MARCH 2023

Hotel Case Study

Business Cases for Data Science

Group H

Tomás Vicente, nº 20221355
Beatriz Carmo, nº 20220685
Lukas Gross, nº 20221363
Karim Miladi, nº 20220720
Tomás Domingos, nº 20221370

NOVA IMS Information Management School

Table of Contents

Table of Figures

1. Executive Summary

To expand any business, it is necessary to gain new customers. Hotel H, located in Lisbon, Portugal, like any other industry, wants to expand its business and make more profit by attracting new clients. The key to that is to analyse and study the hotel's current clients and identify the distinct groups of customers, so that market strategies can be shaped and changed considering the diverse groups' interests and responses to campaigns.

With that in mind, hotel H had already developed a marketing segmentation. However, this original solution can be improved by applying other methods to create the groups of customers and take more variables into consideration.

Following this, and to be able to create new customers' clusters, the data provided was analysed, cleaned, and organized. With those crucial steps done, the data set was finally ready to be process by the clustering algorithm (K-means). The data was also split into two different perspectives (geographical and funnel), so the insights into the clusters would be better and easier to interpret.

To help with interpretation and visualisation, PCA was also applied to the data so its dimensionality would be reduced.

After applying the K-means algorithm to both perspectives and merging them using hierarchical clustering, the solution obtained consisted of five clusters. These clusters were very different from each other and provided fertile soil to develop marketing tactics and targeting strategies.

2. Business Needs and Requirements

A hotel is a commercial establishment that provides lodging, food, and other services in exchange for payment (which is the main goal in every business: high revenue).

As a hotel chain, there are several businesses needs and requirements that are crucial for success:

- High-quality accommodations
- Efficient booking system
- Customer service
- Management
- Sustainability
- Safety and security
- Marketing and branding

In the highly competitive hospitality industry, hotels must operate with constant adaptation to the ever-changing needs of customers and market conditions. Success in the hotel industry depends on prioritizing guest satisfaction while maintaining operational efficiency and profitability. Therefore, it is crucial for hotels to focus on meeting the needs and expectations of guests while considering market trends.

The customer is the most important aspect of the hotel industry as they are the driving force behind the industry's success. Without customers, hotels cannot sustain their operations, even with competent staff and quality accommodations. Therefore, it is essential to maintain current customers while also attracting new ones. To achieve this, it is important to understand the customers' preferences, expectations, and behaviours. The hotel industry can conduct a

thorough customer analysis to answer key questions such as what customers value in a hotel, what they expect, why they come back, and what factors influence their return.

Customer segmentation is one effective way to increase hotel revenue by targeting specific groups of clients with personalized marketing campaigns. However, a segmentation based on only one characteristic is insufficient for a robust segmentation. Therefore, the objective of this project is to develop a better customer segmentation by incorporating more distinctive customer characteristics to personalize marketing campaigns and make a positive impact on the industry. This segmentation will ensure that marketing campaigns are more personalized, resulting in higher customer satisfaction and repeat business.

3. Methodology

3.1. Data Understanding

Prior to performing any modifications on the hotel dataset, it is essential to understand the data by comprehending the meaning of the different variables, their significance to the main problem, and determining if there are any missing values, duplicates, and in which variables they exist.

After understanding the data, the variable ***ID*** was designated as the index of the hotel dataset, resulting in the identification of **111 duplicated rows** and **28 variables**. To obtain additional information about the customer's origin, an extra dataset called 'Regions' was imported, and only the ISO-alpha 3 code column, which corresponds to the Nationality column of the hotel dataset, as well as the Sub-region Name column were retained.

To ensure that the dataset is accurate, missing values were checked, and two variables were discovered to have missing information (***Age*** and ***DocIDHash***) with a total of 5173 missing values. Any unusual values were replaced with missing ones to avoid future issues.

The dataset variables were then classified into three categories: quantitative, qualitative, and binary variables. This grouping enabled the examination of variables distribution through statistical analysis and histograms, resulting in several conclusions such as:

- ***Age***: Most people are around 50 years old.
- ***DaysSinceCreation***: There are four spikes, where more customers were added to the system, compared to the immediately adjacent points in time, with the most customers around 500 days.
- ***AveragedLeadTime***: Most customers have a low lead time. The number of customers gradually decreases with increasing lead time.
- ***BookingsCanceled*** and ***BookingsNoShow***: Vary little customers cancelled a booking or didn’t show.
- Some features might have outliers.

To identify potential outliers, a box plot analysis was performed, revealing the presence of outliers in nine variables within the dataset. Furthermore, a correlation matrix analysis was conducted, which revealed a high correlation between only two variables, namely ***PersonsNights*** and ***RoomNights***.

3.2. Data Pre-processing

3.2.1. Data Cleaning

After conducting an exploratory analysis of the dataset, it became necessary to address any issues and clean the data. The initial step in this process involved removing missing values and duplicated entries.

Given that missing values were not prevalent in the dataset, removing them did not result in any significant loss of information. Additionally, removing duplicated values ensured that the dataset was free of redundancy and improved the accuracy of any subsequent analysis.

When addressing the issue of outliers in the dataset, a set of rules were established to identify and remove observations that deviated significantly from the rest of the data. These rules included:

- Customers with less than 18 years-old or more than 95 were removed, as it is assumed that one cannot book a hotel room under the age of 18 (since in Portugal one is not considered a minor anymore after passing the 18 years mark).
- Clients with ***AverageLeadTime*** above 550 were removed to exclude extreme cases.
- Clients with ***LodgingRevenue*** above 4000 and ***OtherRevenue*** higher than 3400 were removed because they are considered outliers, as these values represent a small portion of the dataset and would cause issues in further analyses.
- Customers with more than 6 ***BookingsCanceled*** and ***BookingsCheckedIn*** above 25 were removed as they represent a small number of customers and are also considered outliers.
- Customers with ***PersonsNights*** and ***RoomNights*** above 50 and 40, respectively, were removed as they also represent outliers in the dataset.

Following the application of these rules, the dataset was significantly filtered, resulting in **93.468%** of its original customers being retained for further analysis.

3.2.2. Feature Engineering

After the data cleaning process, additional features were created to extract as much information as possible from the original dataset variables. These features included:

- **Total_Revenue**: This feature was created by summing ***LodgingRevenue*** with ***Other Revenue***.
- **Service_share**: This feature indicates the proportion of the total revenue spent on other services and was calculated as ***OtherRevenue*** divided by ***Total_Revenue***.
- **Booking_Success_Rate**: This feature measures the percentage of bookings that were actually checked in and was calculated as ***BookingsCheckedIn*** divided by the sum of ***BookingsCheckedIn, BookingsNoShow***, and ***BookingsCancelled***.
- **Preference_Count:** This feature measures the number of preferences that each client has and was calculated as the sum of all the preferences of each client.

Additionally, 'Zombie customers' were removed from the dataset. These were customers who had not had any activity regarding the hotel. After this step, 80.57% of the original data remained for further analysis.

To account for the considerable number of unique values associated with the variable Nationality, a new variable called Region was created to store this information. Region contains nine unique values, which are represented by the following dot-separated country codes:

- FRA/GBR (France and Great Britain)
- Germany
- Western Europe
- South/North America
- Southern Europe
- Portugal
- Northern Europe
- Africa/Asia/Oceania
- Eastern Europe

FRA/GBR and **Germany** were not included in the Europe region category because they represented a considerable proportion of the hotel's clientele. **Portugal** was also excluded from the Europe region. The sparse number of observations from Asia, Africa, and Oceania led to their merging into a single region.

Following these changes, some of the original features were no longer considered relevant and were dropped. This was particularly the case for the ***Nationality, DocIdHash***, and ***NameHash*** variables, which lacked value in the clustering process.

3.2.3. Data Transformation

To apply the K-means clustering method, it was necessary to transform all qualitative features into quantitative ones. This was achieved through the one-hot encoding method, which was applied to variables such as ***Region***, ***MarketSegment***, and ***DistributionChannel***.

To prevent variables with larger scales from being given undue importance in the analysis, it was necessary to scale the data appropriately. Therefore, the **MinMaxScaler** was used to scale the data, ensuring that all variables were on a similar scale.

After completing the necessary data cleaning and feature engineering steps, the dataset is now prepared for clustering analysis. However, since there are diverse types of variables related to various aspects of the customers, two clustering perspectives were developed to gain deeper insights into the clusters:

- **Geo_perspective**: This perspective includes variables such as ***Age, Total_Revenue***, and all the features derived from the original ***Region*** variable using one-hot encoding.
- ***Funnel_perspective***: This perspective includes variables such as ***Age, Total_Revenue***, and all the features derived from the original ***DistributionChannel*** variable using one-hot encoding.

These two perspectives were created to provide a more comprehensive understanding of the customer clusters and their characteristics. With these perspectives in mind, the next step is to apply clustering algorithms to the dataset and evaluate their performance in generating meaningful and useful clusters.

3.3. Modelling

After ensuring the data is ready, the clustering process in both perspectives can begin with confidence. To improve visualization and reduce data dimensionality, **PCA** was applied before the k-means clustering algorithm.

To determine the best number of principal components to retain, two graphics were plotted: a scree plot and one with the explained variance, which were then analysed using the elbow method. This process was repeated when defining the best number of clusters for the K-means algorithm. To gain a better understanding of how the algorithm performed, four additional graphs were plotted: a cluster cardinality plot, a cluster magnitude plot, a cardinality vs. magnitude plot, and a cluster visualization plot.

All the cluster labels created were added to the original data frame to save the information gained during the clustering process.

After performing the k-means algorithm in both perspectives, the resulting views were merged to create a new clustering solution that included all the data in the dataset.

3.3.1. Funnel Perspective

Principal Component Analysis (PCA) was employed on the set of features, and the analysis revealed that the optimal number of principal components to retain is two, as depicted in Figure 3. Consequently, PCA was re-applied, utilizing the ideal number of principal components.

Upon examining the distortion score plot (Figure 4), which is a metric that measures the average distance between data points and their corresponding cluster centroids, it was determined that the optimal number of clusters for the data would be three. The objective of clustering is to minimize this distance or "distortion" in order to achieve a more accurate and meaningful grouping of the data. Subsequently, K-means clustering was performed on the data utilizing three clusters.

Based on our analysis of additional plots (Figure 5 and Figure 7), it is evident that **cluster 2** has substantially fewer observations than **cluster 1**. Additionally, it is apparent that the magnitude of the clusters increases as the cardinality increases, although this relationship is not strictly linear (Figure 6).

The hotel customer segmentation analysis involved accessing the size and distinguishability of the clusters.

Revenue was the key metric used to differentiate between clusters. The objective was to achieve a high inter-cluster distance while maintaining small intra-cluster distances. The analysis of the cluster means of revenue and age revealed that **Cluster 0** had high revenue and low age while **Cluster 1** had medium revenue and high age. **Cluster 2** had low revenue and medium age.

3.3.2. Geographical Perspective

The same series of steps were applied to our second set of features, **Geo_perspective**. The PCA related plots (Figure 8) indicated six principal components, while the distortion score plot suggested that seven clusters were the optimal choice (Figure 9).

Despite following the same method, the relationship between cardinality and magnitude was not as linear as before (Figure 11). Nonetheless, a similar pattern of a small cluster in comparison to the others was observed (Figure 10, Figure 12). Overall, the analysis of the geographical

perspective provided further insights that can help segment hotel customers based on their specific location and travel patterns.

3.3.3. Merged Perspectives

The two different perspectives were then merged using the hierarchical clustering method. A dendrogram was plotted to evaluate the best number of clusters to use in the final solution (Figure 13). The dendrogram indicated four clusters to be the best number. However, the number of clusters used was five since it provided a better cluster differentiation.

3.4. Evaluation

To better access the goodness of the final model, a decision tree classifier was used (Figure 14). The results showed that **72.92%** of new customers were correctly predicted. Although the percentage isn't as high as it would be expected, these findings imply that the clustering method used in this research could be helpful for segmenting customers and predicting their behaviour.

Furthermore, it was discovered that certain features had a greater impact on the cluster to which the customer was assigned. Values in the features ***Region FRA/GBR***, ***Geo Clus***, ***Region Portugal***, and ***DistributionChannel Travel Agent/Operato*r** were found to have the greatest effect on customer clustering. These findings highlight the importance of taking these characteristics into account when segmenting customers in the hotel business.

The final solution consists of five clusters:

Cluster 0 – National Workers:

- Lowest Revenue generated out of all others.
- Mainly Portuguese clients, who are not present on other clusters.
- Segment that books the closest to arrival.
- The lowest amount of People per Night, same for Rooms per Night, although the difference is more marginal.
- Considerable spending on Hotel Services.
- Corporate Funnel is the most appropriate funnel for targeting.
- Most attendance post booking.
- Clients in this segment have a very low special request count.

Cluster 1 – European Spenders:

- Considerably high Revenue generated.
- Mainly French, British, and Southern Europe clients.
- The highest amount of People per Night, 2nd highest Rooms per Night.
- Considerable spending on Hotel Services.
- High attendance post booking.
- Significantly high special request count.
- Travel agencies and other operators are the most appropriate funnel for targeting.
- Direct funnelling could be explored, due to considerable frequency with a discount for booking on arrival, keeping higher margins.

Cluster 2 – Traditional Travelers:

- Oldest segment out of all.
- Segment that books the farthest to arrival.
- Considerably Low Revenue generated.
- German clients, that are not present in other segments, and North and South American.
- Highest spend on Services, these customers should be targeted with tours, restaurant deals, massages, SPAs, and all existing service ranges available.
- Highest preference count.
- Travel Agencies should be the only ones considered for marketing campaigns.

Cluster 3 – North American Clients:

- Highest spend on rooms, and lowest use of services, compared to room spend.
- Most Valuable cluster in Total revenue Count.
- Exclusively North American Clients.
- Use of booking on Arrival with the highest frequency.
- Marketing campaigns should be prioritised towards this cluster.

Cluster 4 – Not so Simple Customers:

- Significantly high Revenue, Persons, and Rooms Booked per Night.
- Low demand for Hotel Services.
- Significantly high special request count.
- Segment with low regional differentiation, englobing Asian, African, Oceanic, and European clients.
- Highest use of booking on Arrival, and a significantly high count of Travel Agency Bookings.

4. Results Evaluation

When segmenting the hotel customers, it is possible to observe what the clients inside each cluster have in common and from that develop marketing strategies to be able to attract new customers with the same patterns and to make the current customers return. Having a good customer base is key to have higher revenue and use part of that income to improve existing services the clients might be interested in.

Because of the importance customers have in this business, it is essential to know them and understand their interests, needs and expectations. The cluster solution provided by the implementation of all the methods mentioned above contributes to this understanding as the final solution divided customers into very different groups (Figure 17).

5. Deployment And Maintenance Plans

Hotels must implement targeted marketing strategies based on the customer segment to which they cater. Personalized offers and promotions can help to make a good impact and develop

customer loyalty. Hotels can benefit from positive word-of-mouth recommendations and increased repeat bookings by making customers feel valued and appreciated.

Keeping this in mind, each cluster should have a distinct marketing approach. Starting with **Cluster 0**. This cluster has mostly Portuguese customers, who generate the lowest revenue for the hotel. It is crucial to keep strong relationships with national clients, particularly in the event of future lockdowns. It is also important to note that most of the customers in this cluster reached to the hotel trough business trips. Therefore, it may be beneficial to reach out to the companies for which they work and offer appealing deals for their workers to stay at the hotel. Furthermore, if these clients are satisfied with their hotel experience, there may be a chance to tap into their leisure market.

For the customers in **Cluster 0** the hotel can implement loyalty programs to promote repeat visits and boost customer loyalty in addition to targeting the businesses that these customers work for and offering them attractive deals. Offering discounts or special benefits to customers who stay at the hotel frequently can be an effective way to encourage them to return for future stays. Furthermore, hotels can work with local businesses and attractions to offer package deals to guests, making the hotel more attractive for both business and leisure travel. Hotels in Cluster 0 can increase revenue and strengthen customer relationships by adopting these strategies.

Cluster 1 differs from the previous segment in that it is primarily made up of European clients who are willing to spend a substantial amount of money on hotel services. These characteristics must be highlighted and promoted in the marketing plan. Offering bundled packages that include eatery, spa, and other services may be profitable. Because this group primarily books through travel agencies, collaborating with these agencies and airlines to offer all-inclusive packages would be a viable strategy. Investing in online advertising campaigns and social media marketing could also help you reach a larger audience and boost bookings.

Another effective approach would be to highlight the hotel's luxurious offerings and attract new customers via social media. Because Instagram and Facebook are especially popular among European tourists, hotels should update their profiles on a regular basis with high-quality pictures and engaging content about their amenities, services, and location.

Cluster 2 is primarily made up of German clients who are older and more traditional in their travel habits. Despite their low revenue contribution to hotels, they have a high number of special requests and are willing to spend money on hotel services. Hotels should consider offering exclusive packages that cater to their preferences in order to tap into this market. Group tours, personalized rooms, and restaurant services could all fall under this category. It is worth mentioning that clients in this segment only book through travel agencies, so collaborations with such agencies could be beneficial as well.

The North American clients, however, are mostly present in **Cluster 3**. These clients are the ones who spend the most in total, despite not spending much on extra hotel services. They also book the closest to arrival, so the hotel should facilitate this process since the customers of this segment are extremely valuable and most likely to react to special offers immediately. Deals with travel agencies shouldn’t be taken into consideration since this cluster is more focused on direct booking. Because of this, discounts for on-site bookings can be an effective marketing strategy for clients in this cluster. This can help the hotel maintain higher profit margins than if it relied on third-party booking systems. Offering a discount for booking directly with the hotel can encourage customers to use the hotel's own booking system while also building trust with customers who value the savings.

Customers in **Cluster 4** have a high income, a high number of room bookings per night, and a low demand for hotel services. They also have a high volume of special requests and originate from various regions, making it difficult to reach them through traditional channels such as travel agencies. This segment is primarily made up of event-based groups, so it is critical to target them before they come.

Offering customized event packages, conference rooms, and catering services can be an effective tactic for attracting more bookings from this cluster. Individualized services can also be provided to satisfy their specific requirements. It is critical to remember that these customers have elevated expectations, and providing exceptional services can result in favourable feedback and repeat business.

Direct marketing to this group can also be a successful strategy because it provides for personalized communication and offers. Keeping note of their preferences and previous stays can also help tailor offers and services to their liking. Overall, emphasizing personalized services and direct marketing can result in increased revenue and client loyalty for this high-revenue cluster.

Hotel H can explore alternative marketing strategies to reach customers in this segment. One method is to use targeted internet advertising. Because customers in this segment have little regional differentiation, hotels can reach out to prospective customers worldwide by using targeted online ads on social media platforms. Referral programs can also be an effective marketing strategy for reaching out to consumers in this demographic. Hotels can encourage current customers to recommend their friends and family by providing rewards or discounts to both the referrer and the referred customer.

It is essential to keep in mind that these strategies might become obsolete over time. Marketing strategies must evolve with the clients and adapt to them, so does the staff of the hotel. Staff plays a crucial role in a hotel, therefore, to keep up with future changes, there might be the need to educate current staff members and even hire new employees.

One must also know that no solution can be effective forever, no matter how great it has proven to be in the present moment. To evaluate and segment customers regularly is important. People change, their habits, likings and preferences change, so a hotel must change to in order to meet the clients' new interests and expectations.

6. Conclusions

After analysing the final clustering solution and all the methods used, one can conclude that K-means clustering is not always the best algorithm since it only provides circular clusters. However, the results were still very satisfactory and were able to highlight the differences between each one of the five final segments.

In addition to that, the outcome proved to be good and also provided evidence that personalized marketing strategies are possible to develop and should be adopted for different customer segments based on their preferences, behaviour, and characteristics. Customer origin, spending habits, booking channels, and preferences for specific hotel services should all be considered in the marketing strategy.

It is also important to emphasize that other customer characteristics can be helpful when developing this tactics: gender, number of children, job, etc. However, and despite not having access to this kind of information, the clusters found were interesting and close to what was needed to organize marketing strategies.

7. Appendix

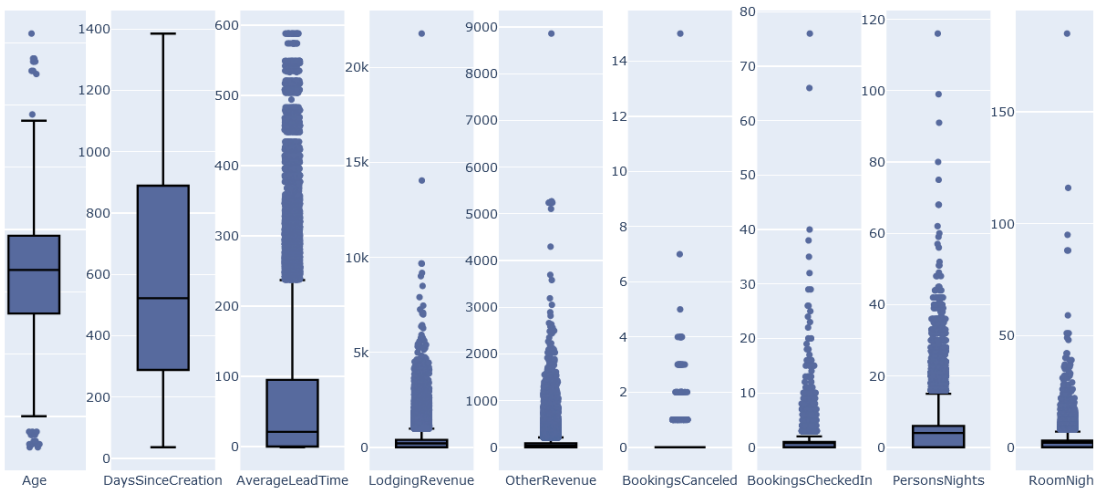

Figure 2 - Boxplots for quantitative variables

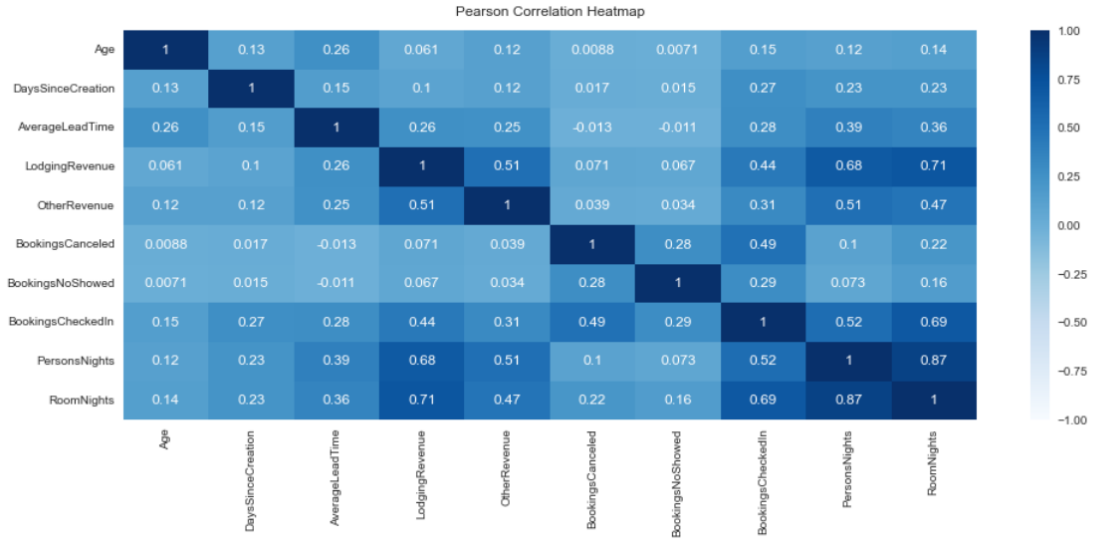

Figure 1 - Pearson Correlation Heatmap

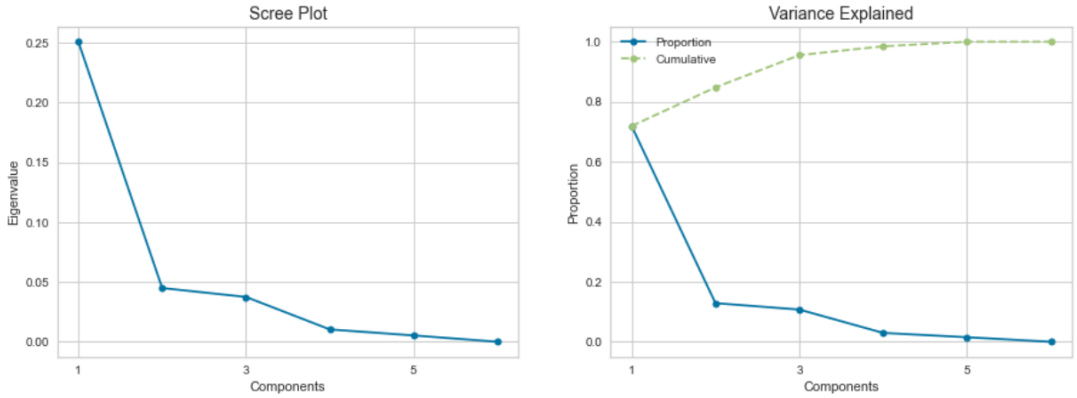

Figure 3 - PCA related plots (Funnel Perspective)

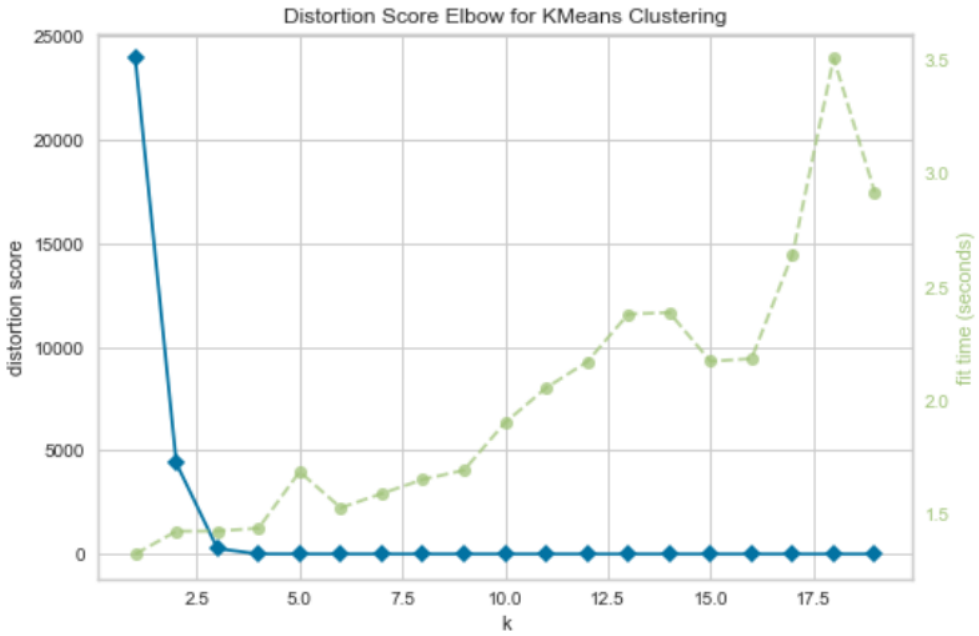

Figure 4 - Distortion score plot (Funnel Perspective)

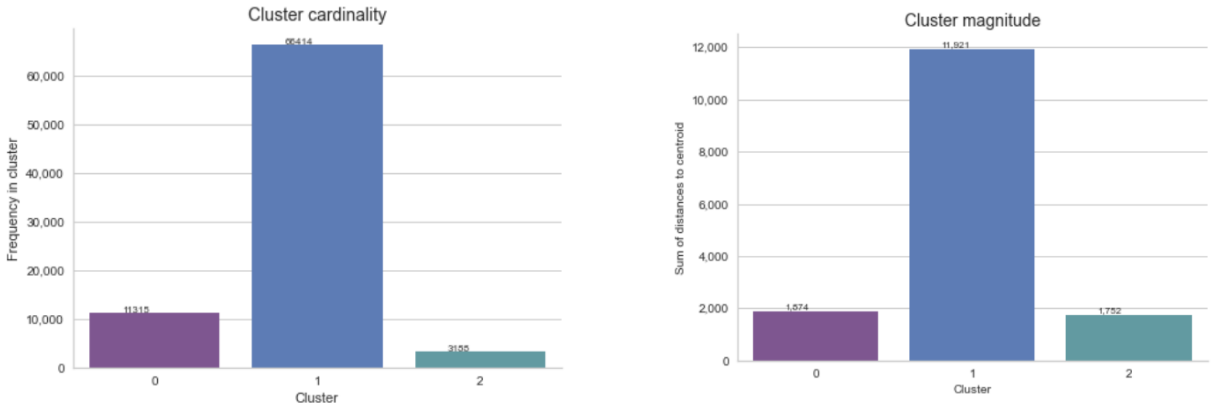

Figure 5 - Cluster cardinality and magnitude plots (Funnel perspective)

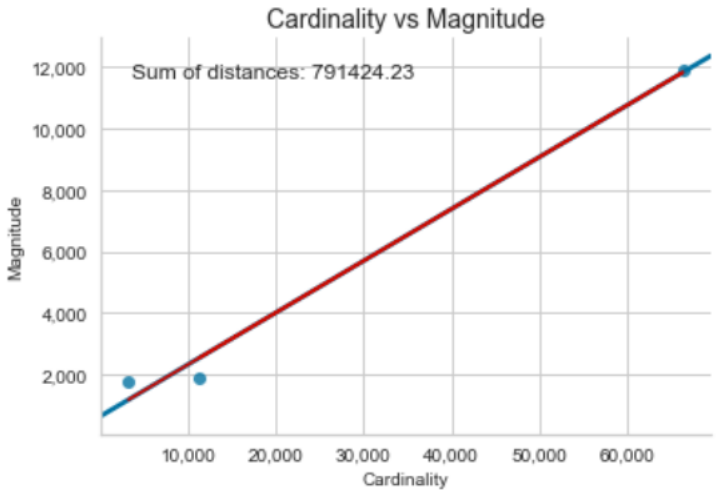

Figure 6 - Cardinality vs. Magnitude plot (Funnel perspective)

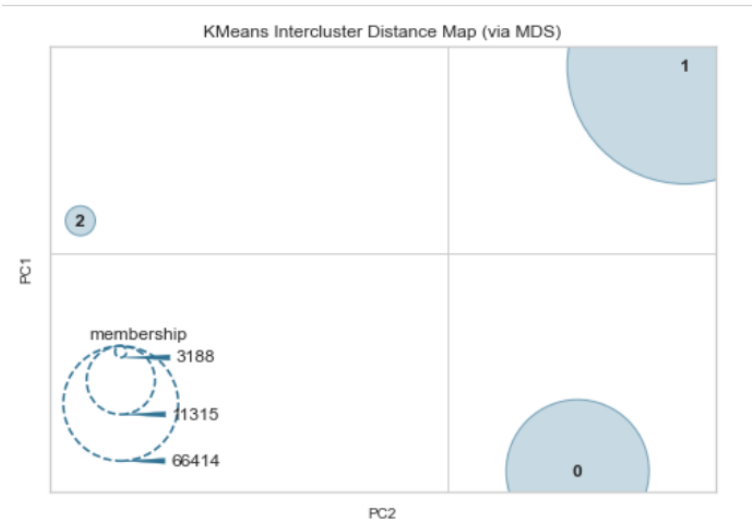

Figure 7 - Cluster visualisation (Funnel perspective)

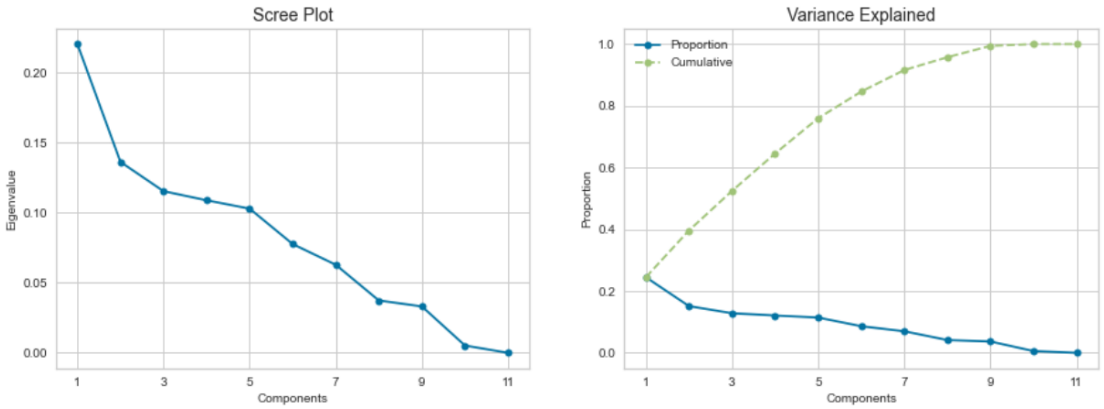

Figure 8 - PCA related plots (Geo. perspective)

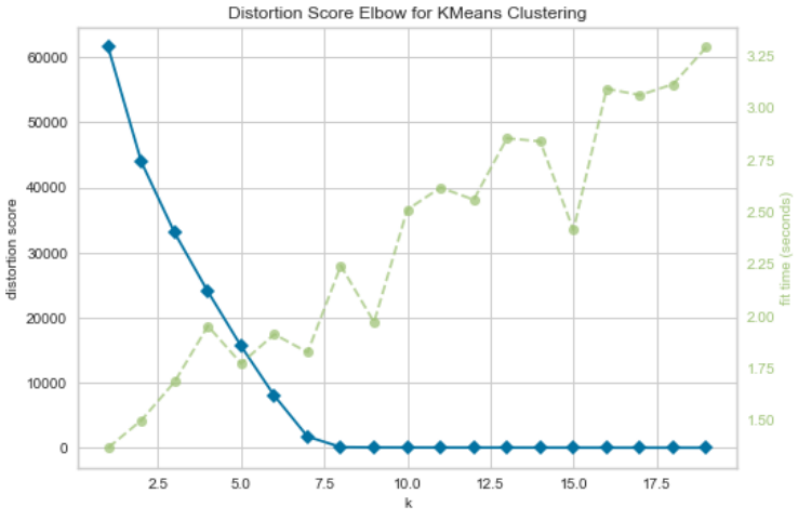

Figure 9 - Distortion score plot (Geo. perspective)

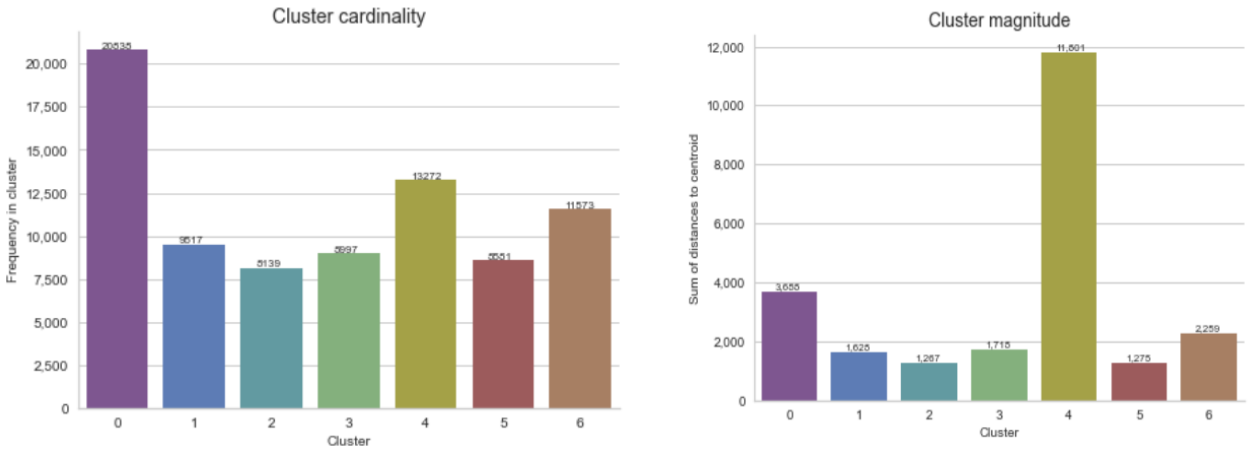

Figure 10 - Cardinality and magnitude plots (Geo. perspective)

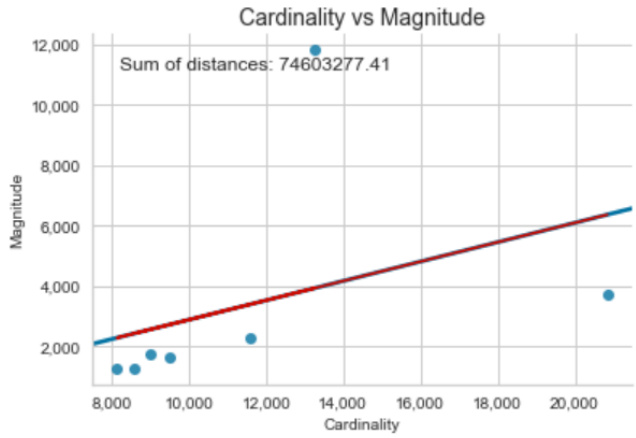

Figure 11 - Cardinality vs. Magnitude plot (Geo. perspective)

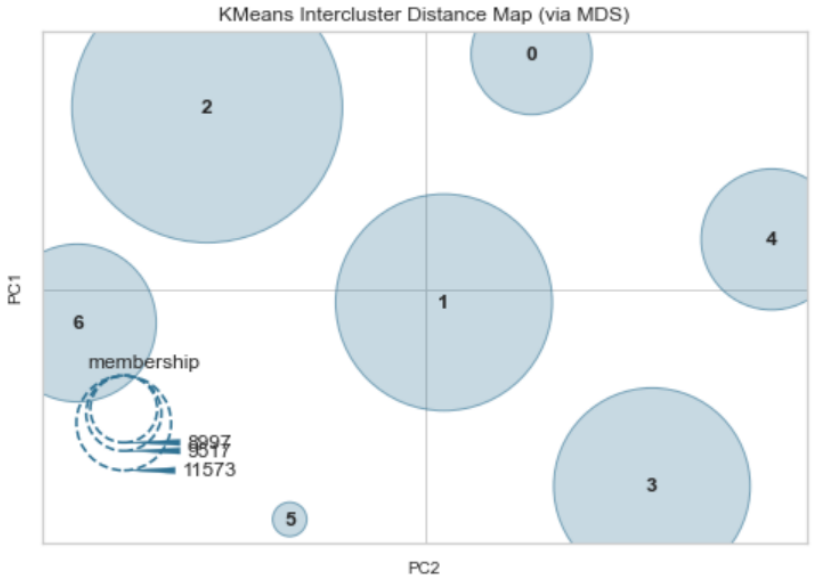

Figure 12 - Clusters visualisation (Geo. perspective)

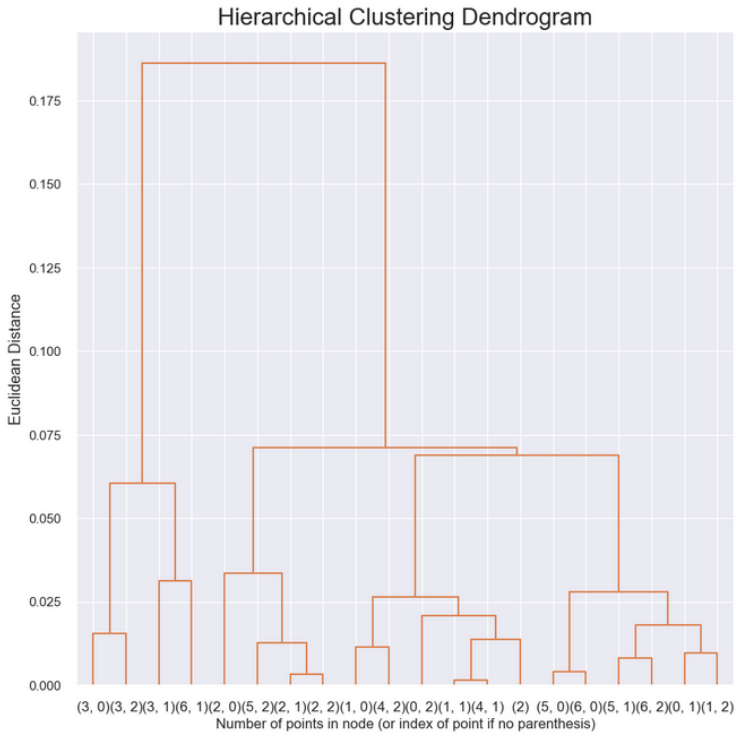

Figure 13 – Dendrogram

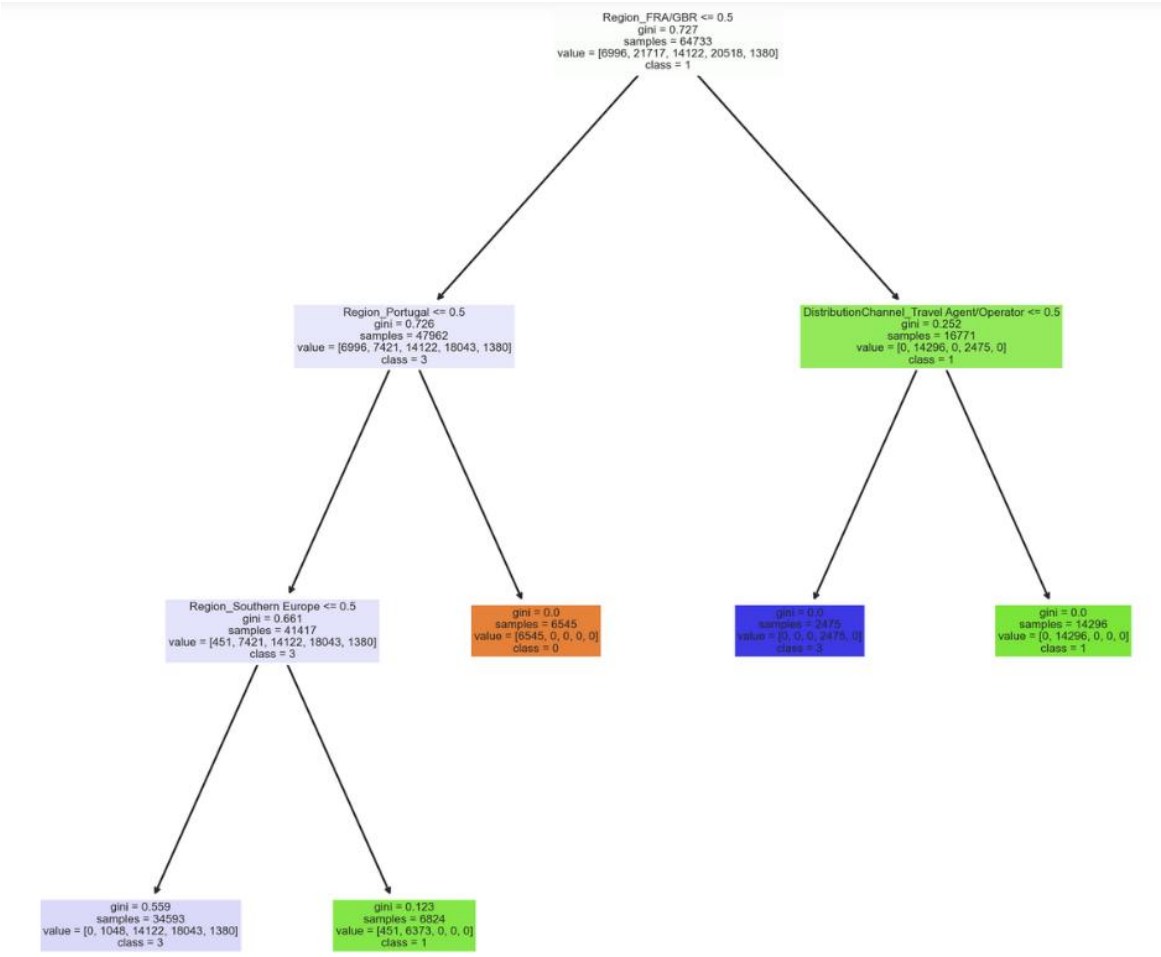

Figure 14 - Decision tree

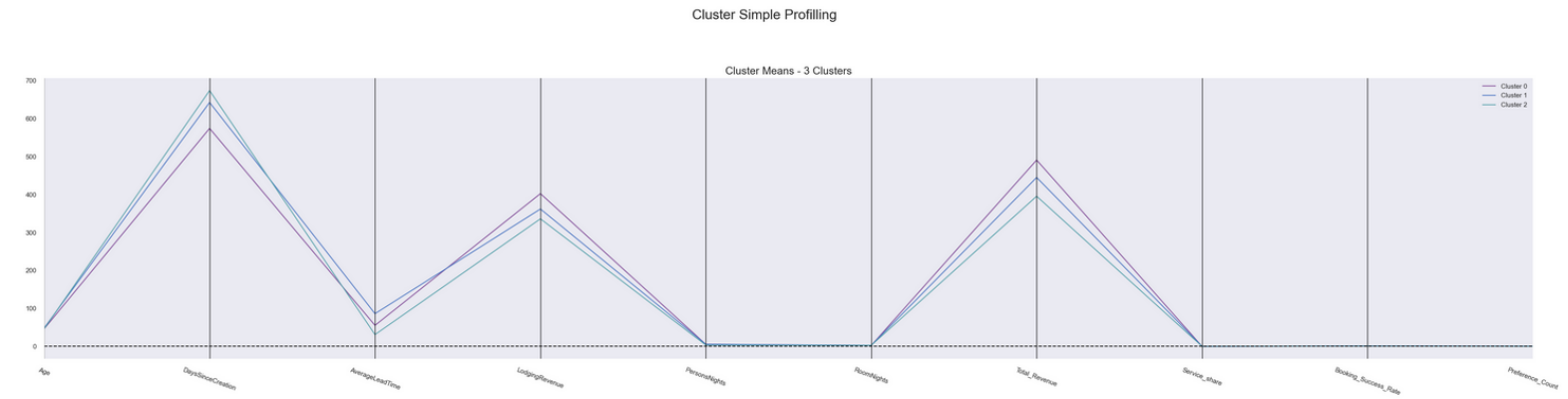

Figure 15 - Cluster profiling (Funnel perspective)

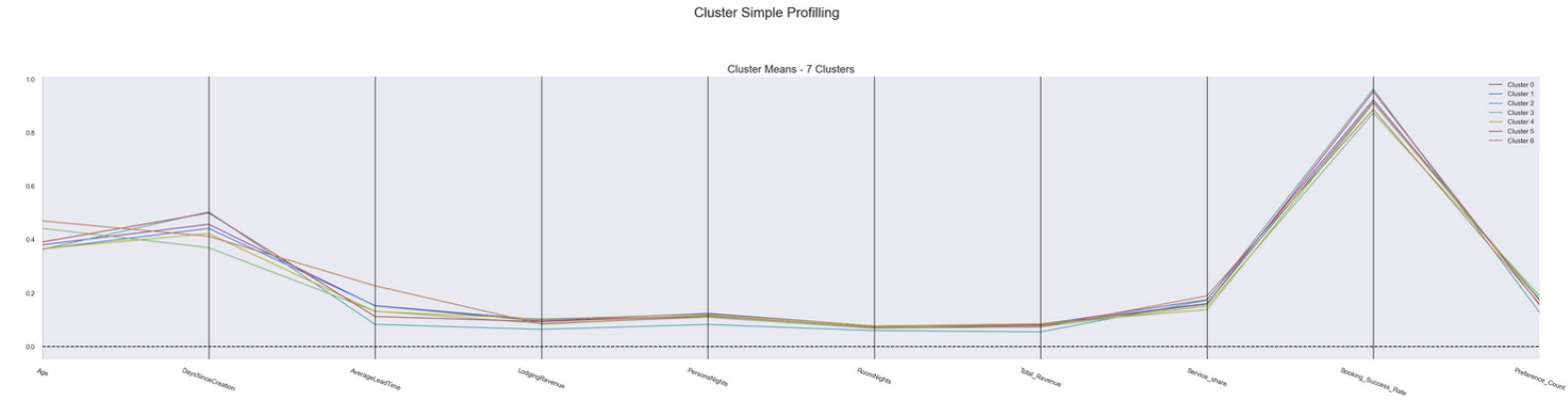

Figure 16 - Cluster profiling (Geo. perspective)

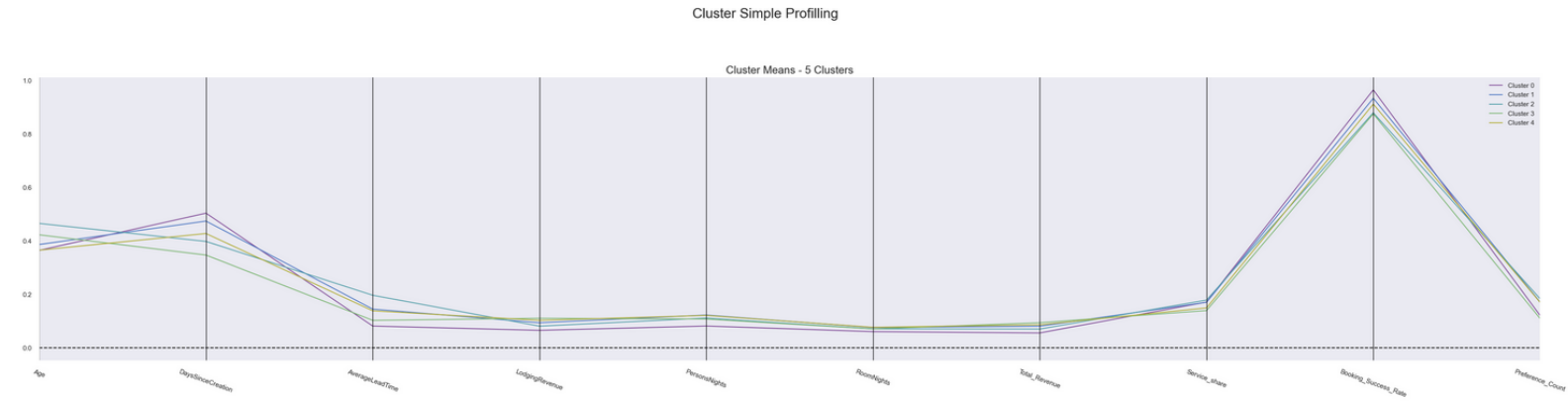

Figure 17 - Cluster profiling (final solution)