. The hotel can develop strategies to attract and retain younger customers, such as offering social media promotions or partnering with local attractions or events.
- Focus on **Cluster 6:** This group of customers represents the most valuable customers for the hotel, with the highest number of bookings and nights. The hotel can develop personalized services and loyalty programs to retain and attract more customers in this segment. Also, providing personalized services, such as welcome amenities and personalized room preferences, could be attractive to this group. Offering incentives for direct bookings, such as discounted rates or complimentary services, could also be useful. Additionally, partnering with travel agencies that specialize in serving this demographic could help attract more customers from this cluster.

4. CONCLUSIONS

The CRISP-DM methodology can enable businesses to comprehend their customers better, customize their offerings to meet their needs, and increase revenue in the long term. Using this analysis, the hotel can obtain insights into their customers' preferences and personalize their offerings accordingly. By consistently reviewing and updating their analysis, the hotel can ensure that they deliver exceptional customer service and generate more revenue over time. The aim of segmentation is to facilitate the development of targeted marketing campaigns that can enhance customer acquisition and retention. By gaining a better understanding of their customers, the hotel can create marketing campaigns that are more pertinent and appealing to their intended audience, leading to increased customer loyalty and satisfaction. Data Vision Analytics can assist businesses like Hotel H in constructing high-quality customer segmentation models that can offer valuable insights into their customer base. By utilizing data analytics techniques and tools, we can assist businesses in making well-informed decisions regarding their marketing strategies and enhancing their overall performance.

4.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Based on the data and the process, here are some suggestions for the hotel manager to have a better analysis next time:

4.1.1. Collect clean data - To avoid a messy data set, the hotel can collect clean data from the beginning. They can do this by ensuring that the data is entered correctly at the point of collection, and by using tools such as data validation and verification to ensure data quality.

4.1.2. Regular data cleaning:

To maintain data quality, the hotel should conduct regular data cleaning exercises. This can include removing duplicates, filling in missing data, and correcting errors. Regular data cleaning can help to improve the accuracy of the data and prevent errors in the analysis.

4.1.3. Standardize data formats:

The hotel can standardize the format of their data to ensure that it is consistent across all sources. This can help to reduce errors in the data and make it easier to analyze.

4.1.4. Collect additional data:

To get a more comprehensive view of their customers, the hotel can collect additional data such as customer feedback, social media data, and online reviews. This can help to provide a more complete picture of their customer base and inform future marketing strategies.

By implementing these suggestions, we believe Hotel H will gain a more comprehensive view of their customers. This can help to improve the accuracy of their analysis and inform future marketing strategies.

5. REFERENCES

https://www.optimove.com/resources/learning-center/customer-segmentation
Last accessed on – 1st March 2023

https://www.datascience-pm.com/crisp-dm-2/
Last accessed on – 5th March 2023

https://www.sv-europe.com/crisp-dm-methodology/
Last accessed on – 1st March 2023

https://www.datascience-pm.com/crisp-dm-2/
Last accessed on – 28th Feb 2023

https://openviewpartners.com/blog/customer-segmentation/
Last accessed on – 28th Feb 2023

https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b
Last accessed on – 5th March 2023

https://www.qualtrics.com/uk/experience-management/brand/customer-segmentation/?rid=ip&prevsite=en&newsite=uk&geo=GB&geomatch=uk
Last accessed on – 5th March 2023

https://pestleanalysis.com/market-segmentation-analysis/
Last accessed on – 28th Feb 2023

https://smallbusiness.chron.com/10-important-business-objectives-23686.html
Last accessed on – 1st March 2023

https://www.youtube.com/watch?v=WhUfhOBKjEU
Last accessed on – 27th Feb 2023

https://www.youtube.com/watch?v=QrU5VHLZ1iw
Last accessed on – 27th Feb 2023

https://www.youtube.com/watch?v=Qz7erR3zVUc
Last accessed on – 3rd Mar 2023

6. APPENDIX (LIST OF FIGURES)