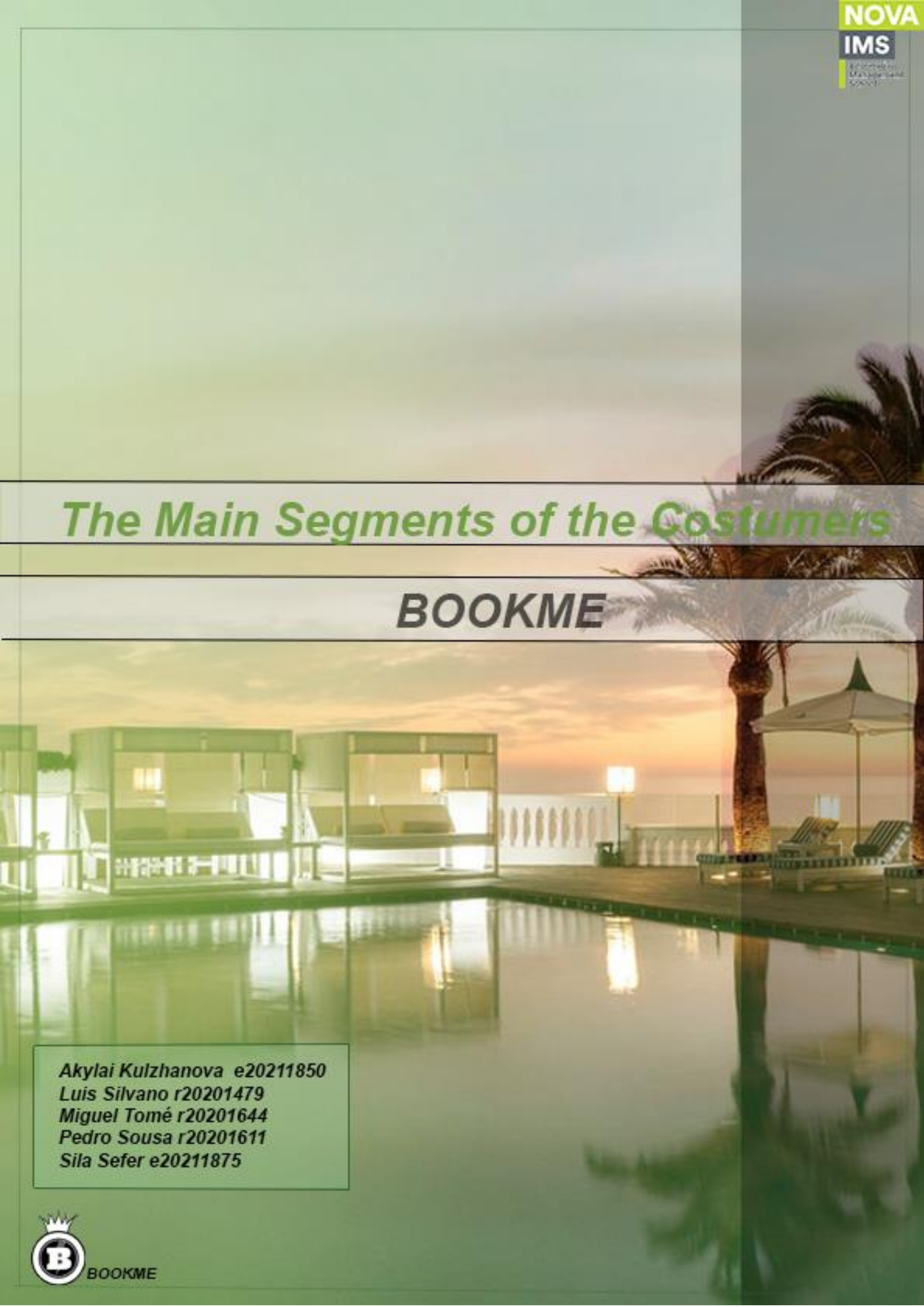


The Main Segments of the Costumers

BOOKME

Akylai Kulzhanova e20211850
Luis Silvano r20201479
Miguel Tomé r20201644
Pedro Sousa r20201611
Sila Sefer e20211875

BOOKME

Index

1. Introduction and Methodology

1.1 Introduction

The company BookMe is a firm that operates in the hospitality sector. It interacts with their clients through their international website. The main objective of this company is to grant quality and conditions in its services (Comfort, Reception Schedule, Food and Drink Available, Accommodation Location, Wi-Fi Service, Accommodation Amenities, Staff, Online Booking, Price Quality Relationship, Room Space, Check-in, Check-Out, Cleanliness and Bar Service) to future traveller's. The company recruited our team to ensure the marketing department spends their annual budget consciously, with our analysis focusing on **satisfaction and customer characteristics**, as they create more efficient campaigns and don't waste their money unwisely.

1.2 Methodology

To support the Marketing Department identifying the satisfaction of the customers and their characteristics, our team firstly explored the given dataset. We analysed the client characteristics, their levels of satisfaction and the information about each service. We observed the training set, to understand which values need to be handled and explored. Secondly, we started pre-processing the raw data. Then we assign the perspectives for the satisfaction in **"satisf_data"(**Table 3: Satisfaction Dataset**)** and for the characteristics in **"charac_data"(**Table 2: Characteristic Dataset**)** of the customers and we implemented different models in our project to test their performance, like **K-Means Clustering (with Elbow Method, Dendrogram Method, Silhouette Method** and **PCA** as a way to see the number of clusters), **SOMs (Self-Organizing Maps)** and as addition, **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)** with the purpose to recognize the **customers' satisfaction and their characteristics,** to understand the profile of the costumers for the Department of Marketing get a more efficient campaigns and use properly their annual budget.

2. Exploration

In the Exploration Criteria, our aim was to explore which were the initial values we had and which attributes we would review from the training dataset (the first look of the dataset). We acknowledged that when we applied the code, we obtained 15589 clients and 22 attributes (Figure 1: Datatypes of our Variables). After recognizing the structure of our dataset, our team decided to read the types of each attribute, noticing some categorical data and numeric characters with decimals that would need to be treated, so we can use our models. Next, we compared the frequency of missing values from the training data set. We noticed that Year Birth has 195 missing values (or 1.25%) so we must remove it (Figure 2: Percentage of Missing Values each Variable). Then, we evaluated the customer's characteristics, so we decided to plot the graphs for the variables ("Clients Churn", "Clients Longevity"," Clients Year Birth", "Clients Type of Travel", "Clients Room Type" and "Clients Reward Points"). With this, we concluded that in **Clients Churn (**Graph 9: Number of Clients Churn Characteristics (Exploration)**)** most of the clients tend to not churn when they are booking one reservation; In **Clients Longevity (**Graph 8: Clients Longevity Characteristics (Exploration)**),** most of the clients are satisfied with BookMe; In **Clients Year Birth (**Graph 5: Clients Year of Birth Characteristics (Exploration)**,** the years that are predominant are 1970 to 2000, which indicates most of the clients are older than 20 years old and younger than 50 years old; In **Clients Type of Travel (**Graph 4: Clients Type of Travel Characteristics (Exploration)**),** most of the travellers prefer to do business travels than leisure travels; In **Clients Room Type**, the clients prefer to reserve single rooms less than to reserve suite rooms; In **Clients Reward Points (**Graph 6: Clients Reward Points Characteristics (Exploration)**)**, most of the clients classify 5000 points to BookMe loyalty. After doing a simple analysis of the variables ("Clients Churn", "Clients Longevity"," Clients Year Birth", "Clients Type of Travel", "Clients Room Type" and "Clients Reward Points") we decided to do further explorations on the variables (Longevity, Type of Travel, Customer Room Type and the Clients Room Type). According to the **Longevity by Churn graph (**Graph 8: Clients Longevity Characteristics (Exploration)**)**, we can conclude that we have values in Churn and in no Churn that were selected into "s". We decided to arrange it because there was an error, by replacing the function to turn the values (s) into (yes). We can conclude as well that when the clients are loyal, their longevity in not Churns is higher than churns longevity. According to the **Longevity Year Birth graph**, **Longevity by Reward Points graph**, **Churn by Type of Travel graph**, the previous speculations that we had remained all the same. According to the **Year Birth by Type of Travel graph** **(**Graph 10: Customer Type of Travel Characteristics (Exploration)**),** all ages between 10 and 70 years old prefer leisure travel rather than business-type travel. According to the **Reward Points by Type of Travel graph** **(**Graph 10: Customer Type of Travel Characteristics (Exploration)**),,** the majority that evaluates the Reward Points is from the business type of travel. According to the **Churn by Room Type** **(**Graph 11: Customer Room Type Characteristics (Exploration)**)**, we concluded that the travellers tend to churn more when the room type is double and tend to no-churn when the room type is single. According to the **Year Birth by Room Type** **(**Graph 11: Customer Room Type Characteristics (Exploration)**,** we terminated that all travellers ages prefer to stay in the single and double room type. According to the **Reward Points by Room Type graph** **(**Graph 11: Customer Room Type Characteristics (Exploration)**,** we concluded all travellers that gave reward points for loyalty are mostly from the single and double room types. According to the **Client´s Room Type by Type of Travel graph** **(**Graph 13: Room Type by Type of Travel Characteristics (Exploration)**),** we concluded that most travellers that stay in suites and single rooms are there for business travel and most travellers that stay in double rooms are there for leisure travel. To further examine and evaluate the customer satisfaction, we decide to plot the graphs for all variables. According to **the Comfort Satisfaction** **(**Graph 20: Comfort Satisfaction (Exploration)**)**, the clients are not very satisfied with the comfort as most travellers attributed 2, 3 or 4 ratings. According to **the Reception Schedule Satisfaction** **(**Graph 24: Reception Schedule Satisfaction (Exploration)**),** the clients are very satisfied with the reception schedule as most of the travellers attributed 4 or 5 ratings. According to **the Food and Drink Satisfaction** **(**Graph 19: Food and Drink Satisfaction (Exploration)**),** the clients are not very satisfied with the food and drinks as most travellers attributed 2, 3 or 4 ratings. According to **the Location Satisfaction** **(**Graph 22: Location Satisfaction (Exploration)**),** most of the clients classifies 3, so they are reasonably satisfied with their location. According to **the Wi-Fi, Amenities, Staff, Online Booking, Price Quality, Room Space, Check-out, Cleanliness, Bar Service Satisfaction** the travellers are very satisfied because most of them attributed in each satisfaction 4 of rating. This means, the clients have easier access to Wi-Fi, they have a higher chance of having amenities in their reservations, they are well cared by all staff and have a good relationship with them, the online booking works well, the size of the rooms fulfils the needs of the travellers, they have a good bar service, and the hotels are always well treated and neat. At last, we have **the Check-In Satisfaction** **(**Graph 15:Check-In Satisfaction (Exploration)**),** the travellers are reasonably satisfied as most travellers classify it as 3 or 4 ratings.

3. Pre-Processing

In the Pre-Processing Criteria, the aim was to take care, clean, transform, and reduce errors on the raw data of the respective datasets. We decided to treat the outliers and missing values instead of deleting them, we changed the data types so we could do extensive analysis with different models that do not allow non-numerical data, we also reviewed the coherence of the data, we performed a feature selection with a correlation test and lastly, we scaled the data.

Starting with the pre-processing itself, we made a copy of our initial table of train dataset, we index our variable 'Cust_ID', to save the information about the client and to have a better view of our dataset. Our next move was to **treat missing values**, very briefly we used the knn imputer, in order to not delete this precious information (more of this process in the final pages of the report).

Then we had to **check the coherence** of our dataset, as noticed before, we had some problems with the variables **'Wi-Fi Satisfaction'** and **'Longevity'.** In the **'Wi-Fi Satisfaction'**, we had ratings equal to six, and after checking the business needs of the company this did not make sense, so, after some consideration we considered this rate as five misplaced and used a **replace** function to do this transformation. In the **'Longevity'**, we used the same thinking process as before, to arrange the values of this variable we used the **replace** function to change the values (y) into (yes). Also, in data coherence, we noted that the minimum age was 7 years old, and the maximum age was 85 years old, we did not consider this as a problem as we consider this as being children of clients who filled in the satisfaction questionnaire.

Our next move was to **create two four variables.** The first one defines if a customer is a female or not and the second one the year of birth variable. For the variable **'Female'** we created a restriction on the first 3 letters **('Mr. or Ms.')** **(**Graph 29;

Customers Gender) from each name in our dataset and saved this as variable. Then, using the dummy variable, we defined that 'Mr.' is zero and 'Ms.' is one, after this we renamed it to **'Female'** and deleted the **'Name'** variable. After a small analysis we saw a small difference and noticed a few more women as clients. The variable **'Age'** was simpler, we created a new variable, already called Age where we subtracted the current year by the year of birth of each customer and then we transformed this float variable into an integer one. Also, with the **'Age'** variable, we decided to be more specific about the costumer's ages, so we split them into three groups of ages (Children, Adults and Seniors). **Children** when **17 years old or younger**, **Adults** with ages of **17 and 60 years old,** and **Seniors** when the customers are **60 years old or older**. At last, we measured the **loyalty (No Loyalty, Low Loyalty, and High Loyalty)** of each customer using the system of Reward Points, the Longevity and Churn variable.

The next step was to **transform categorical data into numerical**. We treated the variables '**Type of Travel', 'Type of Room', 'Churn', 'Longevity'**, and **'Loyalty'** on the first two variables we separated them and used the dummy variables function, creating new columns which we renamed (a leisure travel column which said 1 if a customer did come as leisure, 0 otherwise; a single room said 1 if the room chosen was single, a suite variable which said 1 if the room was a suite and 0 otherwise), for the other two variables we transformed churn and no-churn into 1 and 0, respectively, and yes/no into 1/0.

Then for all variables, except the Reward Points variable we **changed their data type to** "integer8" to be easier to analyse the data. We let the Reward Points with the datatype "integer16" because it has greater values.

We decided to **treat the outliers** from the variable Reward Points as seen before, and since we only had two variables where it was possible to see missing values, it was relatively easy to analyse. We decided that we would not remove outliers as we kept it, we applied multiple methods that will be shown in the in the final,

After handling the outliers, the next step was **feature selection**. The first method we used was simply checking the range, to see if some independent variable had a range equal to zero or close. This was not the case. We checked both Pearson (Graph 37: Pearson Method) and Spearman (Graph 36: Spearman Method) correlation which gave us similar results, with these results we looked for high correlated variables (we opted for a threshold of >70%) and therefore there was only 1 high correlated variable (No Loyalty with the Churns), so we decided to eliminate the variable **'Non_Loyalty'**.

We scaled the data with a robust-scaler scaler and we changed all datatypes from all variables to "float16" to achieve the final table to perform the models.s

To finish the pre-processing of the dataset and prepare for the modelling, **we defined the perspectives**, by dropping the cluster data the variables **'Suite'**, **'Male'**, **'Low_Loyalty'**, **'Children'**, **'Adult'**, **'Seniors'**, **'High_Loyalty',** since this variable did not affect positively our clusters, but later we will use some to create visualizations to show off the final clusters. After organising and cleaning the S_data, we created 2 tables, one is the *satisf_data* (Table 3: Satisfaction Dataset) with the satisfaction variables of the main dataset *(Comfort, ReceptionSchedule, FoodDrink, Location, Wifi, Amenities, Staff, OnlineBooking, PriceQuality, RoomSpace, CheckOut, CheckIn, Cleanliness, and BarService)* and the other is the *charac_data* (Table 2: Characteristic Dataset) with the characteristics variables *(Churn, Longevity, Age, Leisure_Travel, and Single Room).*

4. Modelling

After treating all the data, it was time to access our dataset using different methods to see which one of them would be the most accurate to use in our market campaign using unsupervised learning. By using a cluster base system, we are trying to maximize inter-cluster distance and minimize intra-cluster distance, to create separation by similarity and get better results with our data.

4.1 K-Means for Clustering

To decide the number of clusters that we used in our modelling, we used different approaches such as the Elbow Method, the Dendrogram, the Silhouette method, and PCA (explain later) to compare the different results and fully understand how many clusters were better to use

For the **Elbow Method** **(**Graph 47: Elbow Method (Characteristics and Satisfaction Customers), we analysed both the *(Clients Satisfaction Inertia)* and *(Clients Characteristics Inertia)* and as the number of clusters increases, the cluster members are closer to one another, so, the best solution was regarding (*Clients Satisfaction Inertia)* and *(Clients Characteristics Inertia)* between 2 and 3 clusters.

For the **Dendrogram**, we separated the *(Clients Satisfaction Inertia – Satisfaction Dendrogram)* and *(Clients Characteristics Inertia- Characteristics Dendrogram)* to get a better interpretation of the results. In each of the categories, we subcategorize it by using a complete linkage, average linkage, simple linkage and Ward linkage. We found that the best results for *(Clients Satisfaction Inertia)* would be 2 clusters and for *(Clients Characteristics Inertia)* 3 clusters.

Finally, to complete our analysis to find out how many clusters we should use in our model, we used the **Silhouette Method**. By assuming that the data has already been clustered into k clusters by a clustering technique, for each data point, the cluster assigned to the **ith** data point determined the average Silhouette for each value of k and for the value of k, which has the maximum value. The outcome of this approach led us to point out that *(Clients Satisfaction Inertia)* had 2 clusters and *(Clients Characteristics Inertia)* had 3 clusters.

Now that we have the number of clusters for *(Clients Satisfaction Inertia)* and *(Clients Characteristics Inertia)*, we implemented the **KNN-Means Clustering** model, and explored the results such as each characteristic Cluster, a description of measures and their values. Finally, we plotted the results to get a better visualisation of *(Clients Satisfaction Inertia)* and *(Clients Characteristics Inertia)* attributes.

We decide to implement the **K-Means Clustering** because it gave us more precise results. In the **Clients Characteristics**, we used only 3 clusters, after that, we fitted the model in our data, checked the centroids of the clusters, and created one column to identify the cluster that each customer belongs to. In **Clients Satisfaction**, we used only 2 clusters, after that, we fitted the model in our data, checked the centroids of the clusters, and created one column to identify the cluster that each customer belongs to.

Now, observing the Client's Satisfaction Clustering we evaluate all their variables *(Comfort, ReceptionSchedule, FoodDrink, Location, Wifi, Amenities, Staff, OnlineBooking, PriceQuality, RoomSpace, CheckOut, CheckIn, Cleanliness, and BarService), and after, we evaluated the Clients Characteristics Clustering (Churn, Longevity, Age, Leisure_Travel, Single_Room).*

Started by evaluating the **Satisfaction Cluster** in **Comfort by Cluster** **(**Graph 53: Comfort Satisfaction (KNN Modelling Clustering)**,** we gathered that most of the ratings from cluster zero, were 4 and the least chosen 0. Most of the customers from cluster 0, are comfortable with their room options; on cluster 1, most of the ratings given by the customers were 3 and the least given was the 0 rating. From this graph, the customers do not have the same level of satisfaction as the customers of cluster 0. They are less comfortable with their room options.

In the **Reception Schedule by Cluster** **(**Graph 48: Reception Schedule Satisfaction (KNN Modelling Clustering) in cluster 0, the customers mostly rate 4 or 5. The customers are satisfied with the availability and attendance of the employees. In cluster 1, most of the customers rate 1, 2, 3, or 4, meaning the customers are not so happy with the reception schedule, since we do not have a precise rating that allows taking positive conclusions. In the **Food & Drink by Cluster** **(**Graph 52: Food & Drink Satisfaction (KNN Modelling Clustering)**),** in cluster 0, most of the customers ratings 4, the least chosen rating was 0, so most of the customers are satisfied. Now, in cluster 1 we understood that most of the rates from the customers were 3 and 4 and the least chosen was 0. The customers of cluster 1 are reasonably satisfied with the food and drink available. In the **Location by Cluster** **(**Graph 51: Location Satisfaction (KNN Modelling Clustering)**,** we gathered that in cluster 0 and in 1, the customers mostly rate 3, the least chosen rating was 5, so most of the customers are reasonably satisfied. In **Wi-Fi by Cluster** **(**Graph 63: Wifi Satisfaction (KNN Modelling Clustering)**.** in cluster 0, most of the ratings given by the customers were 4 and 5, the least given rating was 0. Most of the customers are satisfied with the Wi-Fi services. In cluster 1 many of the ratings from the customers were 2 and 3 and the least chosen was the 0 rating. From cluster 1 are not very satisfied with their conditions, the Wi-Fi services. In the **Amenities by Cluster** **(**Graph 58: Amenities Satisfaction (KNN Modelling Clustering)**)**, we gathered that in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 0 rating. Most customers are satisfied with the amenities in their accommodations. In cluster 1 most of the ratings given by the customers were 3 and 4 and the least given was the 0-rating meaning. In the **Staff by Cluster** **(**Graph 62: Staff Satisfaction (KNN Modelling Clustering)**)**, in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 0 rating, most of the customers are satisfied with the Staff. In cluster 1 we gathered that most of the ratings from the customers were 3 and 4 and the least chosen was the rating 5. In the **Online Booking by Cluster** **(**Graph 50: Online Booking Satisfaction (KNN Modelling Clustering)**,** in cluster 0, most of the ratings given by customers were 4 and 5 and the least chosen was the 0 rating meaning that the customers are satisfied with Online Booking accessibility. In cluster 1 we gathered that most of the ratings from the customers were 2 and 3 and the least chosen was the 0 rating not giving good feedback regarding the Online Booking accessibility. In the **Price Quality by Cluster** **(**Graph 49: Price Quality Satisfaction (KNN Modelling Clustering)**)**, in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 1 rating. Most of the customers are satisfied with price-quality. In cluster 1 most of the ratings from the customers were 2 and 3 and the least chosen was the rating 5 meaning that the customers were not very satisfied with price-quality. In the **Room Space by Cluster** **(**Graph 61: Room Space Satisfaction (KNN Modelling Clustering)**)**, in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 0 rating. Most of the customers are satisfied with their room space. In cluster 1 we concluded that most of the ratings from the customers were 2 and 3 and the least chosen was the 0 rating. This shows that the customers were not satisfied with their room space. In the **Check-Out by Cluster** **(**Graph 55: Check-Out Satisfaction (KNN Modelling Clustering) in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 1 rating. Most of the customers are satisfied with their Check-out service. In cluster 1 we concluded that most of the ratings from the customers were 3 and 4 and the least chosen was the rating 5. That means the customers are reasonably satisfied with their Check-out service. In the **Check-In by Cluster** **(**Graph 56: Check-In Satisfaction (KNN Modelling Clustering)**)**, we concluded that in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 1 rating. This means most of the customers are very satisfied with their Check-In service. In cluster 1 we concluded that most of the ratings given by the customers were 1 and 3 and the least chosen was the 5. That means the customers are not at all satisfied with their Check-In service. In the **Cleanliness by Cluster** **(**Graph 54: Cleanliness Satisfaction (KNN Modelling Clustering)**)**, in cluster 0, most of the ratings given by the customers were 4 and 5 and the least chosen was the 1 rating. Most of the customers are very satisfied with their cleaning service. In cluster 1 we grasped that most of the ratings from the customers were 3 and 4 and the least chosen was the rating 5. In the **Bar Service by Cluster** **(**Graph 57: Bar Service Satisfaction (KNN Modelling Clustering)**)**, we gathered that in cluster 0, most of the ratings from the customers were 4 and 5 and the least chosen was the 0 rating. Most of the customers are very satisfied with their bar service. In cluster 1 most of the ratings from the customers were 2 and 3 and the least chosen was the 0 rating. The analysis shows that the customers are not very satisfied with their bar service.

In the **Characteristic Cluster, we concluded that in cluster 0**, most of the customers tend to churn more than no-churn, but the churns and no churns have almost equal values, we could also understand that the customers are loyal to BookMe because their longevity in *(Yes)* is higher than *(No)* results. The type of travel preferred is leisure and the customers prefer single rooms less than they prefer double rooms and suites and we also recognized the group of customer ages, the majority are Adults and single groups of Seniors. **In cluster 1**, most of the customers tend to no-churn, we also understood that the customers are loyal to BookMe because their longevity in *(Yes)* is higher than *(No)* results, the type of travel preferred is business and the customers prefer single rooms, we also recognized the group of customer ages, the majority are Adults, and we have a single group of Seniors. **We gathered that in cluster 2**, most of the customers tend to churn, the customers are loyal to BookMe because their longevity in *(Yes)* is higher than *(No)* results, the type of travel preferred is business and the customers prefer not double rooms and suites rather than single rooms, the group of customer ages are in the majority adults.

4.2 Self-Organizing Maps

The other strategy we used after the KNN-Means Clustering **was the SOM**. The SOM algorithm is another unsupervised technique that helps to understand high dimensional data by reducing the dimensions of data to a map and represents clustering by grouping similar data together. After importing the packages, we converted the package np to arrays and defined a default size of SOM grid with 20 rows and 20 columns. Then we created a SOM instance for both *(Clients Satisfaction)* and *(Clients Characteristics)* and trained the model.

For the interpretation of the **U-Matrix**, we got a better perspective on how many clusters we should take to access our data and make proper segmentation. **U-Matrices**, (*finds the distances in the input space of neighbouring units in the output space to visualize a better solution in the graph, the dark blue areas define possible clusters (lower distance between neurons) while lighter/red areas correspond to larger distances)*, we did this process for all variables, to understand each unit, each unit is coloured according to the weight of each variable in the SOM. By using a large grid and performing another clustering on top of the results, the k-means over the SOM to get clusters and obtained 4 clusters for *(Clients Satisfaction)* and 4 clusters for *(Clients Characteristics)* and analysed their variables outcome with the different measures.

Then we decided to interpret the plain components of each variable considering the characteristics of the clients and the satisfaction of the clients, showing that most plain components that we analysed for the customer satisfaction **have high values** *(larger relationships between two unrelated datasets (neurons) by quantifying their similarity,* that means the surrounding weights are very different from the variables*) (Graph 60: Components Plane Characteristics* (SOM)*).* In the characteristics of the clients, we could understand that the variables **Churn and Longevity have** mostly **higher values**, and the variables **Age, Leisure_Travel, and Single_Room have** mostly **lower values** (*small relationships between two unrelated datasets (neurons) by quantifying their similarity,* that means the surrounding weights are very similar from the variables, this doesn't explain the important differences between the variables*)(*Graph 64: Components Plane Satisfaction (SOM)).

Then, we made a copy of our data from pre-processing but with the 3 clusters obtained from *(Clients Satisfaction)* and 4 clusters obtained from *(Clients Characteristics)* and made an analysis of the variables by each cluster (cluster 0, cluster 1 and cluster 2). Now **using the SOM model**, we analysed the Client's Characteristics Clustering and Clients Satisfaction Clustering. Starting with **Clients Characteristics Clustering**. In **Churn by Cluster (**Graph 82: Churn Characteristics (SOM)**),** in clusters 0 and 1, we have more churns than no churns and in cluster 2, we have more no churns than churns. The customers from cluster 2 tend to churn less than the clients of clusters 0 and 1.

In **Longevity by Cluster (**Graph 83: Longevity Characteristics (SOM)**),** all the customers from clusters 1, 2, and 3 have high longevity, as they all have more responses "Yes" than "No"**.** The customers are loyal to BookMe. The **Age by Cluster (**Graph 77: Age Characteristics (SOM)**,** in cluster 0, we do not have a huge dispersion between the customers age, in clusters 1 and 2, showing we have a significant dispersion from the ages. The **Single Room by the Factor by Cluster (**Graph 84: Single Room Characteristics (SOM)**)**, in clusters 0 and 1, most of the clients prefer not to book single rooms, in cluster 2 all the clients prefer to stay in single rooms. The **Type of Travel by Cluster (**Graph 85: Type of Travel Characteristics (SOM)**),** in clusters 0 and 1, most of the customers prefer to do leisure travel, and in cluster 2 most of the customers prefer to do business-type travels.

Using the SOM in **Clients Satisfaction Clustering** in **Comfort by Cluster (**Graph 66: Comfort Satisfaction (SOM)**),** in cluster 0, most of the clients' rates 4 or 5, showing satisfaction with the comfort of their room. In clusters 1 and 2, most of the clients' rates 2 or 3, customers are not at all satisfied with the comfort of their room. In the **Reception Schedule by Cluster (**Graph 75: Reception Schedule Satisfaction (SOM)**)**, most of the customers from cluster 0, have rated 4 or 5, showing good feedback with the reception schedule. In cluster 2 specifically, most customers' rates 1 or 2, showing that they are not very satisfied with the reception schedule. In the **Food & Drink by Cluster (**Graph 69: Food and Drink Satisfaction (SOM)**)**, we gathered that most of the customers from cluster 0, rates 4 or 5. They are very satisfied with the food and drink services available. In cluster 1, most of the customers were not very satisfied with the food and drink services available, by rating 2 or 3. In cluster 2, most customers rates 1 or 2, being also not at satisfied with the food and drink services available. In the **Location by Cluster** (Graph 68: Location Satisfaction (SOM)), in cluster 0, most of the customers have rated 4 or 5 showing that are very satisfied with the location of their accommodation. In cluster 1, most of the customers have rated 3 or 4. They are satisfied with the location of their accommodation in opposite with cluster 2 where most of the customers have rated 1 or 2. In the **Wi-Fi by Cluster** (Graph 79: Wifi Satisfaction (SOM)), most customers from clusters 1 and 2 are satisfied with their Wi-Fi services but in cluster 0, most customers are not satisfied with their Wi-Fi services. The **Amenities by Cluster** the customers are satisfied with their Amenities. In the **Staff by Cluster** (Graph 78: Staff Satisfaction (SOM)), only clusters 0 and 2 are satisfied with their Staff. In the **Online Booking by Cluster**, **Price Quality by Cluster**, **Room Space by Cluster**, **Check-Out by Cluster, Check-In by Cluster, Cleanliness by Cluster, Bar Service by Cluster,** most customers from clusters 0 and 2 are overall satisfied with the hotel services comparing with cluster 1.

5. Description of Customer Segments

For further analysis, we decided to use **K-Means Clustering in our model decision.** We started by organizing the datasets and concatenate in a single table all data from Characteristic Cluster and Satisfaction Cluster. We created 6 clusters, saving all the clusters in data frames and created two single tables explaining the descriptive statistics from the 6 clusters, one table for Characteristic Clusters and another one for Satisfaction Clusters.

Starting with **Final Clients Characteristics Clustering**, we concluded in **Churn by Cluster** (Graph 88: Churn Characteristics (Final Model)) in the clusters 10 and 00, most customers no-churn respectively with (4426 customers), and (1345 customers), then in clusters 11, 20, 01, 21, most customers tend to churn respectively with (1050 customers), (1389 customers), (1281 customers), (2192 customers), Observing all clusters, most of the customers prefer mostly to churn than no-churn. In the **Longevity by Cluster** (Graph 81: Longevity Characteristics (Final Model))**,** all the customers from each cluster are loyal to BookMe, as we have in all clusters a higher number of answers (Yes) than answers (No). The cluster 10, had the most customers from all clusters respectively with (4724 customers). In the **Type of Travel by Cluster** (Graph 90: Type of Travel Characteristics (Final Model))**,** most of the customers that are in clusters 10, 11 and 21 prefer to do business travel respectively with (4664 customers), (1907 customers), (1940 customers), and the customers that are in clusters 20, 01, 00 prefer to do leisure travel respectively with (1171 customers), (1041 customers), (1863 customers). In the **Single Room Factor by Cluster**, most of the customers that are in clusters 10 and 11 prefer single rooms respectively with (4617 customers), (1796 customers) representing 86% of all customers that prefer single rooms, and most of customers that are in clusters 20, 01, 21 and 00 prefer not to stay in single rooms respectively with (1775 customers), (1742 customers), (2206 customers), (2190 customers). In the **Low Loyalty by Cluster** (Graph 80: Low Loyalty Characteristics (Final Model)), from all clusters, except in cluster 10, we do not have a low loyalty. That means most of the customers of each cluster that are not in cluster 10, are loyal to BookMe, representing 81% of high loyalty total, respectively with (1413 customers), (1683 customers), (1667 customers), (2337 customers) and (1928 customers). In the **High Loyalty by Cluster** (Graph 89: High Loyalty Characteristics (Final Model)), in all clusters we have more customers with high loyalty than those with not high loyalty. Most customers from all clusters, are highly loyal to BookMe. In the **Age Group – Children, Adults, and Seniors** (Graph 87: Children Group Characteristics (Final Model)), (Graph 86: Adult Group Characteristics (Final Model)), (Graph 98: Seniors Group Characteristics (Final Model)) **by Cluster,** most of the clients from all clusters are adults. Two important notes, in clusters 01 and 00 we do not have children, and in the clusters 20 and 21 we do not have seniors. In the **Reward Points by Cluster** (Graph 103: Reward Points Characteristics (Final Model)), we have similar dispersions from Reward Points in the clusters 11 and 01, or 10 and 00, or 20 and 21.

Now, interpreting the **Final Clients Satisfaction Clustering.** In the **Comfort by Cluster** (Graph 91: Comfort Satisfaction (Final Model))**,** on clusters 10 and 00, most of the customers have rated 4 or 5 showing that are very satisfied with the comfort of the room unlike in clusters 11, 01, and 21, where most of the customers rates 2 or 3. The **Reception Schedule**

by Cluster **(**Graph 104: Reception Schedule Satisfaction (Final Model)**)**, on clusters 10, 20, 01 and 00 most of the customers rated 4 or 5 comparing with clusters 21 and 11 where most of the customers rated 2 or 3. The **Food & Drink by Cluster** **(**Graph 102: Food and Drink Satisfaction (Final Model)**)**, on cluster 10 most of the customers have rated 4 or 5 showing that they are very satisfied with the Food & Drink availability unlike on clusters 11 and 21, most customers have rated 2 or 3. The **Location by Cluster** on cluster 11 most of the customers have rated 2 or 3. In the **Wi-Fi by Cluster** **(**Graph 107: Wifi Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers have rated 4 or 5 compared with clusters 11, 01, and 21 where most of the customers have rated 2 or 3. In the **Amenities by Cluster** **(**Graph 96: Amenities Satisfaction (Final Model)**)**, on clusters 10, and 00 most of the customers have rated 4 or 5 approving the accommodation amenities. On clusters 11, 20, most of the customers have rated 3 or 4 and on cluster 21 most of the customers have rated 2 or 3. In the **Staff by Cluster** **(**Graph 106: Staff Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 we can conclude, that most of the customers have rated 4 or 5, on clusters 11, 01 most of the customers have rated 3 or 4 and on cluster 21 most of the customers have rated 2 or 3 having a big discrepancy in the results. In the **Online Booking by Cluster** **(**Graph 100: Online Booking Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers have rated 4 or 5. On clusters 11, 01, and 21 most of the customers have rated 2 or 3 showing that are not satisfied with the Online Booking service. In the **Price Quality by Cluster** **(**Graph 99: Price-Quality Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers are very satisfied with the Price Quality relationship. On clusters 11, 01, most of the customers are not satisfied with the Price Quality relationship. On cluster 20 most of the customers have rated 1 or 3. In the **Room Space by Cluster** **(**Graph 105: Room Space Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers have rated 4 or 5. They are very satisfied with the room space. On clusters 11, and 21 most of the customers have rated 2 or 3. They are not satisfied with the room space. On cluster 01 most of the customers have rated 2 or 4. In the **Check-Out by Cluster** **(**Graph 94: Check-Out Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers have rated 4 or 5, giving great feedback. On clusters 01, and 21 most of the customers have rated 3 or 4. On cluster 11 most of the customers are not satisfied with the Check-Out service. In the **Check-In by Cluster** **(**Graph 93: Check-In Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 most of the customers are very satisfied with the Check-In service. On clusters 01, and 21 most of the customers have rated 3 or 4 and on clusters 11, 01, and 21 most of the customers have rated 1 or 3 being not very satisfied with the Check-In service. In the **Cleanliness by Cluster** **(**Graph 92: Cleanliness Satisfaction (Final Model)**)**, on clusters 10, 20, and 00 they are very satisfied with the Cleanliness service. On clusters 01, and 21 most of the customers have rated 3 or 4. On cluster 11 most of the customers have rated 2 or 3 are not being satisfied with the Cleanliness service. In the **Bar Service by Cluster** **(**Graph 95: Bar Service Satisfaction (Final Model)**)**, on clusters 10, 20, and 00, most of the customers have rated 4 or 5 being very satisfied with the Bar service. On clusters 01, and 21 most of the customers have rated 2 or 3 and on cluster 11 most of the customers have rated 3 or 4.

6. Marketing Plan

After interpreting the results from our clusters, our strategy, after observing the **Final Client Characteristics and Final Client Satisfaction results**, were give more emphasis on the customers from cluster 11 and the cluster 21. To improve and attract these BookMe clients, the company should figure out the main problems in that departments that let customers unsatisfied (they should improve their technology, contract collaborators more qualified, provide a better accommodation environment, provide formation to new employees for they be able do their jobs perfectly, hire a new food and drink retailer, they should contract new employees to develop their website, provide better clean and bar services conditions to customers), for that BookMe will do an investment in these specific areas the comfort of the rooms, the reception schedule, the food and drink availability, accommodation in location, Wi-Fi services, accommodation amenities, their Staff, Online Booking, price-quality services, the room space, Check-In, Check-Out, Cleanliness conditions, and their bar-services for solve more detailly the main problems in these departments.

7. Conclusions

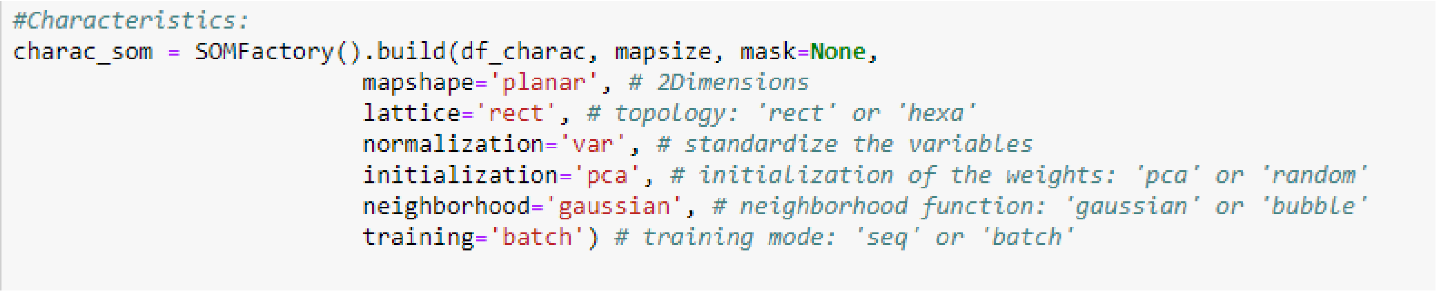
Given the problem regarding the marketing department, to create more efficient campaigns and don't waste **wrongly** their money, we realized that most customers were middle-aged adults on business travels and the customers that rate worst in any satisfaction level, usually churn on the company. We also noticed that we had to make some adjustments to have a more precise and correct model by treating missing values and outliers, checked the coherence of the dataset, created some new variables and defined the perspectives. Then, we did the preparation of both validation and test datasets and write different algorithms for the different models and adjust them with different parameters to improve them and get better results. With this, we obtained different outcomes so after some analysis and comparison, we evaluated each one of them, and based on the values we decided to get the results of the test dataset using the KNN-Clustering model.

In conclusion, we strongly believe that the model we implemented in our project will be a good solution to help in future decisions of the Marketing Department and improve future sales.

8. Other Cluster Techniques and Theorical Explanation

8.1. PCA

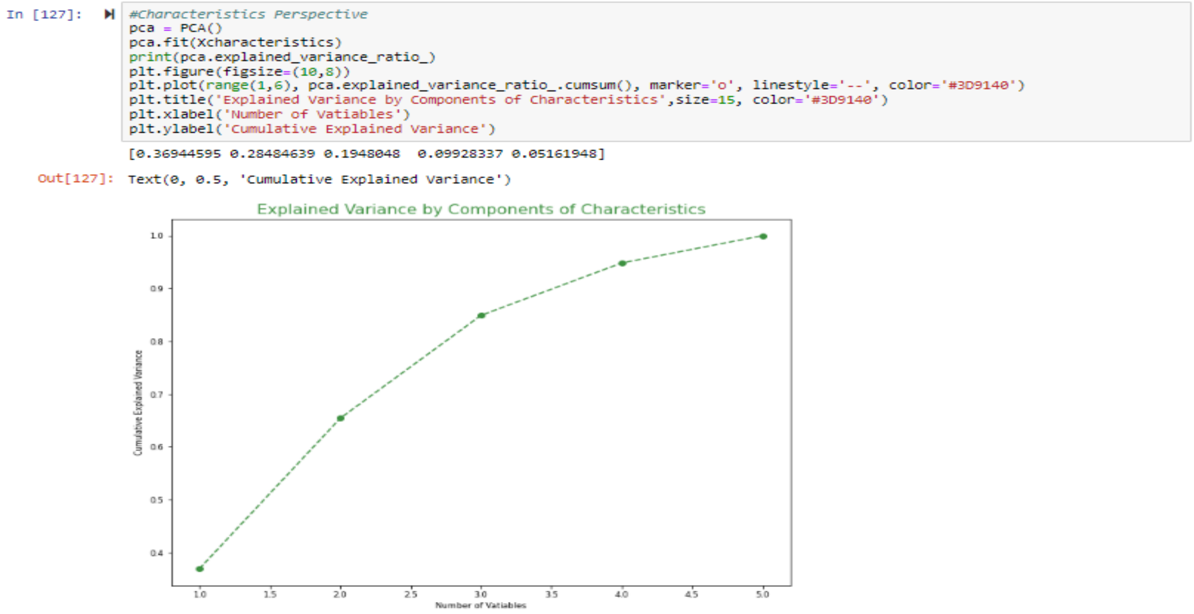
Regarding the PCA (Principal Component Analysis), a method to reduce the dimension of the data in larger data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. With this method, we segregate the variables in fewer ones correlated with each other to achieve better results. The goal using the PCA was knowing the number of seed to use in our project to give a further and more accurate results. In our model, we used the PCA as initiator of the SOM to separate the data according to their characteristics and provide us a better results and better execution of our model. And we did the same for the satisfaction SOM analysis.



```
#Characteristics:
charac_som = SOMFactory().build(df_charac, mapsize, mask=None,
                         mapshape='planar', # 2Dimensions
                         lattice='rect', # topology: 'rect' or 'hexa'
                         normalization='var', # standardize the variables
                         initialization='pca', # initialization of the weights: 'pca' or 'random'
                         neighborhood='gaussian', # neighborhood function: 'gaussian' or 'bubble'
                         training='batch') # training mode: 'seq' or 'batch'
```

Code 1: Characteristics (SOM)

In addition, to complement our study and have a deeper and more precise analysis of the number of clusters needed, we decided to use the PCA once again, to verify the number of cluster to use in KNN.

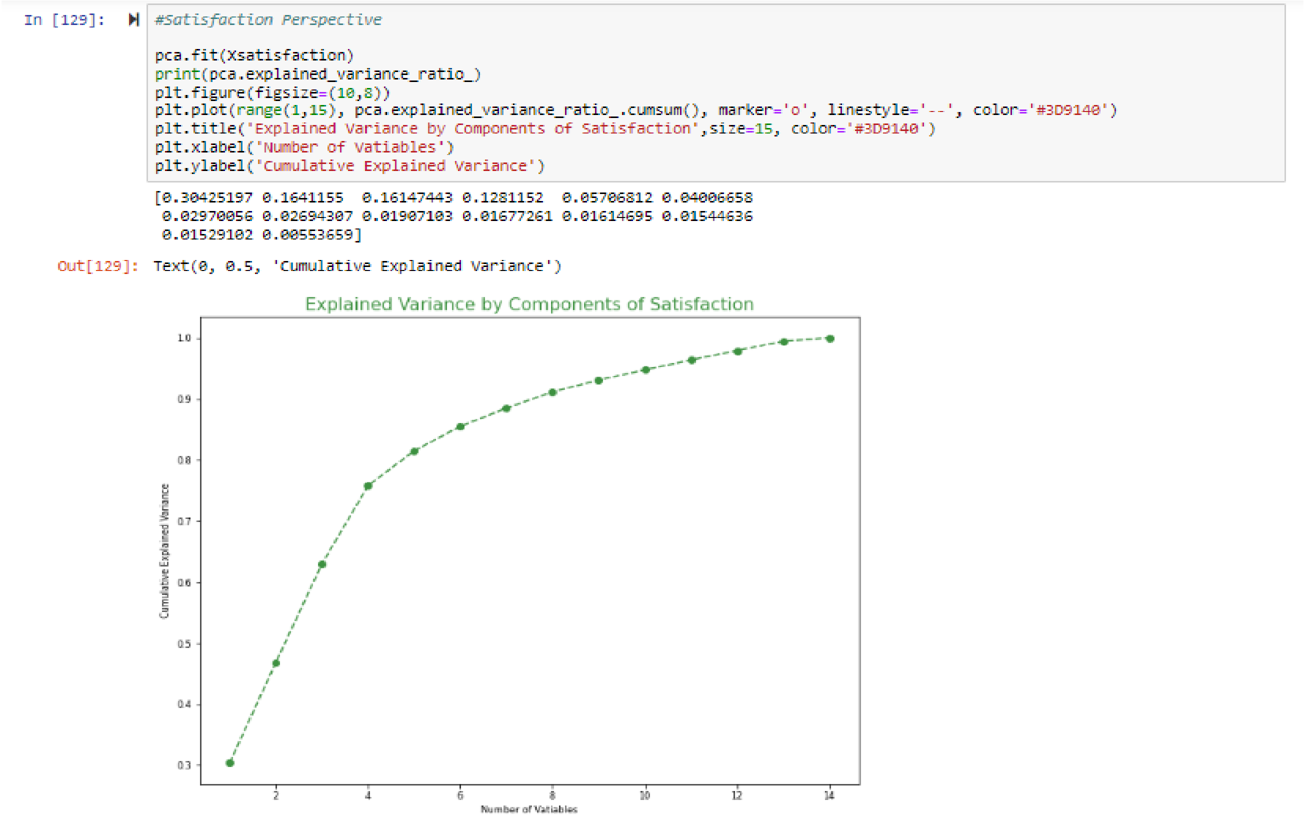


```
In [127]:  #Characteristics Perspective
           pca = PCA()
           pca.fit(Xcharacteristics)
           print(pca.explained_variance_ratio_)
           plt.figure(figsize=(10,8))
           plt.plot(range(1,6), pca.explained_variance_ratio_.cumsum(), marker='o', linestyle='--', color='#3D9140')
           plt.title('Explained Variance by Components of Characteristics',size=15, color='#3D9140')
           plt.xlabel('Number of Vatiables')
           plt.ylabel('Cumulative Explained Variance')

           [0.36944595 0.28484639 0.1948048  0.09928337 0.05161948]

Out[127]:  Text(0, 0.5, 'Cumulative Explained Variance')
```

Explained Variance by Components of Characteristics
Cumulative Explained Variance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of Vatiables

Graph 1: PCA Characteristics Method

From the characteristic's perspective, we use the variance of each observation to explain how much the components are correlated with the variables. By plotting a graph were the x axis from 1-6, corresponding with the number of variables and the y axis from 0-1 corresponding with the Cumulative Variance explained, the best result came where the number of clusters in the Characteristics were 3 since the variance explained it must be greater than 0.7.



```
In [129]:  #Satisfaction Perspective

           pca.fit(Xsatisfaction)
           print(pca.explained_variance_ratio_)
           plt.figure(figsize=(10,8))
           plt.plot(range(1,15), pca.explained_variance_ratio_.cumsum(), marker='o', linestyle='--', color='#3D9140')
           plt.title('Explained Variance by Components of Satisfaction',size=15, color='#3D9140')
           plt.xlabel('Number of Vatiables')
           plt.ylabel('Cumulative Explained Variance')

           [0.30425197 0.1641155  0.16147443 0.1281152  0.05706812 0.04006658
            0.02970056 0.02694307 0.01907103 0.01677261 0.01614695 0.01544636
            0.01529102 0.00553659]

Out[129]:  Text(0, 0.5, 'Cumulative Explained Variance')
```

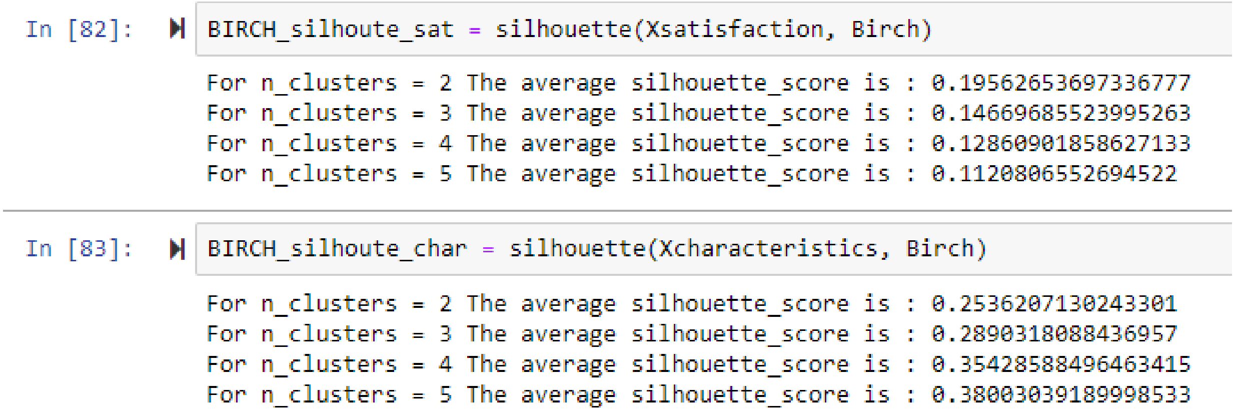
Explained Variance by Components of Satisfaction
Cumulative Explained Variance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
2
4
6
8
10
12
14
Number of Vatiables

Graph 2: PCA Satisfaction Method

From the Satisfaction perspective, we did the same thing. We perform the PCA regarding the data related to the satisfaction and on the x axis we put the number of variables (1-15) and on the y axis we put the cumulative explained variance. In this case the number of clusters were 4 since the variance of the variables explained were above 0.75.

8.2. BIRCH

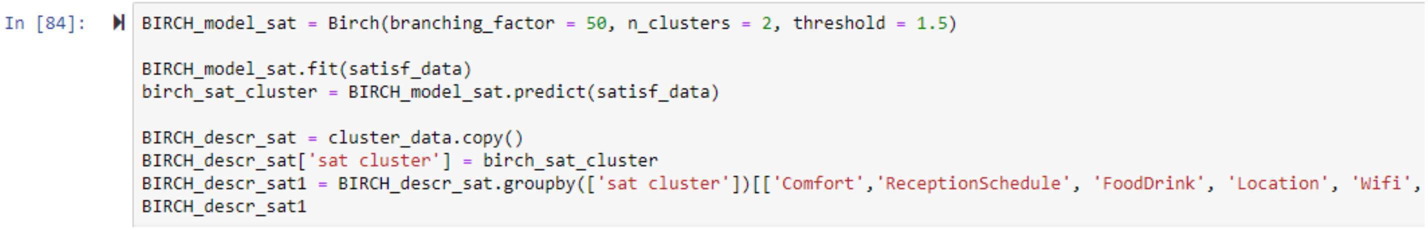
The other method that we used for cluster analysis was the Birch Model. The Birch Model (Balanced Iterative Reducing and Clustering using Hierarchies) is a cluster algorithm that performs in large data sets by first generating a small and compact summary of the large dataset that retains as much information as possible and then clustering those small summaries. The BIRCH doesn't directly cluster the dataset only indirectly. We started the algorithm by importing and access the best average silhouette score regarding the satisfaction, which was two clusters. We did the same thing to access the best average silhouette score regarding the characteristics, given us 3 clusters.



```
In [82]:  BIRCH_silhoute_sat = silhouette(Xsatisfaction, Birch)

          For n_clusters = 2 The average silhouette_score is : 0.19562653697336777
          For n_clusters = 3 The average silhouette_score is : 0.14669685523995263
          For n_clusters = 4 The average silhouette_score is : 0.12860901858627133
          For n_clusters = 5 The average silhouette_score is : 0.1120806552694522

In [83]:  BIRCH_silhoute_char = silhouette(Xcharacteristics, Birch)

          For n_clusters = 2 The average silhouette_score is : 0.2536207130243301
          For n_clusters = 3 The average silhouette_score is : 0.2890318088436957
          For n_clusters = 4 The average silhouette_score is : 0.35428588496463415
          For n_clusters = 5 The average silhouette_score is : 0.38003039189998533
```

Graph 3: Birch Method Satisfaction and Characteristics

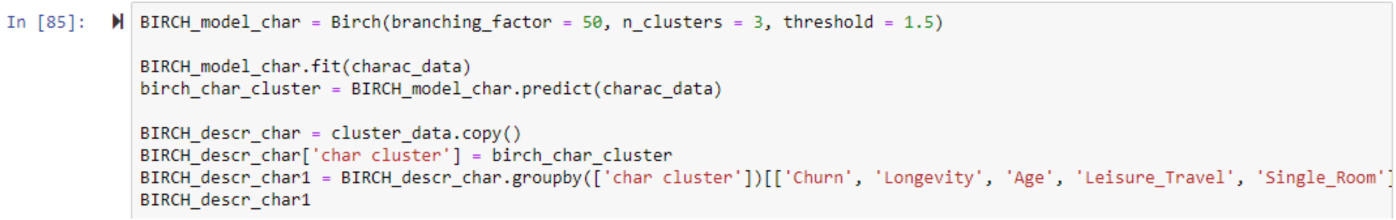
Then, we perform the Satisfaction Model with two clusters and analyse how much each variable is correlated with each cluster to separate the costumers based on the variables associated.



```
In [84]:  BIRCH_model_sat = Birch(branching_factor = 50, n_clusters = 2, threshold = 1.5)

          BIRCH_model_sat.fit(satisf_data)
          birch_sat_cluster = BIRCH_model_sat.predict(satisf_data)

          BIRCH_descr_sat = cluster_data.copy()
          BIRCH_descr_sat['sat cluster'] = birch_sat_cluster
          BIRCH_descr_sat1 = BIRCH_descr_sat.groupby(['sat cluster'])[['Comfort','ReceptionSchedule', 'FoodDrink', 'Location', 'Wifi',
          BIRCH_descr_sat1
```

Code 2: BIRCH Satisfaction

We did the same thing now with the Characteristics model but with three clusters since it performed better with this option, given us better results and values more correlated with the variables.



```
In [85]:  BIRCH_model_char = Birch(branching_factor = 50, n_clusters = 3, threshold = 1.5)

          BIRCH_model_char.fit(charac_data)
          birch_char_cluster = BIRCH_model_char.predict(charac_data)

          BIRCH_descr_char = cluster_data.copy()
          BIRCH_descr_char['char cluster'] = birch_char_cluster
          BIRCH_descr_char1 = BIRCH_descr_char.groupby(['char cluster'])[['Churn', 'Longevity', 'Age', 'Leisure_Travel', 'Single_Room']
          BIRCH_descr_char1
```

Code 3: Birch Characteristics

8.3. KNN-Means Theorical Explanation

For K-means Clustering, in order to have the complete model we had to choose several clusters and randomly assign each data point to a specific cluster until the clusters stopped changing to compute the centroid or assign data points to the closest centroid. Being a supervised learning algorithm, it joins clusters accordingly to their similarity, in unlabelled clusters. In practice it uses random seeds or better initialized seeds (using other methods) to assign every single point in the dataset to a centroid or one of these seeds. After every point is assign to a cluster, the centroid of that cluster moves to the average of all points assign to it. Then the model reassigns all points again to the nearest centroid (after the change in the centroid), and then we will have different cluster from before, and again it moves the centroid accordingly to the average of the points of the data from that cluster. The model only stops iterating, when after grouping the data, the centroid doesn't move anymore, and we achieve a stable centroid and cluster. For this, the choice of the seeds of the centroids is very important because it impacts on how well the model runs.

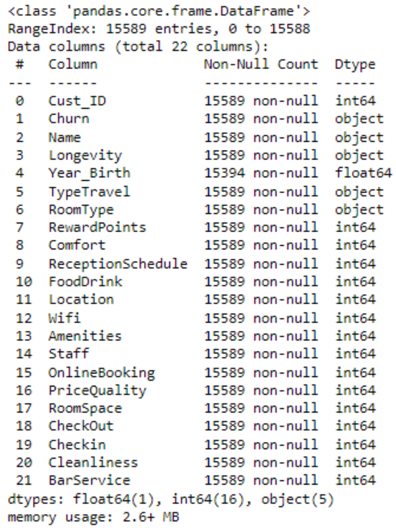
9. *Creativity & Other-Self Study*

As decided, in order to not delete any customer, we use self-study techniques in order to obtain the desirable results, so to start to with the KNN imputer we used it to fill the missing values. The point of KNN method is to identify 'k' samples in the variable (in this case was Year of Birth) that are similar or close in the space. This 'k' was the number of neighbours or samples we will use to estimate the value of the missing data, and for k we end up using five, since it presented the best results. Before using this type of method is important to scale the data since it was necessary to produce good results, we scale it from 0 to 1, and afterwards we inversed the scale to get the true values of all data. (this step was similar to the one made in predictive report)

Then, in order to treat outliers, we search for methods of treating them without deleting them, to do that we saw the skewness of the data, to understand if the data is symmetrically distributed, which showed that the skewness is negative, meaning the data is not equally distributed, so we applied a few methods with the objective to reduce the right skewness. We used the **Square Root Transformation** (Graph 34: Square Root Method) which transforms the range of values in such form that if the value of x is four times bigger than w, the result is only multiplied by two and not by 4. The results given here were better than the previous one, but still worse than the original, so we also rejected it. Then we tried **Cube Root Transformation** (Graph 31: Cube Transformation Method), which elevates the value x one-third times. It can be applied to zero or negative values but with gave us a bad result. We tried the technique **Log Scale** (Graph 30: Log Transformation Method) which is a way of showing numerical data over significantly large range of values in a compact way, the results given were not the best since the skew was even more marked by the log function. At last, we used the **Sine Transformation** (Graph 33: Sine Transformation Method)**,** only selecting the values that were not the outliers. After realizing which method was better, we replaced the older Reward Points with the new one.

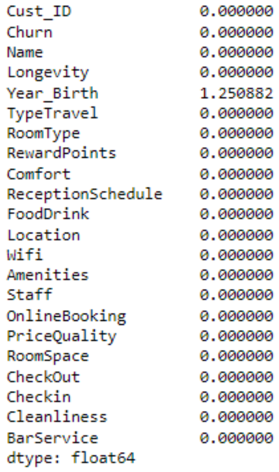
Besides the treatment made in pre-processing we also looked for improving the visualization of the graphs made during the realization of the project. To start, we used a mixed of the libraries of matplotlib and seaborn, using seaborn to create the graph itself and matplotlib to add some features, and group the graphs in a grid. facilitating the view and understanding of each group of graphs.

10. Annex



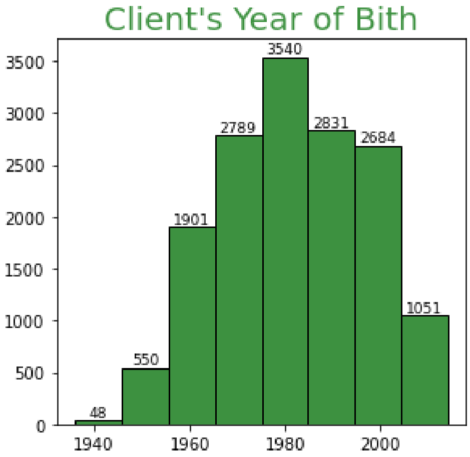
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15589 entries, 0 to 15588
Data columns (total 22 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Cust_ID            15589 non-null  int64
 1   Churn              15589 non-null  object
 2   Name               15589 non-null  object
 3   Longevity          15589 non-null  object
 4   Year_Birth         15394 non-null  float64
 5   TypeTravel         15589 non-null  object
 6   RoomType           15589 non-null  object
 7   RewardPoints       15589 non-null  int64
 8   Comfort            15589 non-null  int64
 9   ReceptionSchedule  15589 non-null  int64
 10  FoodDrink          15589 non-null  int64
 11  Location           15589 non-null  int64
 12  Wifi               15589 non-null  int64
 13  Amenities          15589 non-null  int64
 14  Staff              15589 non-null  int64
 15  OnlineBooking      15589 non-null  int64
 16  PriceQuality       15589 non-null  int64
 17  RoomSpace          15589 non-null  int64
 18  CheckOut           15589 non-null  int64
 19  Checkin            15589 non-null  int64
 20  Cleanliness        15589 non-null  int64
 21  BarService         15589 non-null  int64
dtypes: float64(1), int64(16), object(5)
memory usage: 2.6+ MB
```

Figure 1: Datatypes of our Variables

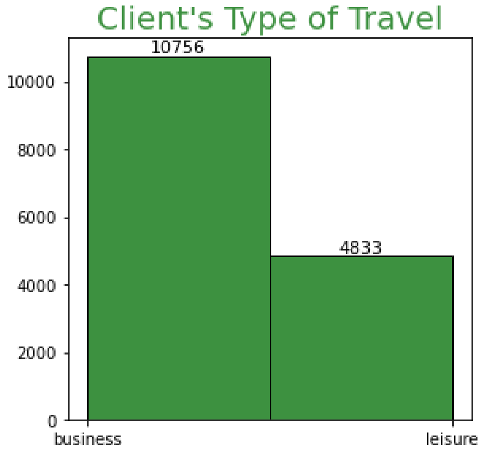


```
Cust_ID              0.000000
Churn                0.000000
Name                 0.000000
Longevity            0.000000
Year_Birth           1.250882
TypeTravel           0.000000
RoomType             0.000000
RewardPoints         0.000000
Comfort              0.000000
ReceptionSchedule    0.000000
FoodDrink            0.000000
Location             0.000000
Wifi                 0.000000
Amenities            0.000000
Staff                0.000000
OnlineBooking        0.000000
PriceQuality         0.000000
RoomSpace            0.000000
CheckOut             0.000000
Checkin              0.000000
Cleanliness          0.000000
BarService           0.000000
dtype: float64
```

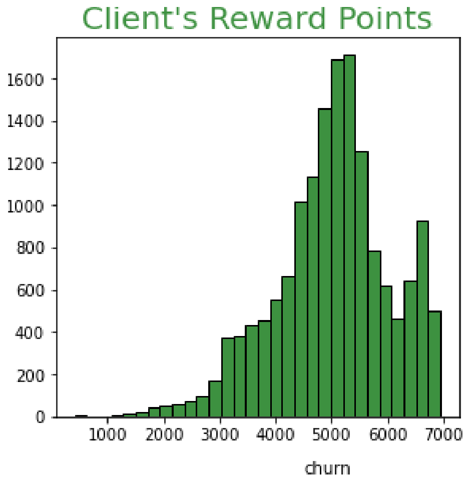
Figure 2: Percentage of Missing Values each Variable



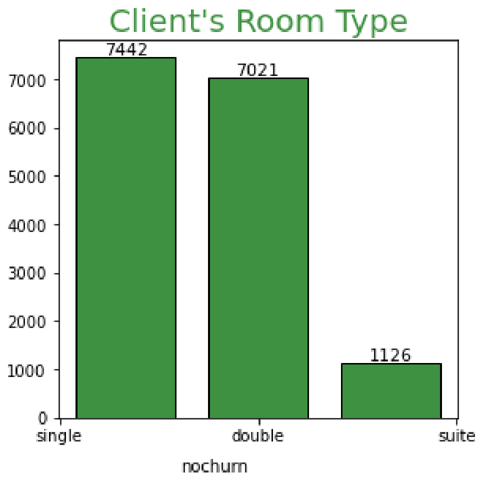
Graph 5: Clients Year of Birth Characteristics (Exploration)



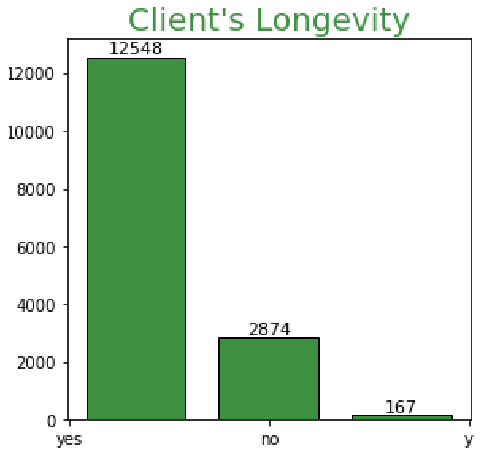
Graph 4: Clients Type of Travel Characteristics (Exploration)

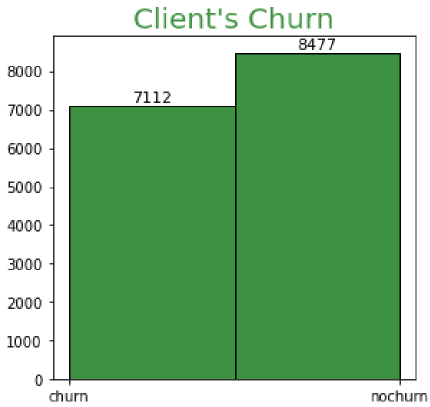


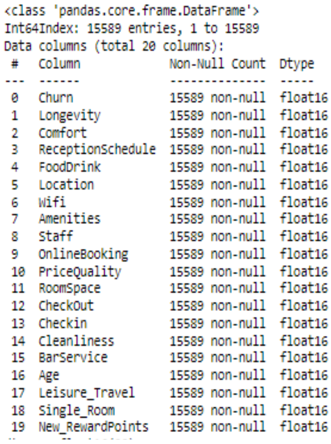
Graph 6: Clients Reward Points Characteristic (Exploration)



Graph 7: Clients Room Type Characteristics (Exploration)





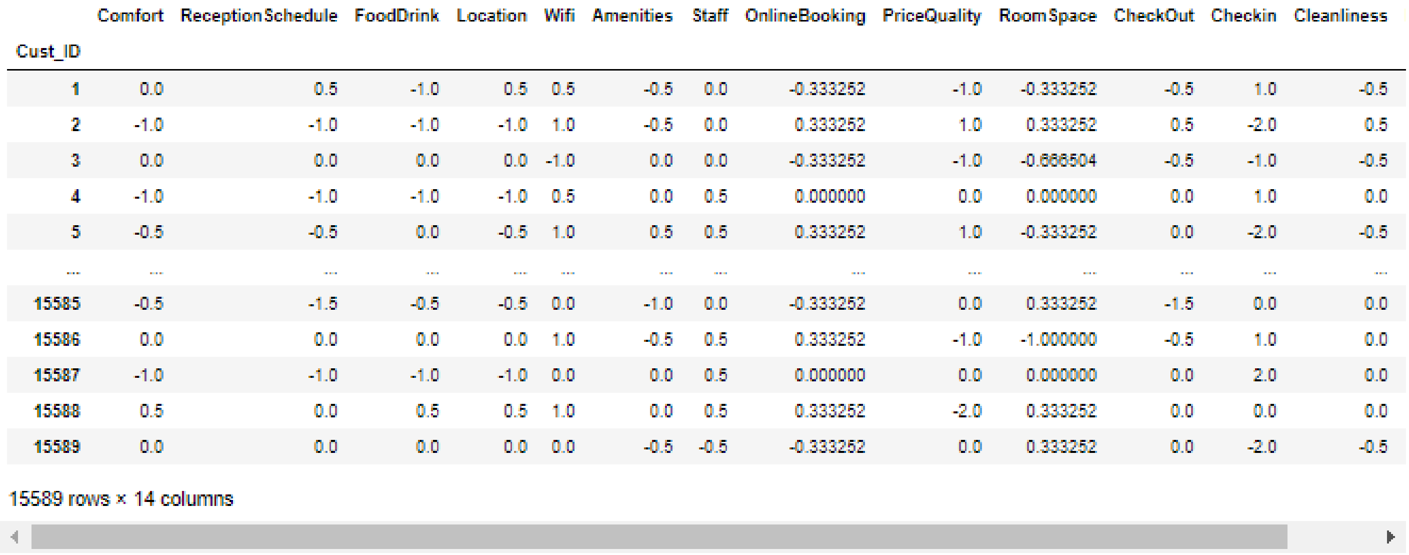


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15589 entries, 1 to 15589
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Churn              15589 non-null  float16
 1   Longevity          15589 non-null  float16
 2   Comfort            15589 non-null  float16
 3   ReceptionSchedule  15589 non-null  float16
 4   FoodDrink          15589 non-null  float16
 5   Location           15589 non-null  float16
 6   Wifi               15589 non-null  float16
 7   Amenities          15589 non-null  float16
 8   Staff              15589 non-null  float16
 9   OnlineBooking      15589 non-null  float16
 10  PriceQuality       15589 non-null  float16
 11  RoomSpace          15589 non-null  float16
 12  CheckOut           15589 non-null  float16
 13  Checkin            15589 non-null  float16
 14  Cleanliness        15589 non-null  float16
 15  BarService         15589 non-null  float16
 16  Age                15589 non-null  float16
 17  Leisure_Travel     15589 non-null  float16
 18  Single_Room        15589 non-null  float16
 19  New_RewardPoints   15589 non-null  float16
```

Table 1: Modifying Data types to float16

Graph 8: Clients Longevity Characteristics (Exploration)

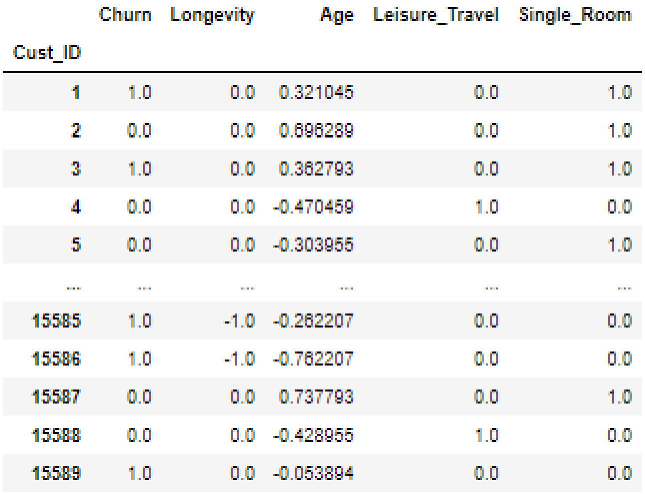
Graph 9: Number of Clients Churn Characteristics (Exploration)



	Comfort	ReceptionSchedule	FoodDrink	Location	Wifi	Amenities	Staff	OnlineBooking	PriceQuality	RoomSpace	CheckOut	Checkin	Cleanliness
Cust_ID													
1	0.0	0.5	-1.0	0.5	0.5	-0.5	0.0	-0.333252	-1.0	-0.333252	-0.5	1.0	-0.5
2	-1.0	-1.0	-1.0	-1.0	1.0	-0.5	0.0	0.333252	1.0	0.333252	0.5	-2.0	0.5
3	0.0	0.0	0.0	0.0	-1.0	0.0	0.0	-0.333252	-1.0	-0.666504	-0.5	-1.0	-0.5
4	-1.0	-1.0	-1.0	-1.0	0.5	0.0	0.5	0.000000	0.0	0.000000	0.0	1.0	0.0
5	-0.5	-0.5	0.0	-0.5	1.0	0.5	0.5	0.333252	1.0	-0.333252	0.0	-2.0	-0.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...
15585	-0.5	-1.5	-0.5	-0.5	0.0	-1.0	0.0	-0.333252	0.0	0.333252	-1.5	0.0	0.0
15586	0.0	0.0	0.0	0.0	1.0	-0.5	0.5	0.333252	-1.0	-1.000000	-0.5	1.0	0.0
15587	-1.0	-1.0	-1.0	-1.0	0.0	0.0	0.5	0.000000	0.0	0.000000	0.0	2.0	0.0
15588	0.5	0.0	0.5	0.5	1.0	0.0	0.5	0.333252	-2.0	0.333252	0.0	0.0	0.0
15589	0.0	0.0	0.0	0.0	0.0	-0.5	-0.5	-0.333252	0.0	0.333252	0.0	-2.0	-0.5

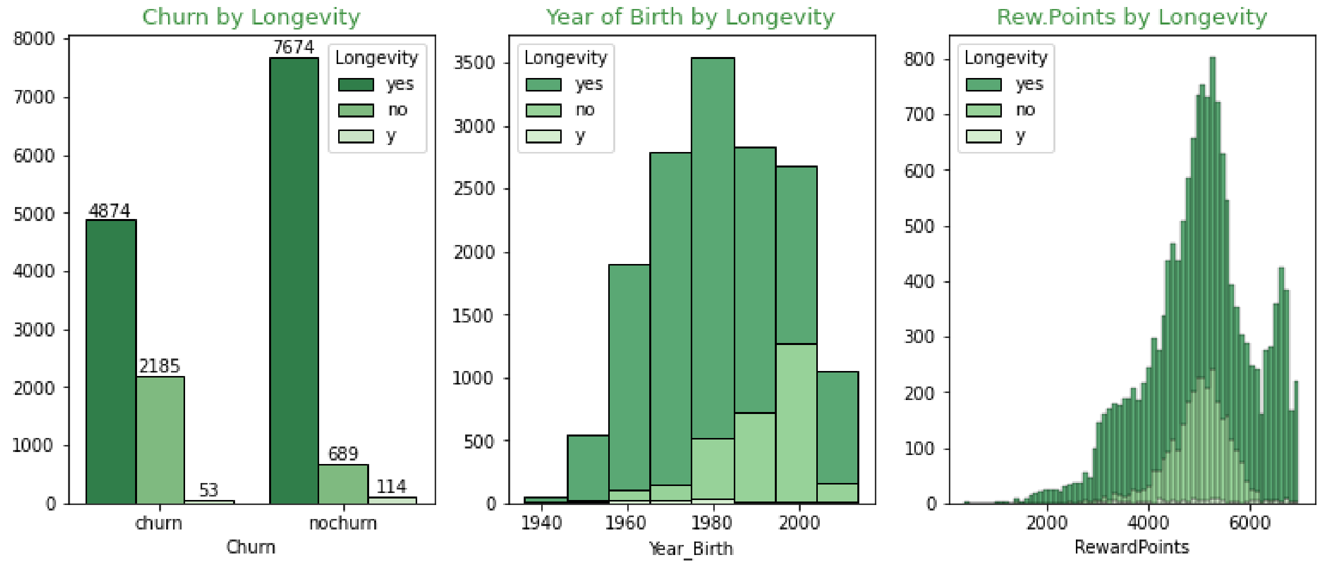
15589 rows × 14 columns

Table 3: Satisfaction Dataset

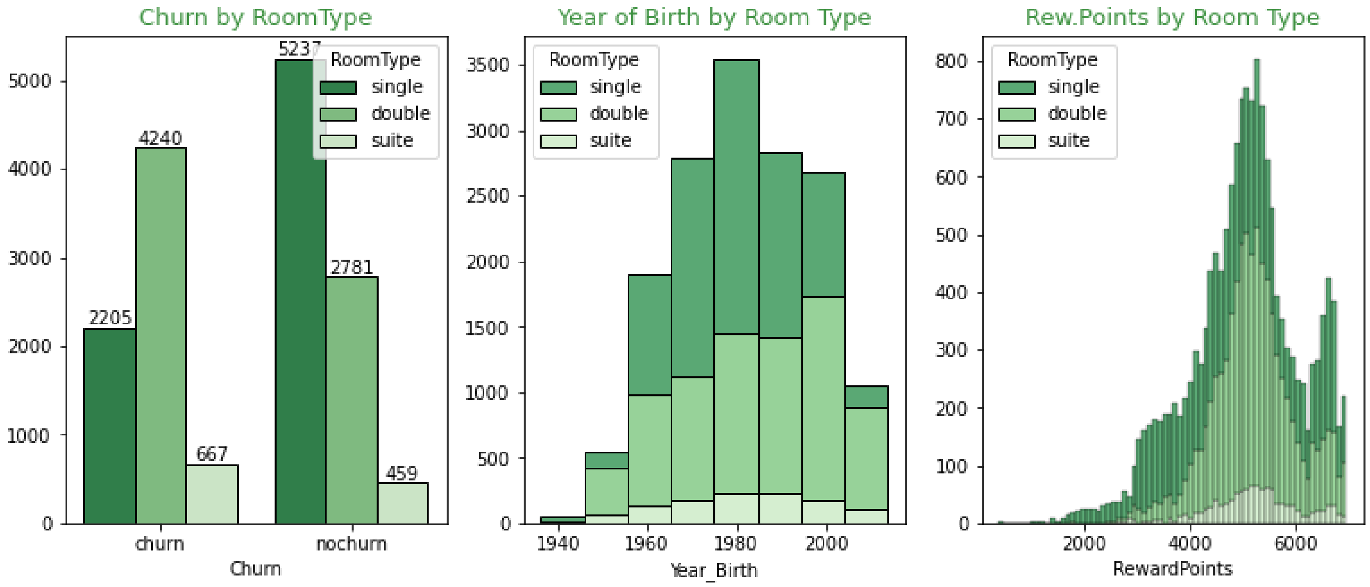


	Churn	Longevity	Age	Leisure_Travel	Single_Room
Cust_ID					
1	1.0	0.0	0.321045	0.0	1.0
2	0.0	0.0	0.696289	0.0	1.0
3	1.0	0.0	0.362793	0.0	1.0
4	0.0	0.0	-0.470459	1.0	0.0
5	0.0	0.0	-0.303955	0.0	1.0
...	...	...	...	...	...
15585	1.0	-1.0	-0.262207	0.0	0.0
15586	1.0	-1.0	-0.762207	0.0	0.0
15587	0.0	0.0	0.737793	0.0	1.0
15588	0.0	0.0	-0.428955	1.0	0.0
15589	1.0	0.0	-0.053894	0.0	0.0

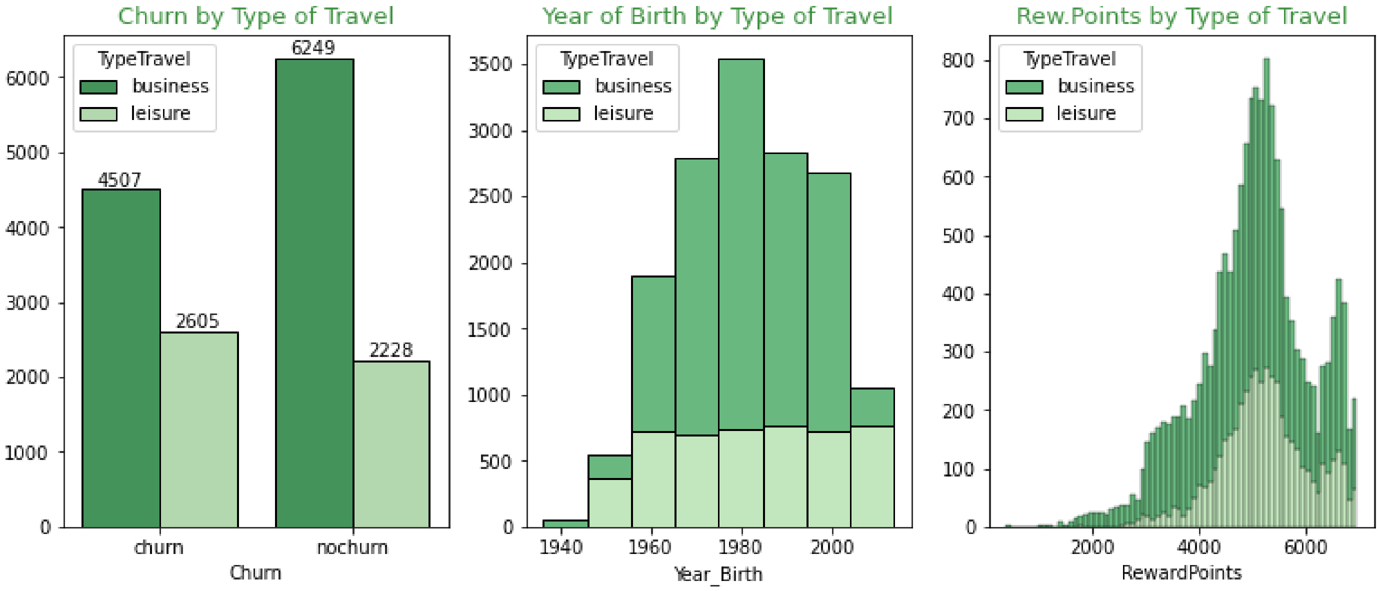
Table 2: Characteristic Dataset



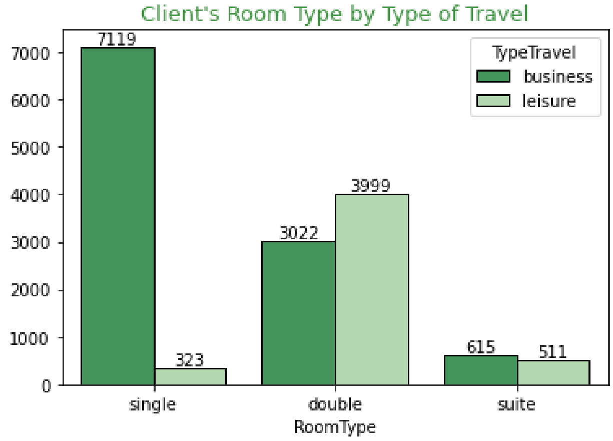
Graph 12: Customer Longevity Characteristics (Exploration)



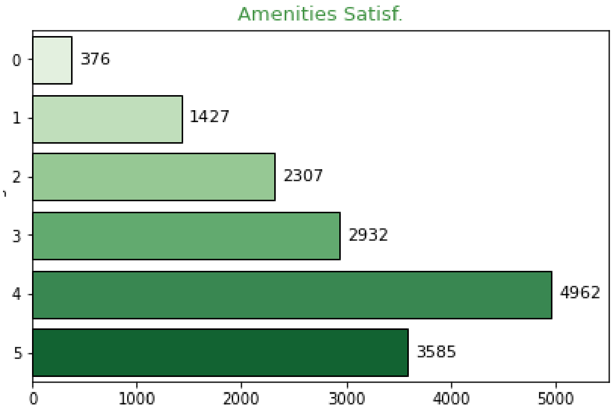
Graph 11: Customer Room Type Characteristics (Exploration)



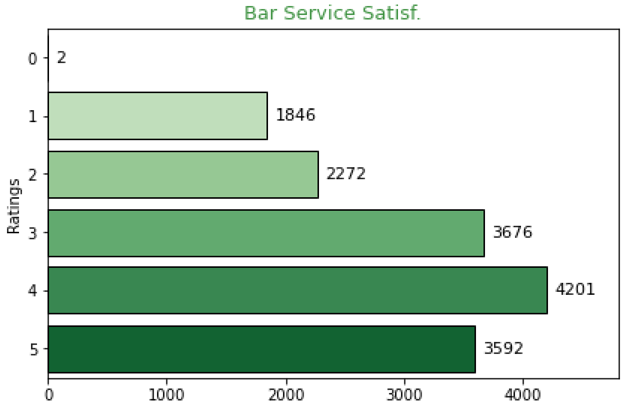
Graph 10: Customer Type of Travel Characteristics (Exploration)



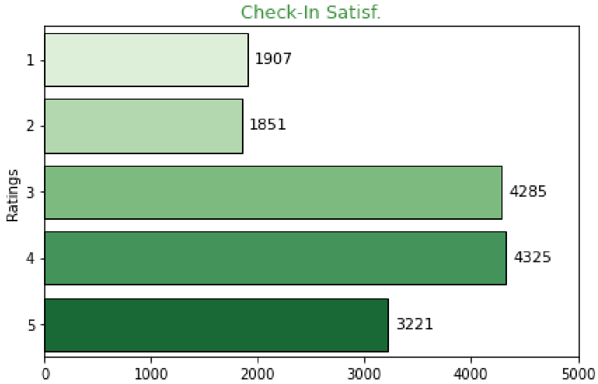
Graph 13: Room Type by Type of Travel Characteristics (Exploration)



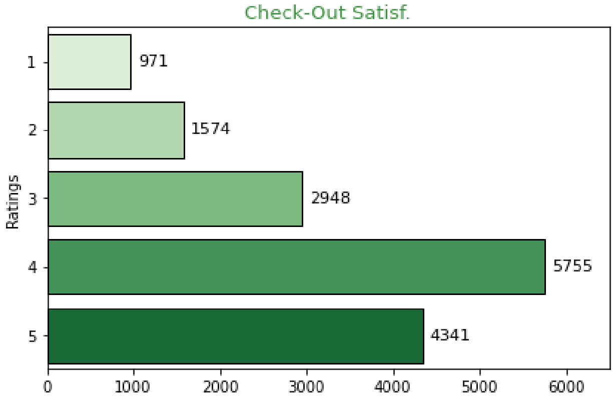
Graph 14: Amenities Clients Satisfaction (Exploration)



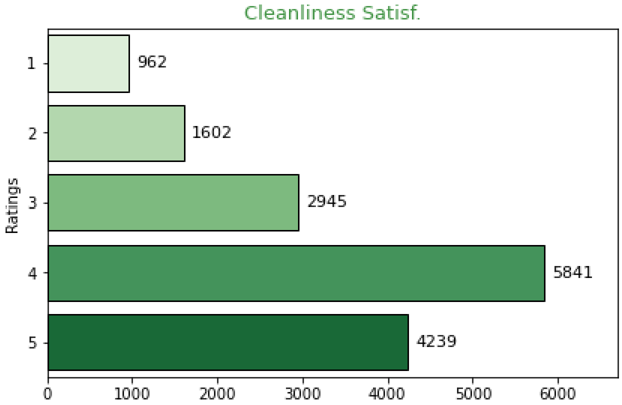
Graph 16: Bar Service Satisfaction (Exploration)



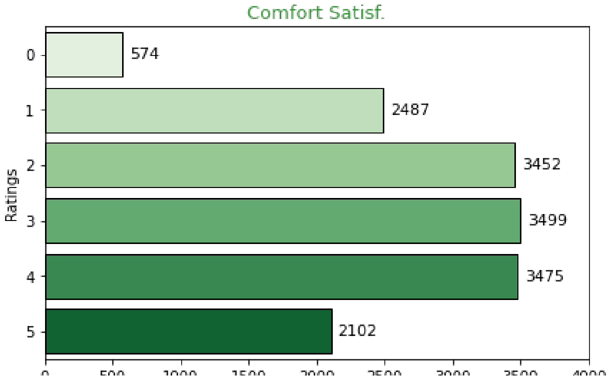
Graph 15:Check-In Satisfaction (Exploration)



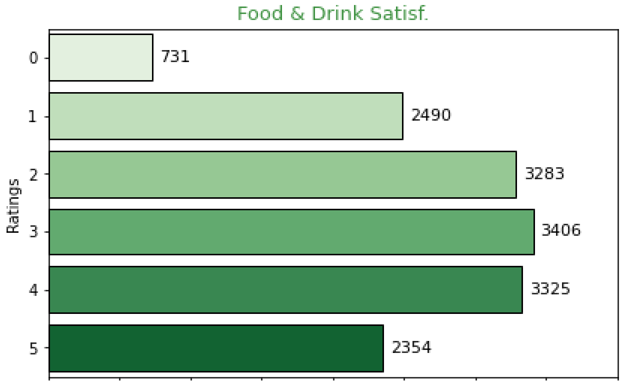
Graph 18: Check-Out Satisfaction (Exploration)



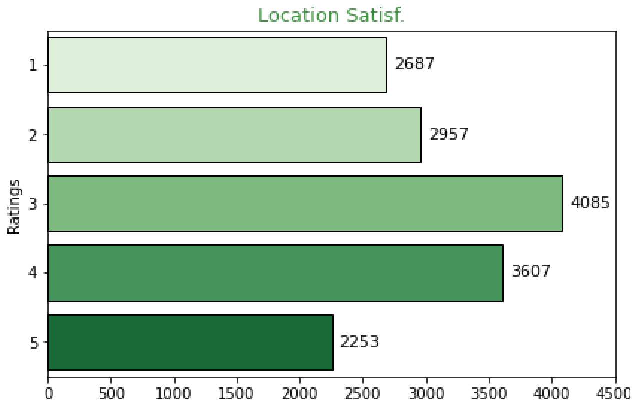
Graph 17: Cleanliness Satisfaction (Exploration)



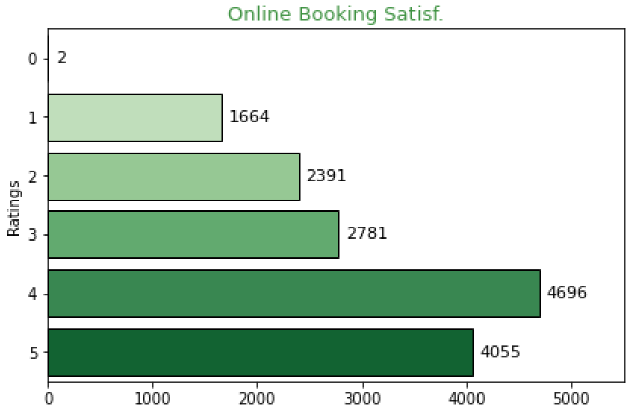
Graph 20: Comfort Satisfaction (Exploration)



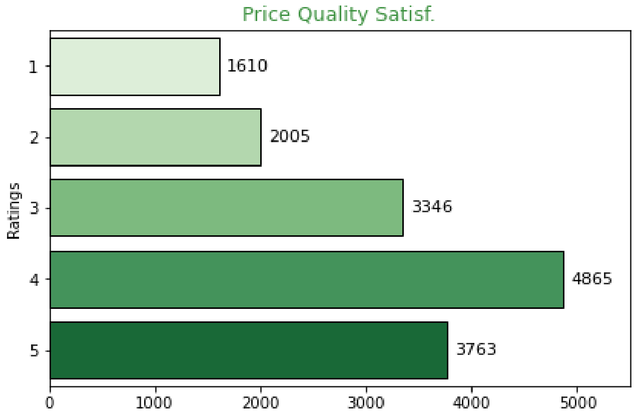
Graph 19: Food and Drink Satisfaction (Exploration)



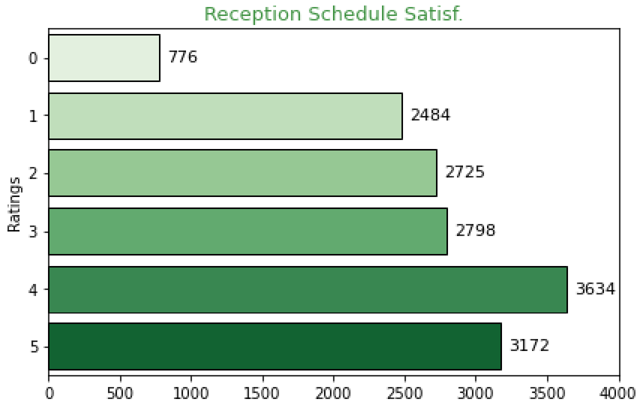
Graph 22: Location Satisfaction (Exploration)



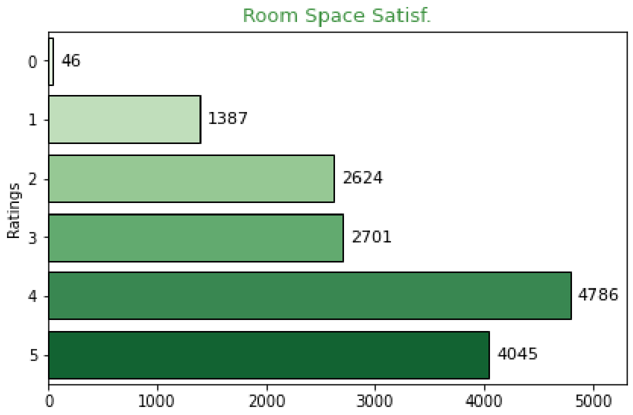
Graph 21: Online Booking Satisfaction (Exploration)



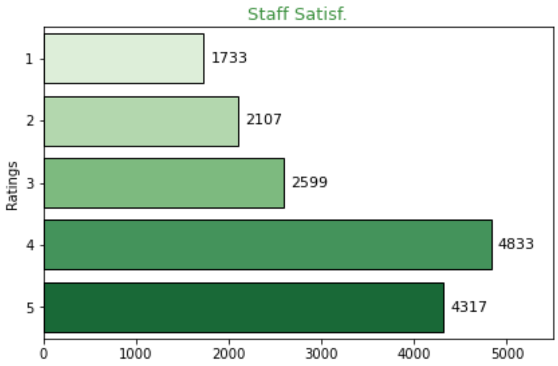
Graph 25: Price- Quality Satisfaction (Exploration)



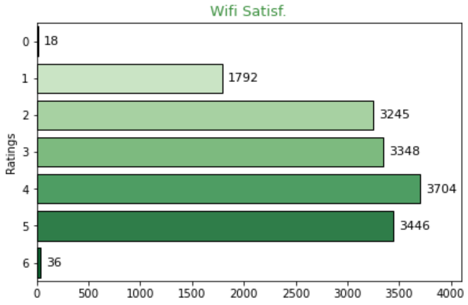
Graph 24: Reception Schedule Satisfaction (Exploration)



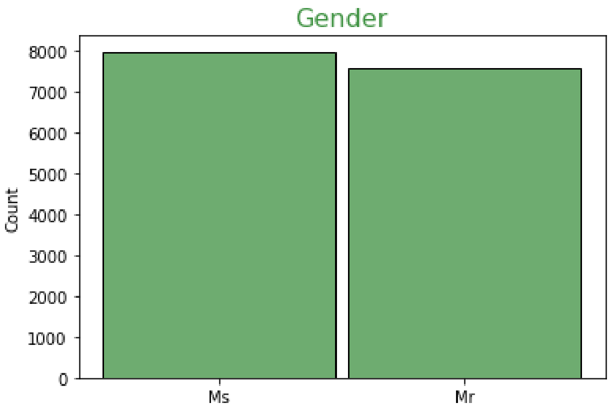
Graph 23: Room Space Satisfaction (Exploration)



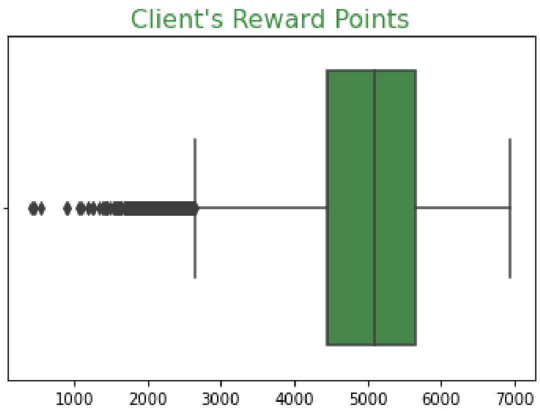
Graph 26: Staff Satisfaction (Exploration)



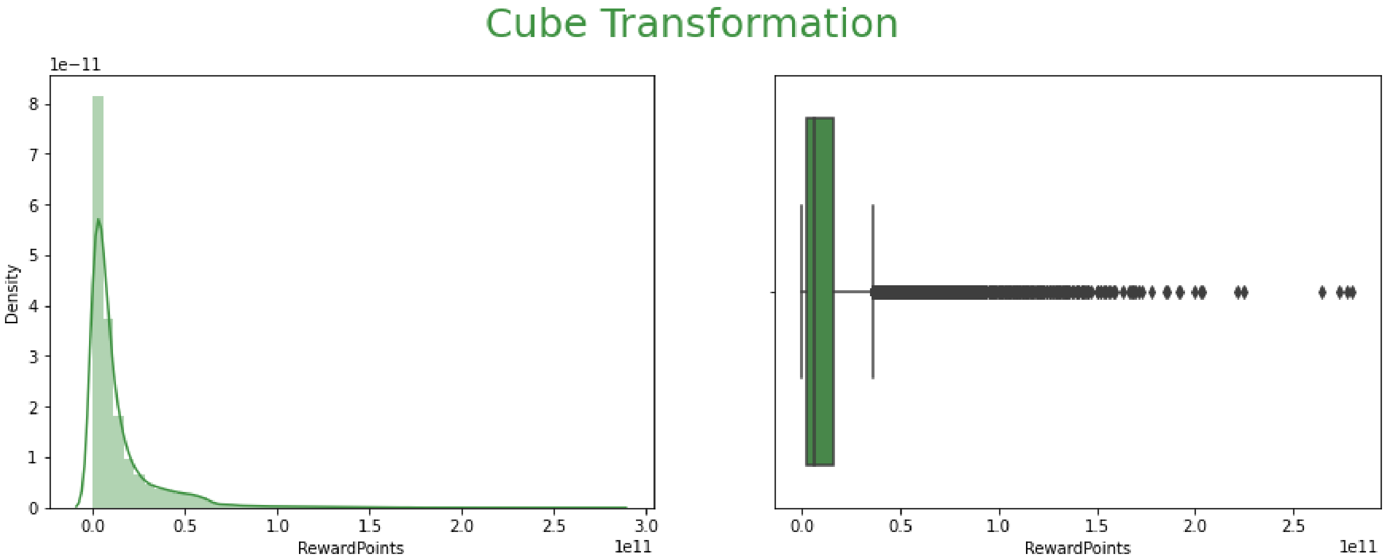
Graph 27: Wifi Satisfaction (Exploration)


Gender

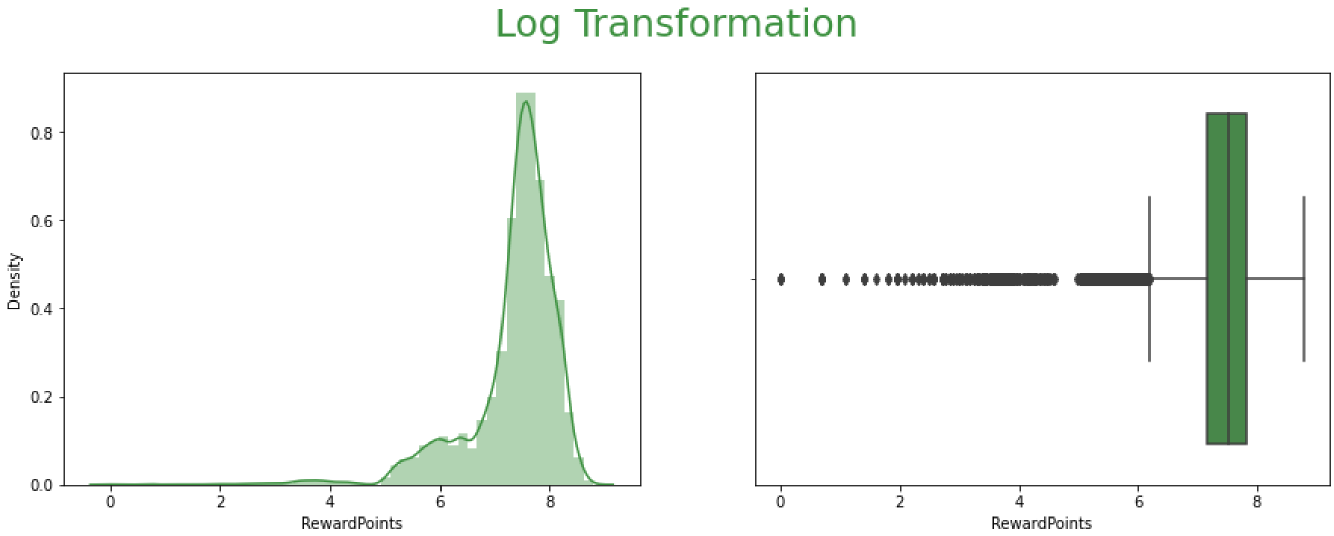
Graph 29; Customers Gender


Client's Reward Points

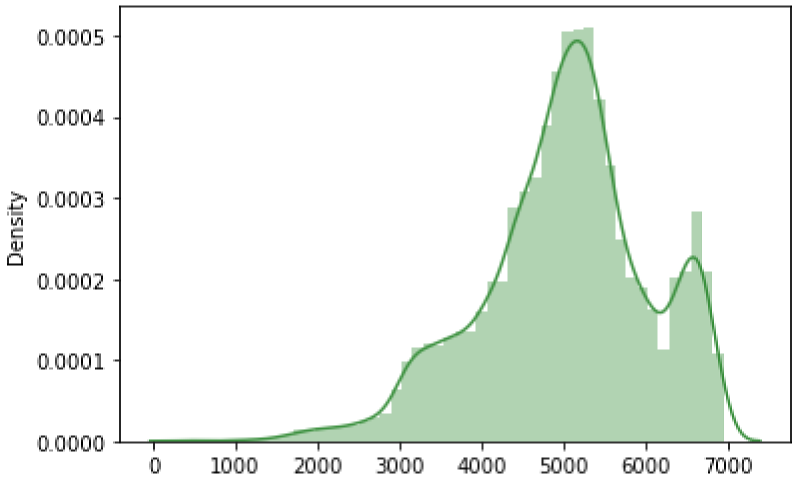
Graph 28: Clients Reward Points Boxplot


Cube Transformation

Graph 31: Cube Transformation Method

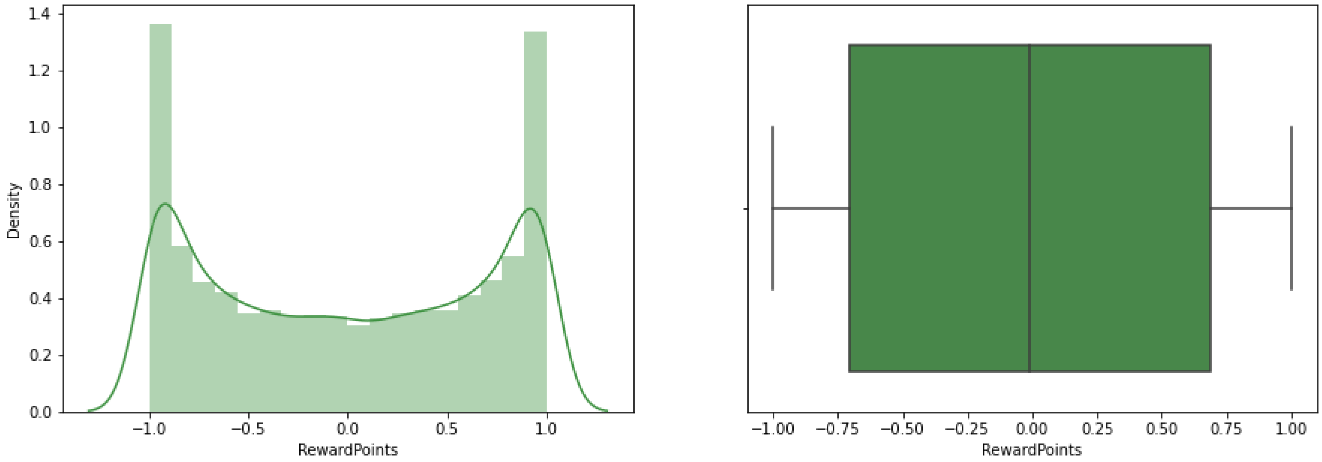

Log Transformation

Graph 30: Log Transformation Method

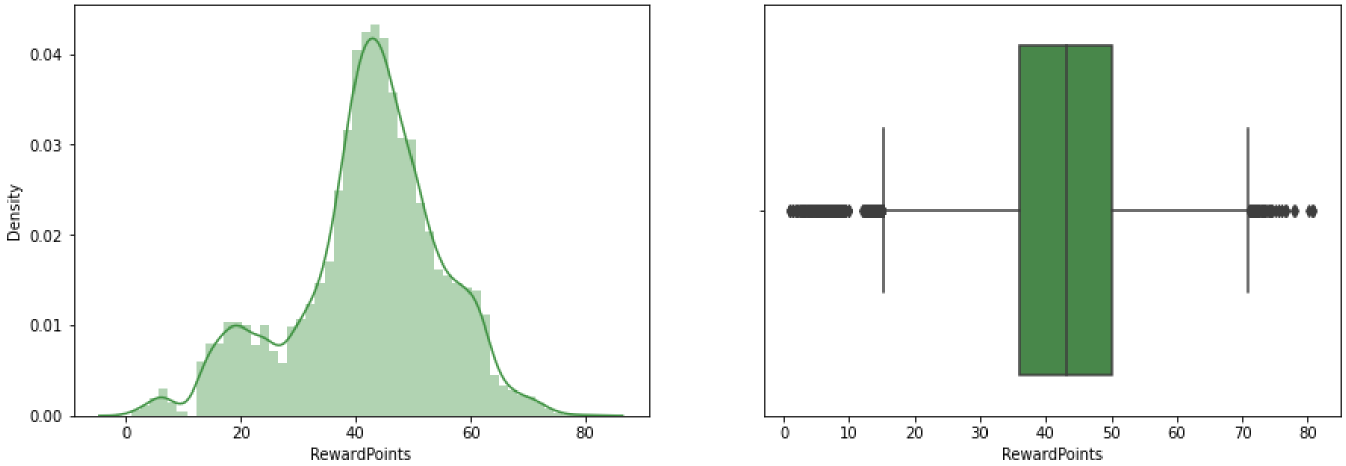


Graph 32: Original Graph Reward Points

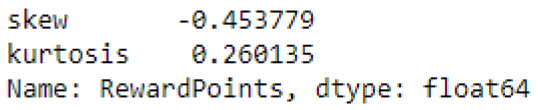
Sine Transformation



Graph 33: Sine Transformation Method

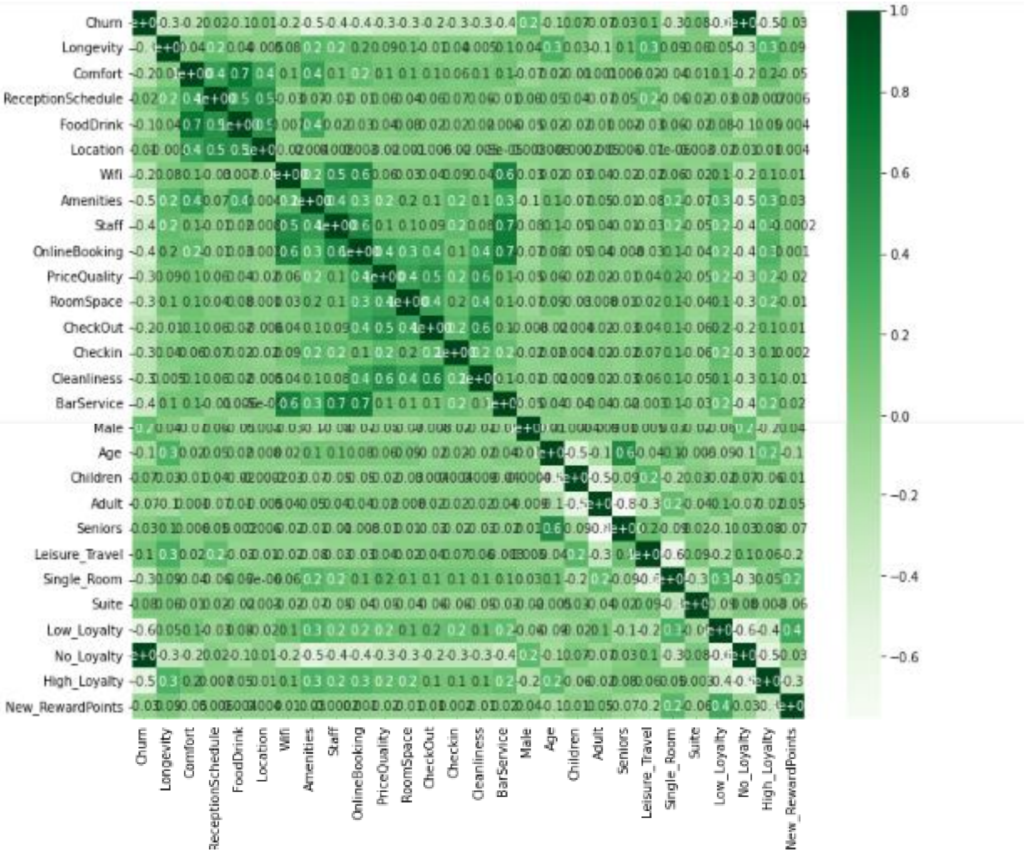


Graph 34: Square Root Method

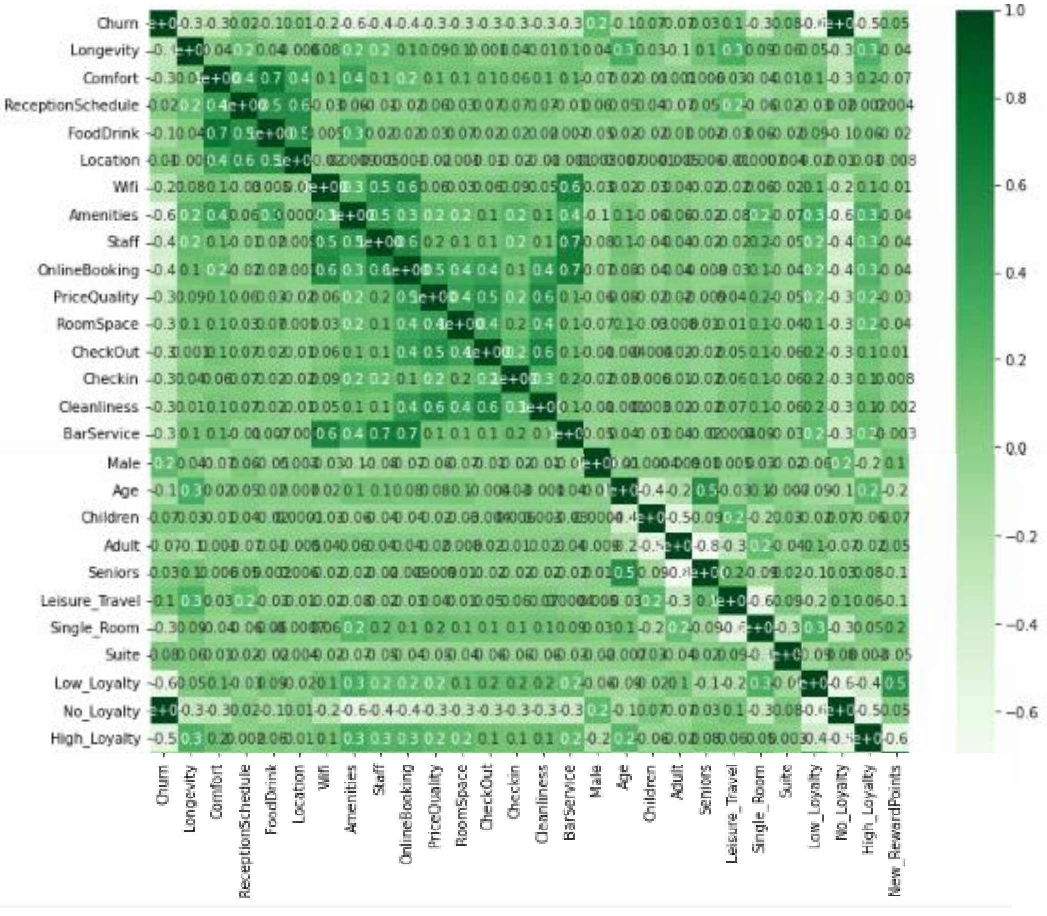


```
skew        -0.453779
kurtosis     0.260135
Name: RewardPoints, dtype: float64
```

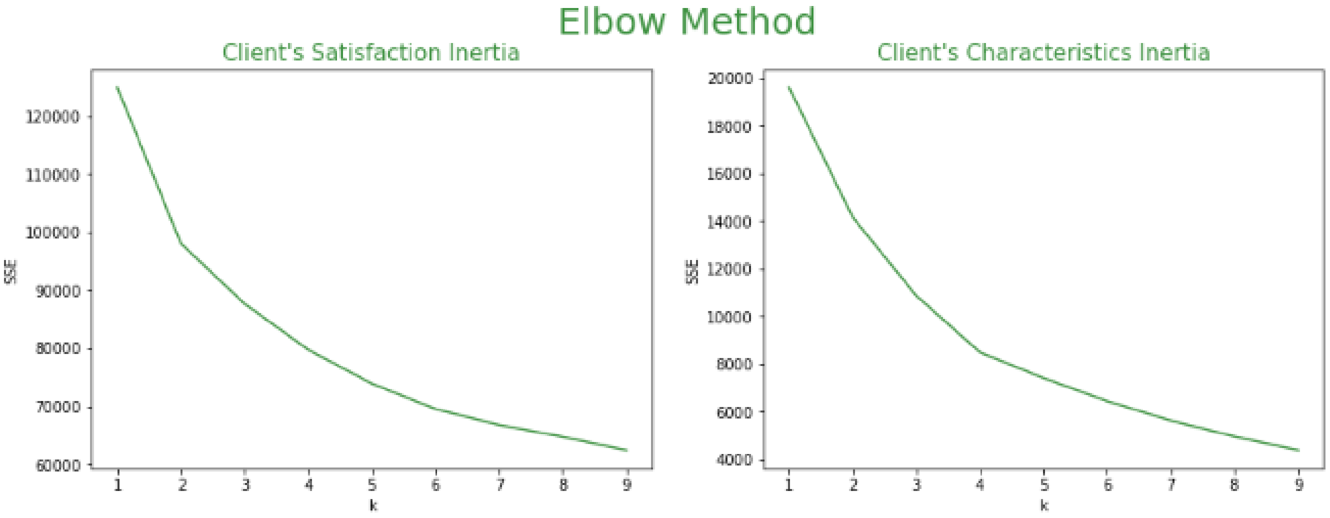
Graph 35: Skewness of the Dataset



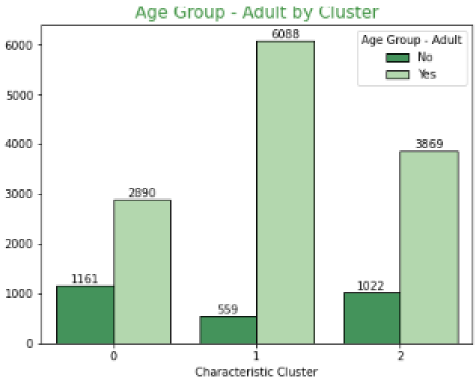
Graph 37: Pearson Method



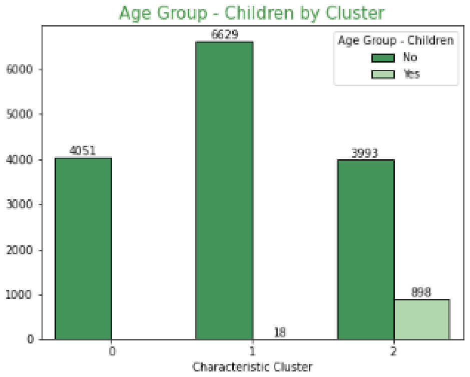
Graph 36: Spearman Method



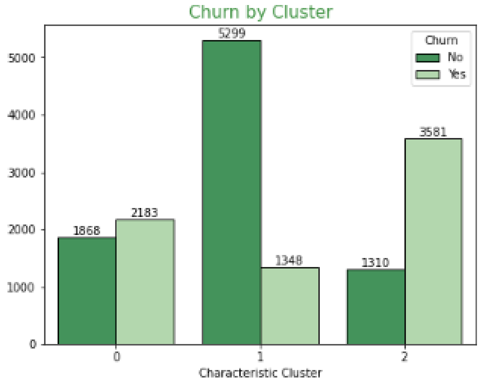
Graph 47: Elbow Method (Characteristics and Satisfaction Customers)



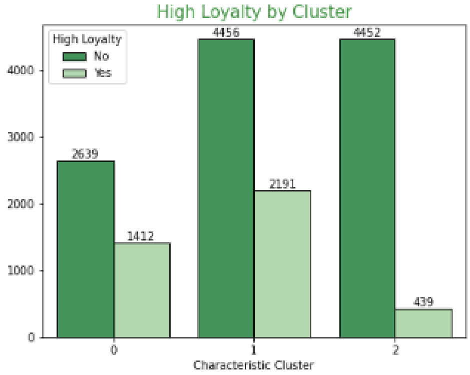
Graph 43 Adult Group Characteristics (KNN Modelling Clustering)



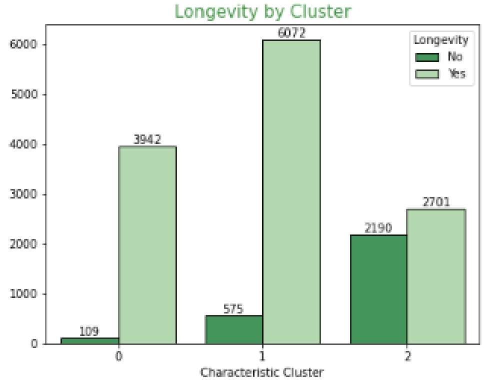
Graph 44 Children Group Characteristics (KNN Modelling Clustering)



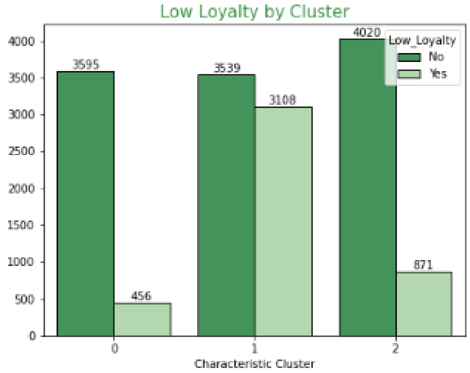
Graph 45: Churn Characteristics (KNN Modelling Clustering)



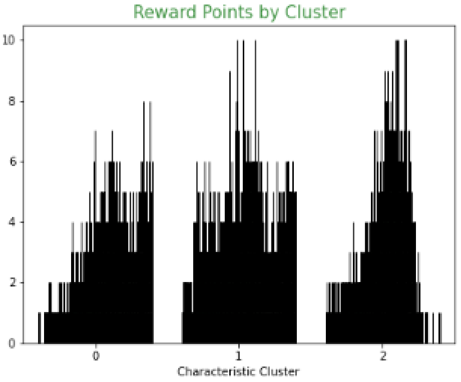
Graph 42: High Loyalty Characteristics (KNN Modelling Clustering)



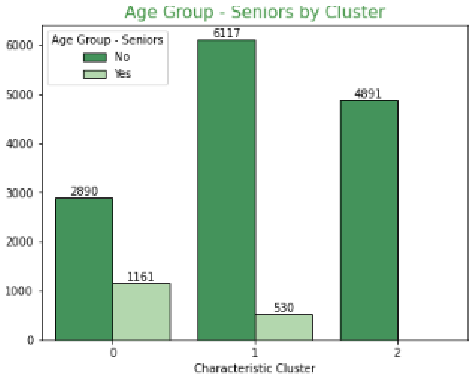
Graph 41: Longevity Characteristics (KNN Modelling Clustering)



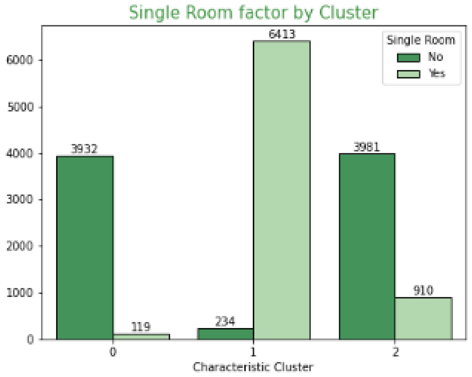
Graph 40: Low Loyalty Characteristics (KNN Modelling Clustering)



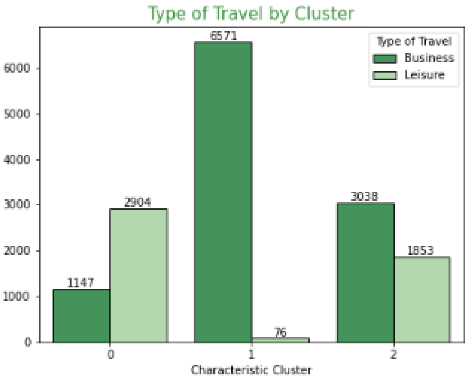
Graph 46: Reward Points Characteristics (KNN Modelling Clustering)



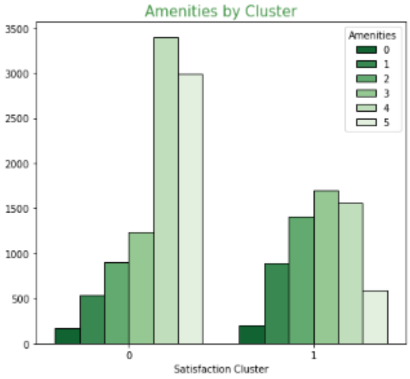
Graph 39: Senior Group Characteristics (KNN Modelling Clustering)



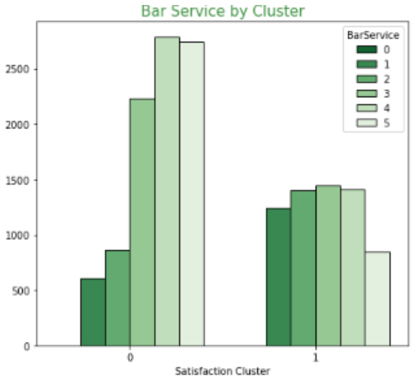
Graph 38: Single Room Characteristics (KNN Modelling Clustering)



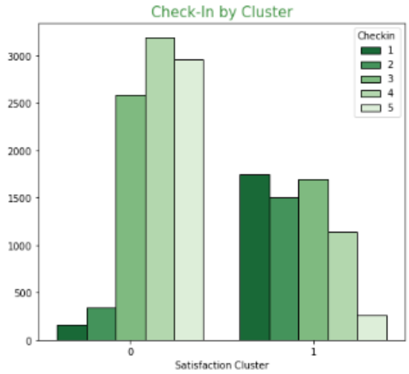
Graph 59: Type of Travel Characteristics (KNN Modelling Clustering)



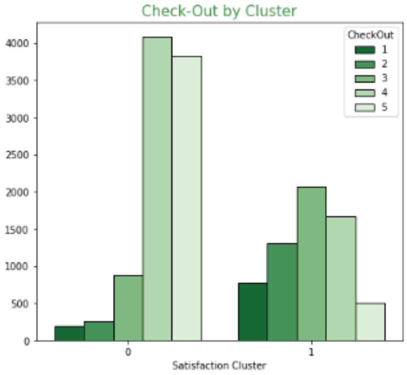
Graph 58: Amenities Satisfaction (KNN Modelling Clustering)



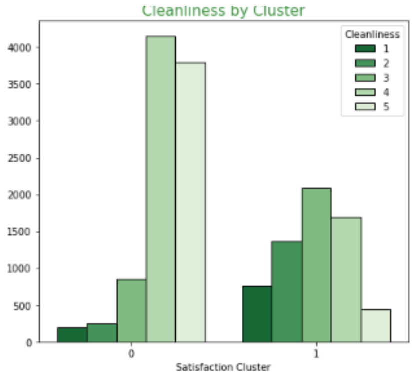
Graph 57: Bar Service Satisfaction (KNN Modelling Clustering)



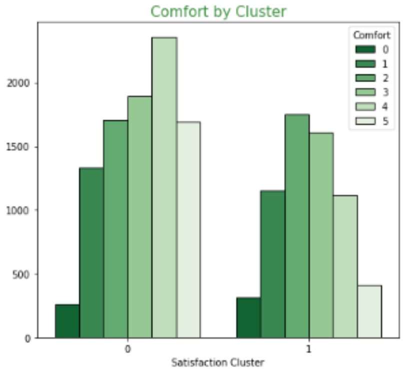
Graph 56: Check-In Satisfaction (KNN Modelling Clustering)



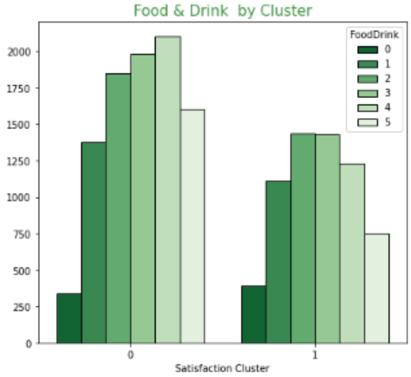
Graph 55: Check-Out Satisfaction (KNN Modelling Clustering)



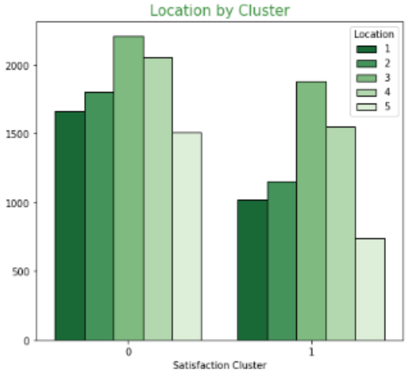
Graph 54: Cleanliness Satisfaction (KNN Modelling Clustering)



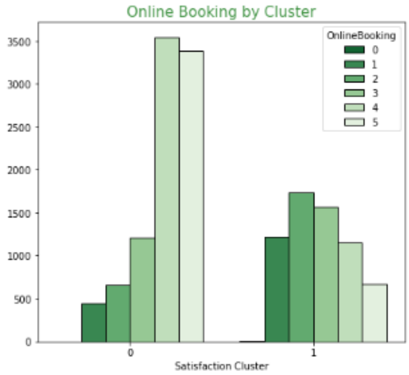
Graph 53: Comfort Satisfaction (KNN Modelling Clustering)



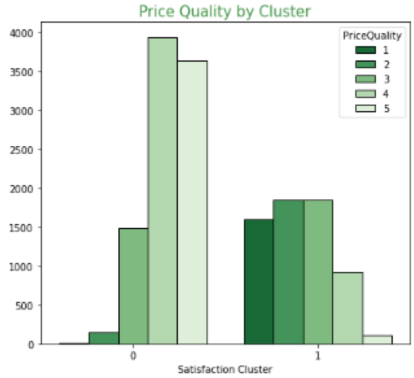
Graph 52: Food & Drink Satisfaction (KNN Modelling Clustering)



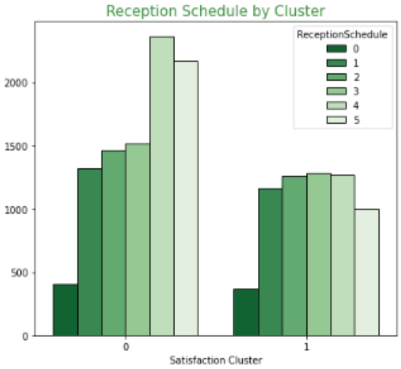
Graph 51: Location Satisfaction (KNN Modelling Clustering)



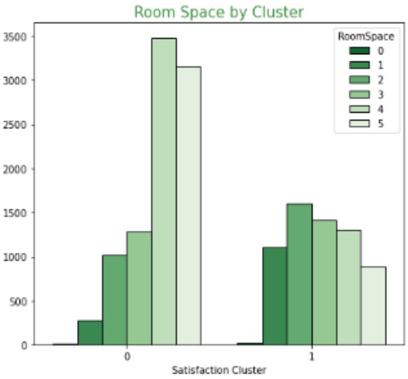
Graph 50: Online Booking Satisfaction (KNN Modelling Clustering)



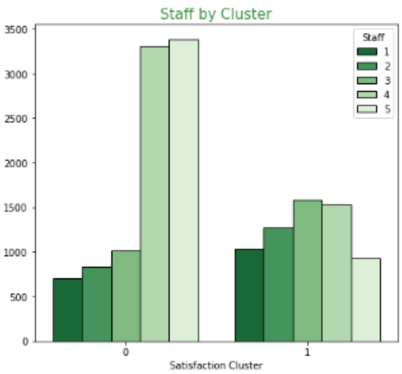
Graph 49: Price Quality Satisfaction (KNN Modelling Clustering)



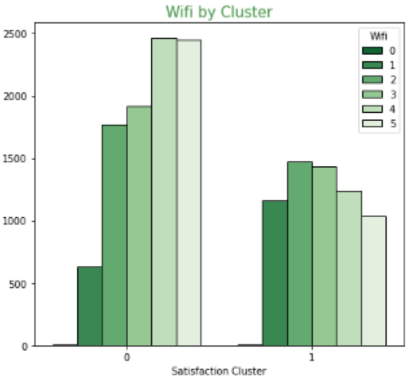
Graph 48: Reception Schedule Satisfaction (KNN Modelling Clustering)



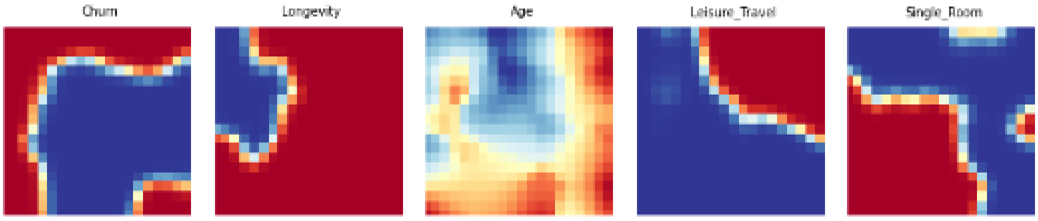
Graph 61: Room Space Satisfaction (KNN Modelling Clustering)



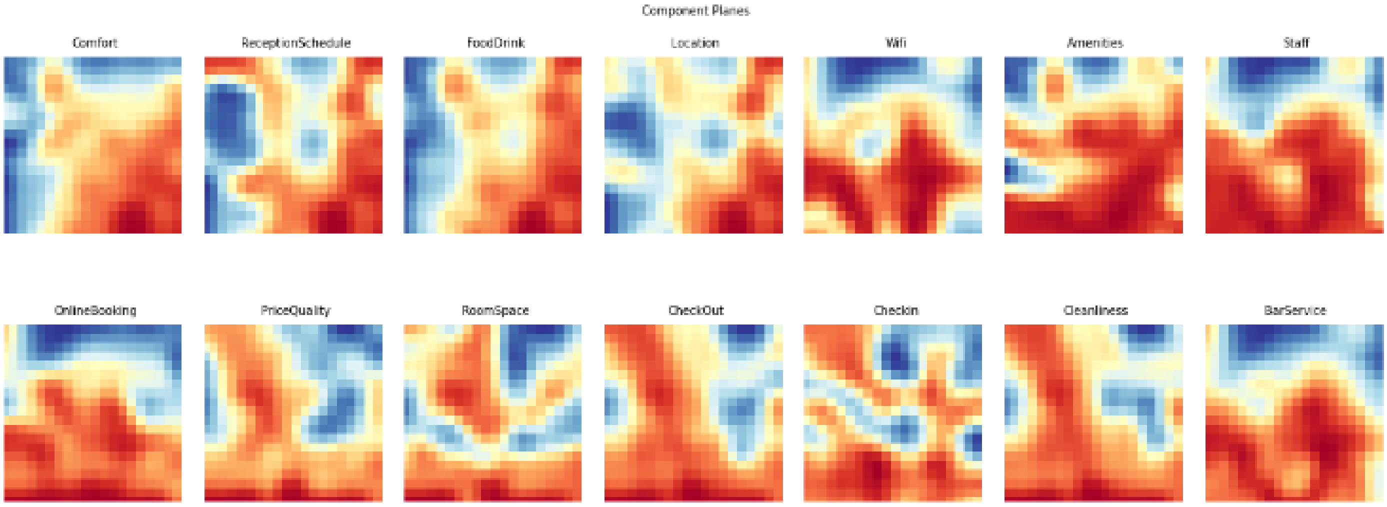
Graph 62: Staff Satisfaction (KNN Modelling Clustering)



Graph 63: Wifi Satisfaction (KNN Modelling Clustering)



Graph 60: Components Plane Characteristics (SOM)



Graph 64: Components Plane Satisfaction (SOM)

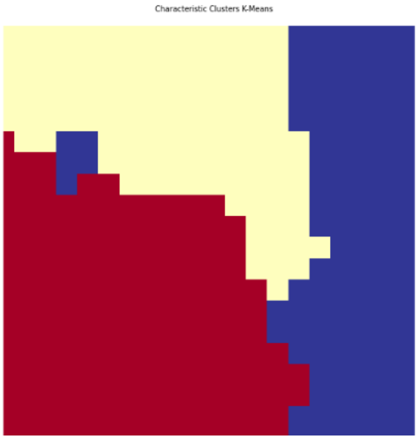


Figure 4:K-Means Over SOM Characteristic

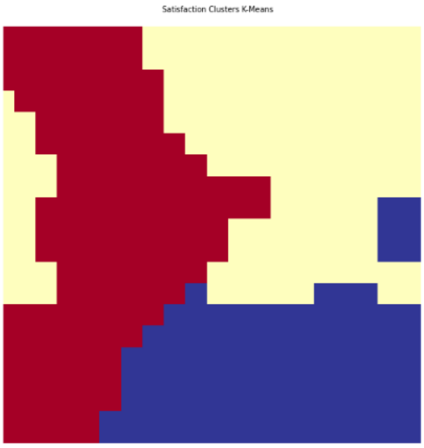


Figure 3: K-Means Over SOM Satisfaction

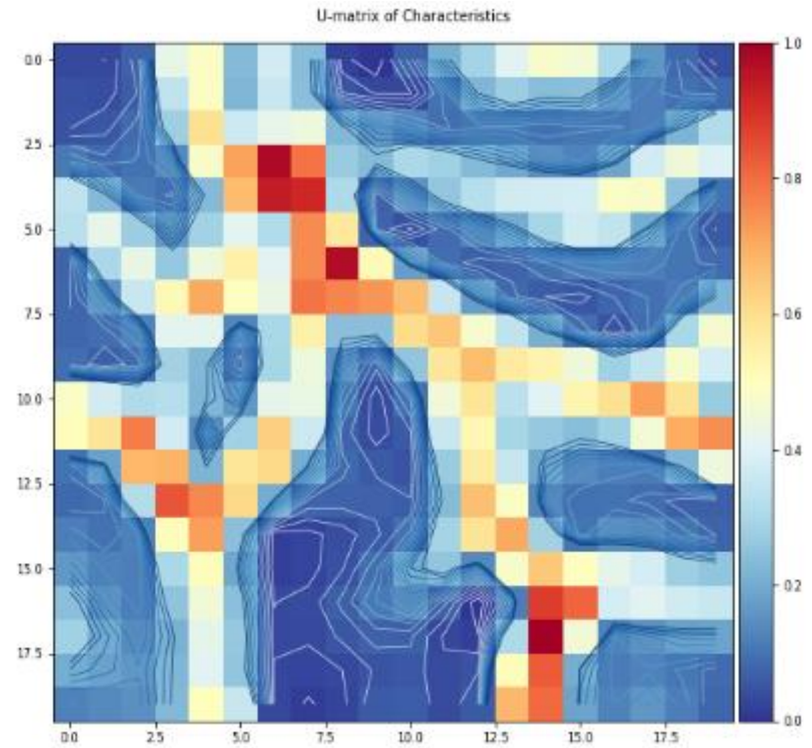


Figure 5: U-Matrix Characteristics

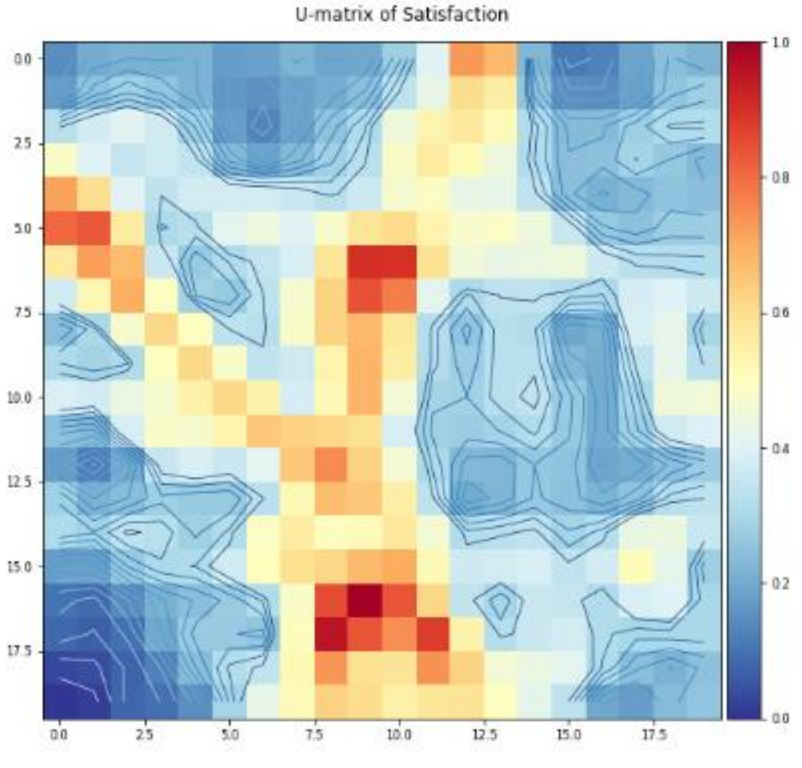
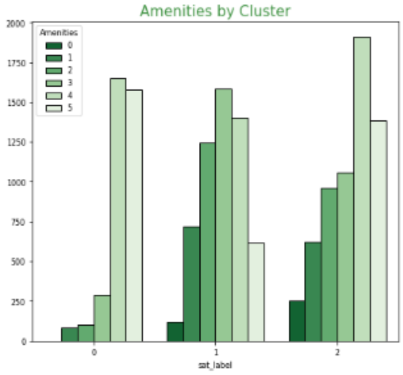
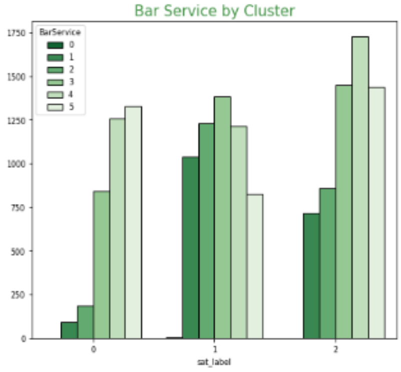


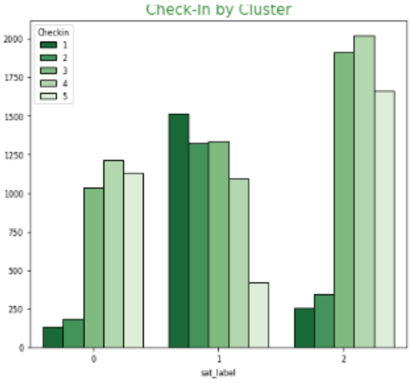
Figure 6: U-Matrix Satisfaction



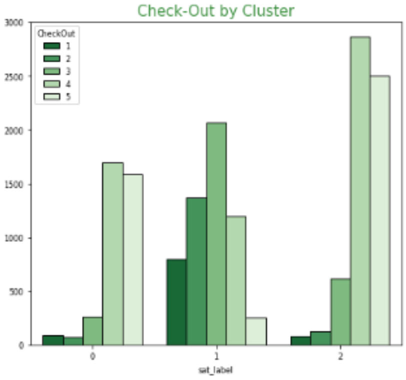
Graph 72: Amenities Satisfaction (SOM)



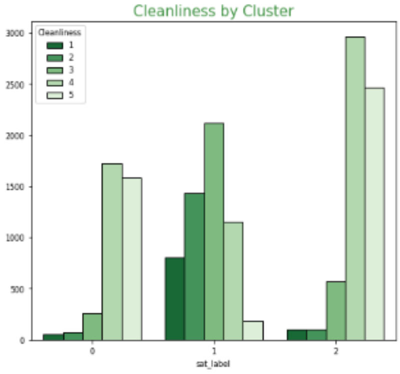
Graph 71: Bar Service Satisfaction (SOM)



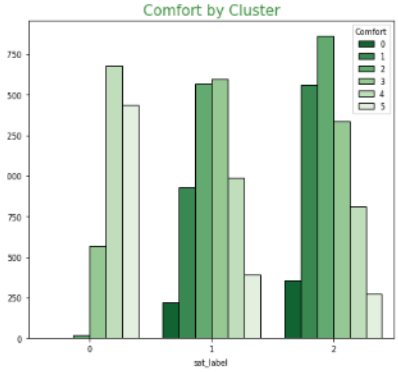
Graph 67:Check-In Satisfaction (SOM)



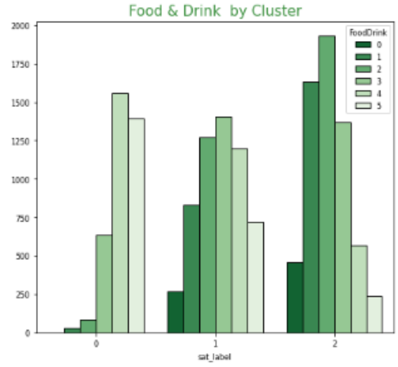
Graph 70: Check- Out Satisfaction (SOM)



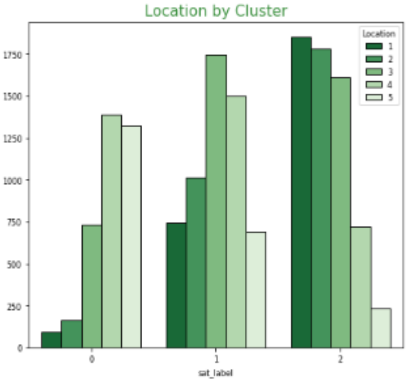
Graph 65: Cleanliness Satisfaction (SOM)



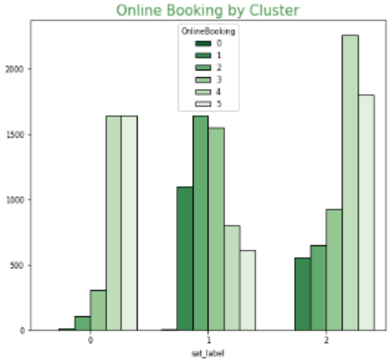
Graph 66: Comfort Satisfaction (SOM)



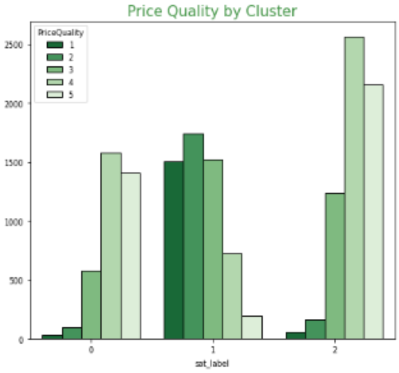
Graph 69: Food and Drink Satisfaction (SOM)



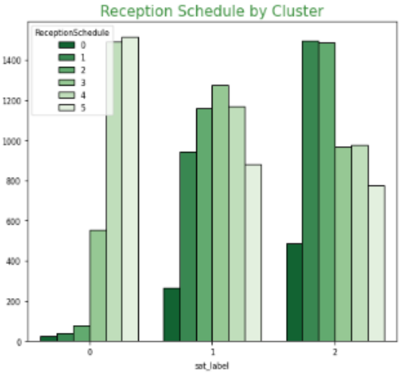
Graph 68: Location Satisfaction (SOM)



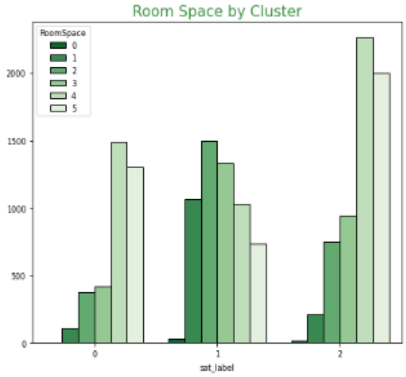
Graph 73: Online Booking Satisfaction (SOM)



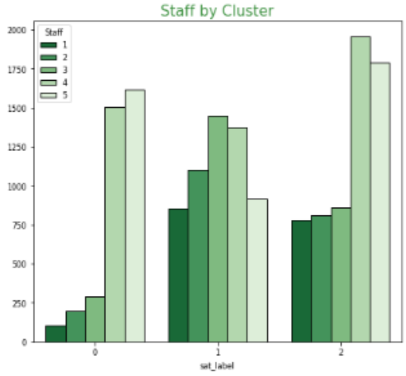
Graph 76: Price Quality Satisfaction (SOM)



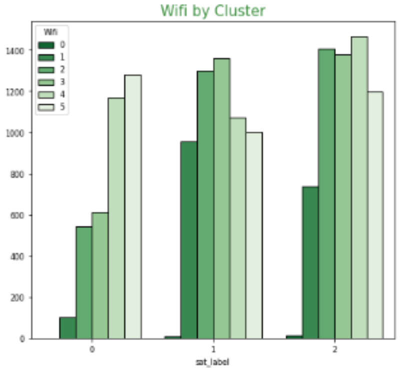
Graph 75: Reception Schedule Satisfaction (SOM)



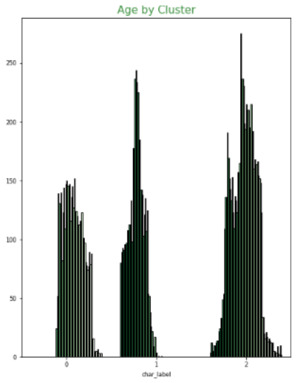
Graph 74: Room Space Satisfaction (SOM)



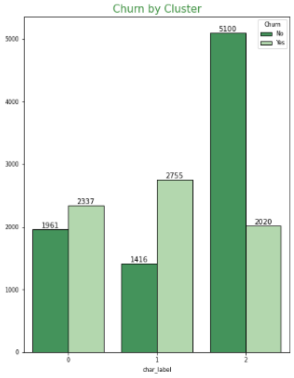
Graph 78: Staff Satisfaction (SOM)



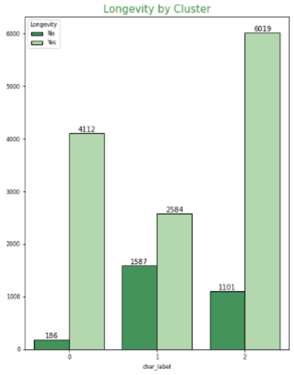
Graph 79: Wifi Satisfaction (SOM)



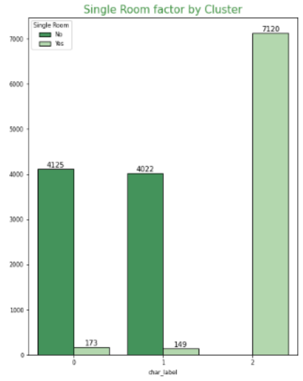
Graph 77: Age Characteristics (SOM)



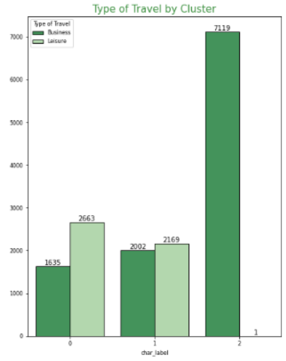
Graph 82: Churn Characteristics (SOM)



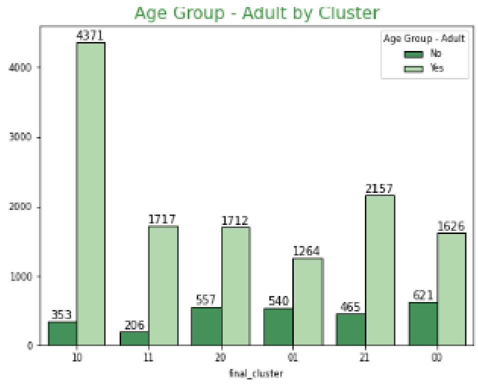
Graph 83: Longevity Characteristics (SOM)



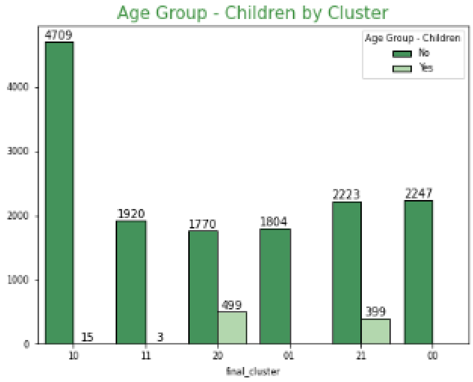
Graph 84: Single Room Characteristics (SOM)



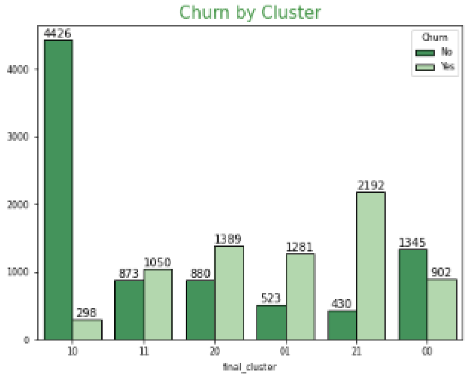
Graph 85: Type of Travel Characteristics (SOM)



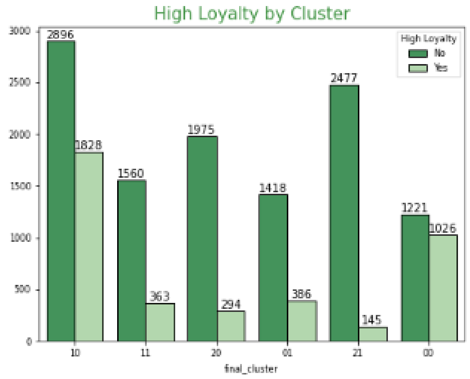
Graph 86: Adult Group Characteristics (Final Model)



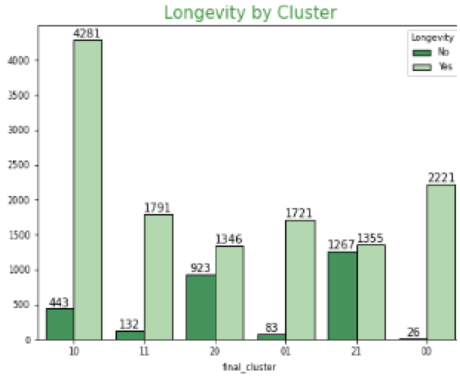
Graph 87: Children Group Characteristics (Final Model)



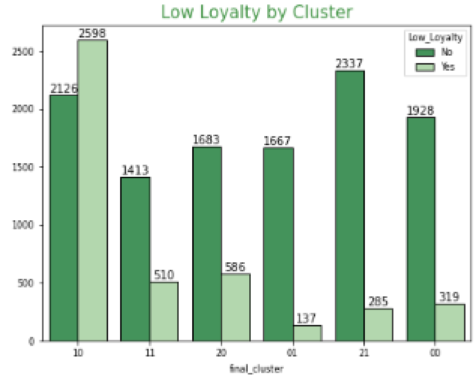
Graph 88: Churn Characteristics (Final Model)



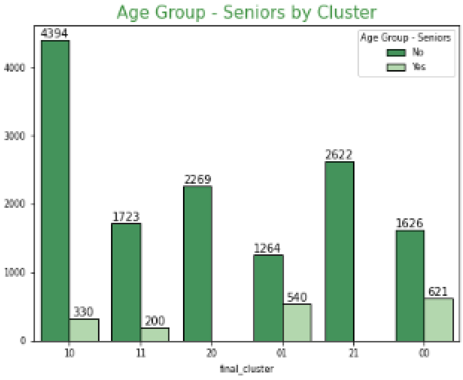
Graph 89: High Loyalty Characteristics (Final Model)



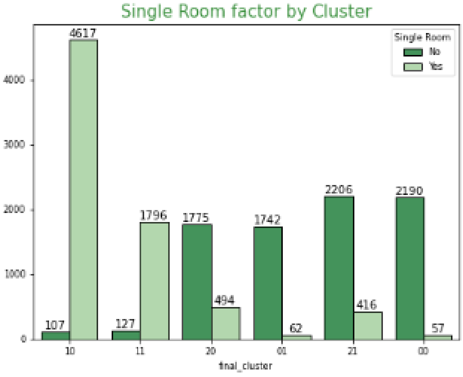
Graph 81: Longevity Characteristics (Final Mod



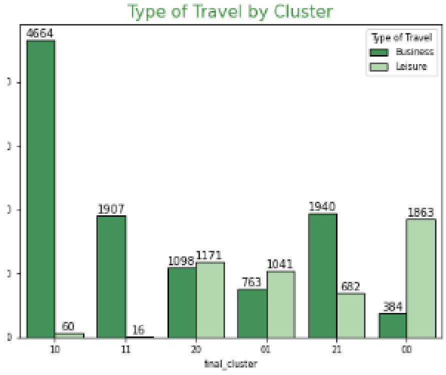
Graph 80: Low Loyalty Characteristics (Final Model)



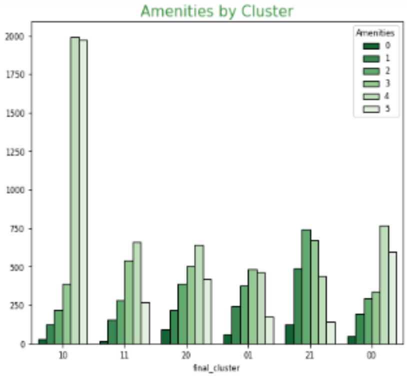
Graph 98: Seniors Group Characteristics (Final Model)



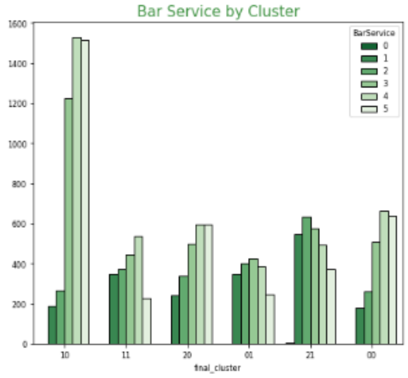
Graph 97: Single Room Characteristics (Final Model)



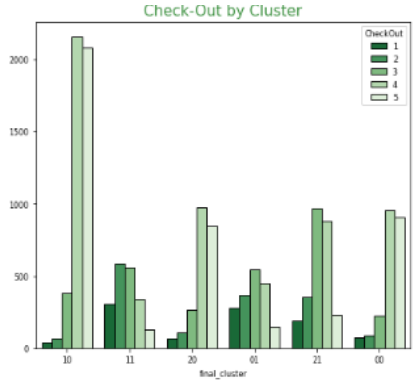
Graph 90: Type of Travel Characteristics (Final Model)



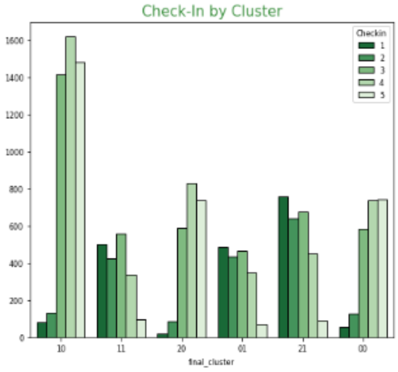
Graph 96: Amenities Satisfaction (Final Model)



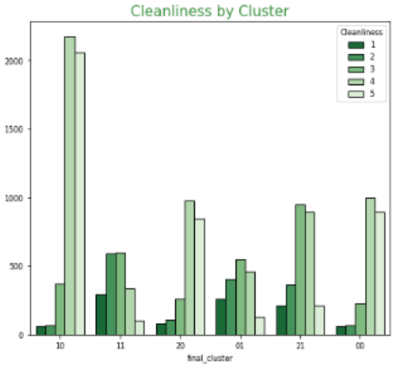
Graph 95: Bar Service Satisfaction (Final Model)



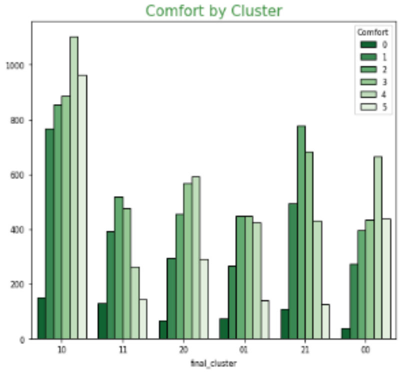
Graph 94: Check-Out Satisfaction (Final Model)



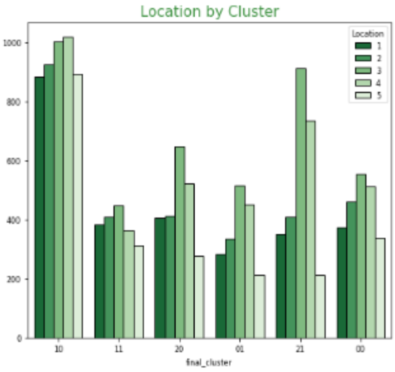
Graph 93: Check-In Satisfaction (Final Model)



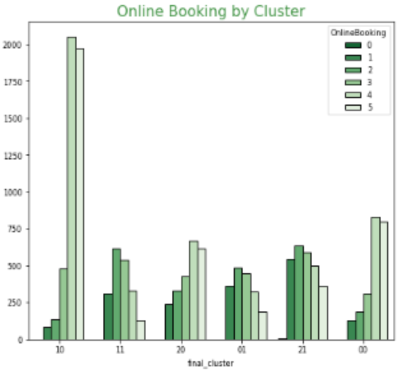
Graph 92: Cleanliness Satisfaction (Final Model)



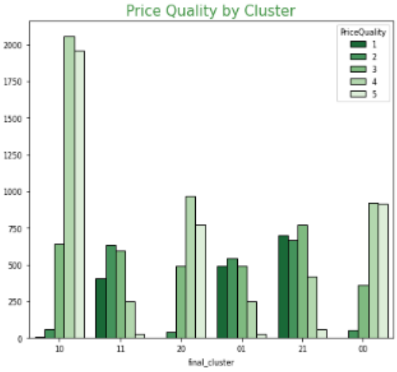
Graph 91: Comfort Satisfaction (Final Model)



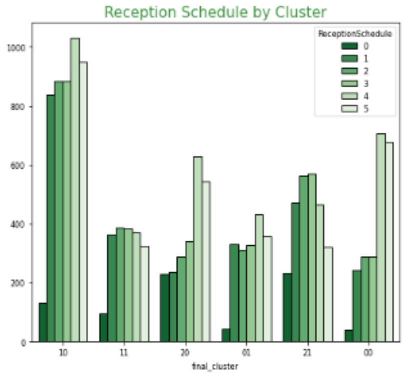
Graph 101: Location Satisfaction (Final Model)



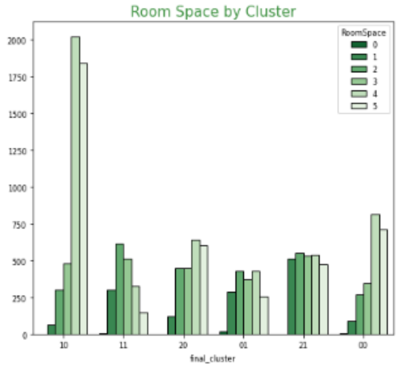
Graph 100: Online Booking Satisfaction (Final Model)



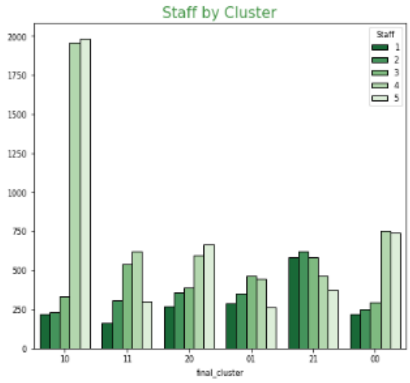
Graph 99: Price-Quality Satisfaction (Final Model)



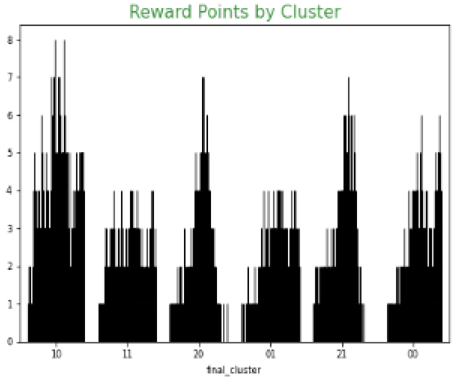
Graph 104: Reception Schedule Satisfaction (Final Model)



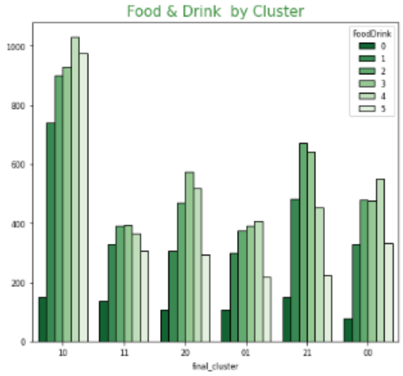
Graph 105: Room Space Satisfaction (Final Model)



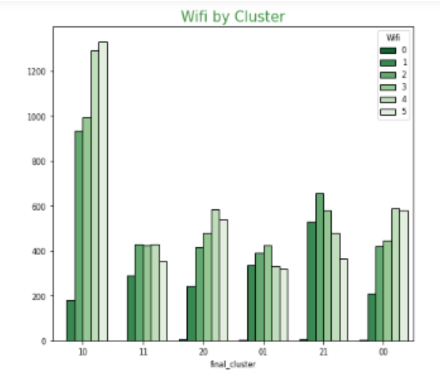
Graph 106: Staff Satisfaction (Final Model)



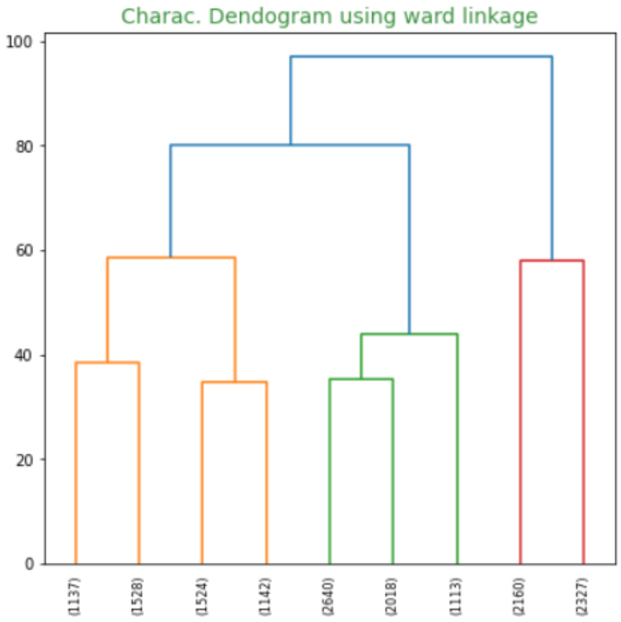
Graph 103: Reward Points Characteristics (Final Model)



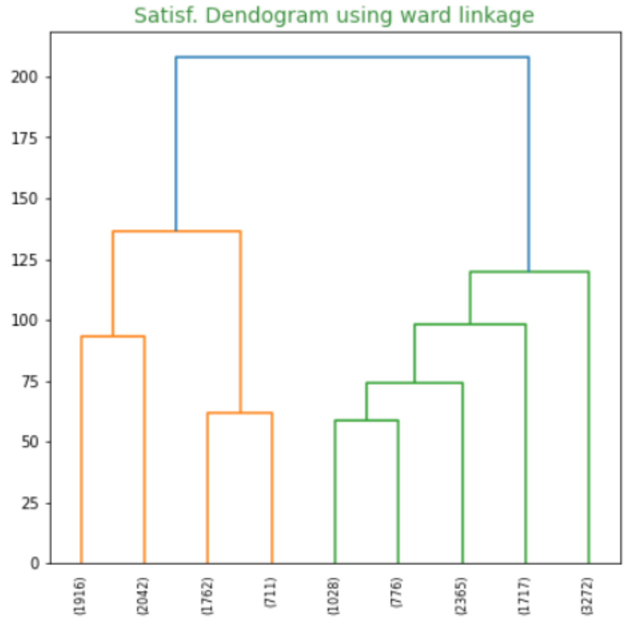
Graph 102: Food and Drink Satisfaction (Final Model)



Graph 107: Wifi Satisfaction (Final Model)



Graph 108: Characteristics Dendrogram, ward linkage



Graph 109: Satisfaction Dendrogram, ward linkage