

**Course Name:** BC2407 Analytics II

AY22/23 (Semester 2)

**Seminar Group 4**

**Team 7**

**Students’ Names and Matriculation Numbers:**

| Phyllis Yen Chi Hsuen | U2110949L |
| --- | --- |
| Khoo Yu Zhen | U2110948J |
| Grace Lim Wan Yu | U2010635G |
| Brennan Tay Yu Zhe | U2110523D |
| Tan Yi Hui Brenda | U2110207C |

**Table of Contents**

[**1. Executive Summary 4**](#_rnla2mhlxtao)

[**2. Business Problem 5**](#_on0hiam3xpnh)

[**3. Existing Solutions 6**](#_ljqapk15o2qe)

[3.1 Revenue Optimising System 6](#_7omjdjn4qcbq)

[3.2 Advanced Analytics in Hospitality 6](#_3yjnfedhharj)

[3.3 Opera Cloud System 6](#_pav5wxr7lm69)

[**4. Desired Outcomes 7**](#_5ww7p438sw83)

[4.1 Improved Guest Management 7](#_wp0n7rybmcst)

[4.2 Decreased Overheads 7](#_rw6n7ft2fge)

[4.3 Improved Staff Retention 7](#_3j8vngqp1ja3)

[4.4 Optimised Pricing Strategy 7](#_zez10ni6jpfx)

[**5. Data Description and Cleaning 8**](#_j2q2ljbfbsx8)

[5.1 Data Sources 8](#_li1bamuqbcpt)

[5.2 Data Cleaning 8](#_hfjmhbjl46zz)

[5.2.1 Customer Dataset 8](#_jmjgto1t31h2)

[5.2.2 Hotel Booking Demand Dataset 10](#_eh5gmcdpaioq)

[**6. Data Exploration 12**](#_kdeb3oqclr2e)

[6.1. Customer Data Set 12](#_5o8tb286p8i6)

[6.2. Hotel Booking Demands Data Set 13](#_x0h1jwmzc8q5)

[6.2.1. Cancellation Attributes 14](#_e4564ib2wi3y)

[**7. Data Modelling 16**](#_7f24knefe7o4)

[7.1 Customer Segmentation with Clustering 16](#_8l2v0179rh6v)

[7.1.1 Optimal Cluster 16](#_u18sm6tfxece)

[7.1.2 Clustering Results 16](#_3jn11cpnsq4j)

[7.2 Cancellation Modelling with Logistic Regression 17](#_f3o800jw3agh)

[7.3 Cancellation Modelling with CART 18](#_vr98r3lgcqhi)

[7.4 Cancellation Modelling with Random Forest 18](#_jeu5enqc0wzk)

[7.6 Model Evaluation 19](#_yjzbz2s75phf)

[7.6 Model Implementation 20](#_t3a1f6c7ch3c)

[7.6.1 Logistic Regression 20](#_njbswos52ydb)

[7.6.2 CART 20](#_eclsmol2unl2)

[7.6.3 Random Forest 21](#_cjsavmexotbd)

[**8. Solutions 22**](#_5utnsc14frzj)

[8.1 Overbooking 22](#_w4csrhrg49fg)

[8.2 Targeted Marketing Strategies 22](#_7lg0arjrob1m)

[8.3 Pricing Strategies and Policies 23](#_hj9dyk2bzxsi)

[8.4 Staff Rostering 24](#_lfh8zwk9tuij)

[**9. References 25**](#_vnl2w7sa13yr)

[**10. Appendices 27**](#_j1k263qcgody)

[**11. Data Dictionary 52**](#_ij71e2kbj010)

[11.1 Hotel Bookings Demand Dataset 52](#_h3enaqvgbkjv)

[11.2 Customer Booking Dataset 58](#_krvgjplm8g03)

# 

# 

# **Executive Summary**

The global hotel and resort industry, worth USD 1.52 trillion pre-pandemic, is projected to grow at a compounded annual growth rate of 4.28% over the next 5 years. However, rising costs, high staff turnover, and last-minute cancellations remain major concerns for hoteliers. Inflation has caused overhead costs to increase by 50%, and the leisure and hospitality industry has the highest staff turnover rate, which can negatively impact productivity. Moreover, even with travel restrictions and safe distancing measures lifted, hoteliers may not be able to fully capitalise on the travel boom due to operational shortfalls. Lastly, last-minute cancellations cause significant financial losses for hoteliers, making it crucial for them to anticipate guest demand and arrival to improve cost efficiency, operating efficiency, and customer service.

There are currently various solutions available in the hospitality industry, such as revenue optimising systems, advanced analytics, and the opera cloud system. However, these solutions lack the ability to analyse customer segmentation to deliver a more refined personalised experience, as well as the ability to predict cancellation rates. Therefore, there is a growing need to develop a predictive model to address this gap in the industry.

Our team hopes to use analytics to tackle these issues faced by hoteliers through implementing preventive and reactive measures. In doing so, 2 datasets have been employed to facilitate our analysis. (1) The Hotel Customer dataset includes three full years' worth of demographic, personal, behavioural, and geographic data for a hotel in Lisbon, Portugal, from 2015 to 2018. Due to the close proximity of the years to the current year (2023), the dataset is still applicable in the present. It will be used to help the hotelier understand their customer demographic and behaviour through the use of K-means clustering. This will assist our team in creating preventive measures for customer cancellation. (2) The Hotel Booking dataset contains booking information for a city hotel and resort hotel. The dataset has a total of 31 features, including information such as the date when the booking took place, length of stay, the number of adults and children, the number of available parking spaces, distributional channels, and many others. It will be used to conduct logistic regression, CART and random forest to predict likelihood of cancellation by customers. Findings from the above mentioned models are then used to create reactive measures to tackle cancellations.

By the end of the report, the following 4 solutions will be introduced for the hotelier’s consideration. Firstly, the practice of overbooking which can help reduce opportunity cost in the event of a last-minute cancellation. Secondly, the introduction of differentiated marketing strategies to target its 5 main customer segments. Thirdly, the use of pricing strategies to capitalise on seasons where there may be higher demand. Lastly, the implementation of staff rostering according to peak periods to ensure maximum operational capacity throughout the year.

# Business Problem

Hotels play an integral role in our modern-day economy. Prior to the pandemic in 2019, the global market size for hotels and resorts peaked at USD 1.52 trillion (Statista, 2022). Despite the toll of Covid-19 on the tourism and hotel industry, analysts forecast a compounded annual growth rate (CAGR) of 4.28% for hotel revenue over the next 5 years (Statista, 2022). On the surface, it may seem like the road ahead for hoteliers is optimistic. Yet, new obstacles await as rising costs, high staff turnover rates and an influx of customers swarm the industry.

Hotels generally have high fixed and variable costs due to the large amount of space and manpower required to carry out their operations. Given the current cost of living crisis, rising inflation is reported to increase their overheads by 50%, making operations more costly than pre-pandemic periods (The Drinks Business, 2022). Hence, there is a dire need for hotels to better manage their costs to improve their bottom-line even as their revenues increase. Another pertinent issue plaguing the industry would be its high staff turnover. The leisure and hospitality industry was reported to have the highest staff turnover in 2019, 2020 and 2021, with a whopping average of 98% (Award, n.d.). Even with staffing numbers nearing pre-pandemic levels, staff retention still remains a challenge for most hoteliers worldwide. This is concerning as hotel employees form the backbone of the industry by supporting daily operations and providing quality service. The lack of a competent pool of staff will result in decreased productivity which will compound cost issues. Thus, there is also a need to improve staff retention in hotels. The final issue pertains to the potential travel boom that hotels will have to cope with as the industry recovers from the toll of Covid-19. With travel restrictions and safe distancing measures lifted, hotels are likely to experience a spike in demand for their services. Considering the earlier two issues related to cost and manpower, hoteliers may not be able to fully capitalise on the travel boom due to operational shortfalls, which could negatively impact their top line. Additionally, even as hotels aim for a high volume of bookings, these may not always translate to a greater number of guests due to the flexibility of cancellation policies. Last minute cancellations are undesirable for hoteliers as it leaves them with more unoccupied rooms leading to a loss of income. With the increase in the rate of hotel reservation cancellations and no-shows, this trend is troublesome because it is causing significant financial losses (Duetto, 2016; Pederson, 2018).

With the above mentioned points in mind, our team finds that there is a need for hoteliers to better anticipate guest demand and arrival, so as to improve their cost efficiency, operating efficiency and customer service. Therefore, this report will expound more on the factors affecting demand for hotel rooms and examine possible reasons that could give rise to booking cancellations.

# **Existing Solutions**

## 3.1 Revenue Optimising System

Data Analytics has been successfully integrated in Revenue Management for InterContinental Hotel Group (IHG), who co-developed Concerto, a guest-reservation platform system to increase revenue and gather more data for decision-making. Concerto allows hoteliers to drive direct bookings by providing personalised offers based on past reservations and suggesting a wider set of booking options.

## 3.2 Advanced Analytics in Hospitality

Hospitality companies have deployed predictive analytics to better anticipate and meet customer needs and preferences. In 2013, the US economy-hotel chain Red Roof Inn used public weather and flight data to predict which customers would face flight cancellations. Based on the results of this predictive analysis, Red Roof Inn launched a targeted marketing campaign aimed at mobile-device users in the areas most likely to be affected by harsh weather. In those areas where the strategy was deployed, Red Roof Inn saw a significant increase in business. With advances in artificial intelligence and predictive analytics, hospitality companies have been able to create unique offers and experiences in real time that appeal to the needs and desires of each individual traveller. For instance, advances in machine learning have improved hotels’ ability to optimise pricing through more accurate analyses and predictions based on market demand signals, local room availability, and a deep understanding of the individual customer’s willingness to pay. At the same time, advances in predictive analytics will drive improvements in forecasting. Right now, companies throughout the value chain, from global hotel brands to technology start-ups, are using self-learning algorithms that incorporate historical data on millions of searches to predict future price movements based on multiple factors, including seasonal trends, demand growth, and limited-time special offers, as well as consumer preferences and purchase patterns. As demand forecasts become more accurate, pricing and yield strategies can become more sophisticated to capture greater value.

## 3.3 Opera Cloud System

Opera Cloud System, run by Oracle, provides hotels a one-stop cloud-based solution to help them manage day-to-day activities from guest management to room availability and prices. Presently, Opera is used by both customer-facing and back-end teams which seamlessly facilitates coordination between all departments. However the system takes a reactive approach to managing guest cancellation which limits its efficacy in predicting guest demand.

# **Desired Outcomes**

## 4.1 Improved Guest Management

Hoteliers should be able to predict the volume and type of customers they are expected to receive during different periods of the year. For instance during peak vacation periods, hotels may expect more families with children to book their services. This indicates that facilities such as the pool and playground will have to be cleaned more often, and rooms which are bigger in size may be in higher demand. With this knowledge, hoteliers can then set aside rooms that its guests are more likely to book, and allocate sufficient manpower to tend to its guests.

Additionally, hoteliers should also be able to minimise booking cancellations by forecasting the likelihood of a customer backing out of their reservation. This includes preventive measures such as adopting the right marketing strategy, and reactive measures like overbooking a room to reduce opportunity cost.

## 4.2 Decreased Overheads

Since variable costs change with demand and are harder to reduce, hoteliers should aim to minimise their fixed costs. For instance, payroll expenses are usually high as staff are scheduled to work on days where they are not needed. By using data to forecast demand, it can help hoteliers to estimate the amount of manpower needed to bear a given workload. This maximises the operational capacity of staff and ensures cost efficiency for the hotel.

## 4.3 Improved Staff Retention

Ideally, employee turnover rates faced by the hotel industry should be reduced by half over the next 5 years. On average, companies experience a 18% turnover in its workforce every year. The hotel industry is currently five-fold of the average, hence addressing structural issues could help to improve this figure significantly.

## 4.4 Optimised Pricing Strategy

With the ease of online bookings and different channels of reservations, hotels often find themselves mitigating operational changes, such as changes to bookings and cancellations. Existing implementations by hotels include their deposit and cancellation policies as well as dynamic pricing, which aim to optimise their revenue. By utilising customer data to segment and model customer behaviour, it enables hotels to mitigate their losses.

# **5. Data Description and Cleaning**

## 5.1 Data Sources

To tackle the business problem holistically, the team scoured Kaggle to obtain 2 key datasets that will help to enrich our analysis. For both datasets, it is important to note that each row of data only has information on 1 person - the organiser who makes the reservation. However, each booking can be made for more than 1 person. For instance, Person A can make a single booking for 10 people over a period of 3 days. With this in mind, the number of rows is not an accurate representation of the number of guests a hotel has, rather, one should refer to the number of occupants indicated in the booking.

The first dataset – [**Customer Dataset**](https://www.sciencedirect.com/science/article/pii/S2352340920314645?via%3Dihub) provides a granular view into the customers’ information. It is retrieved from a hotel in Lisbon, Portugal but is relevant in the context of Singapore since travel demand in both countries tend to peak around the same time each year, during the summer vacation period. It comprises 31 attributes which reveal insight on the customer’s nationality and age, as well as their distribution channels. This dataset will be used to conduct clustering for customer segmentation which can help hotels forecast the likelihood of customers booking a reservation, and the services they require based on their segmentation. With improved preparation, our team hopes that the rate of cancellations may also be reduced.

The second dataset – [**Hotel Booking Demand**](https://www.sciencedirect.com/science/article/pii/S2352340918315191) consists of de-identified hotel booking records from a city hotel in Lisbon and a resort hotel in Algarve. With a total of 31 booking attributes, we hope to utilise this dataset to identify relevant factors which contribute to a booking cancellation in a City Hotel, which is largely relevant to Singapore’s context.

Overall, this report will synthesise both datasets to provide a more in-depth and comprehensive study on the subject matter. The data dictionary of the datasets, which details its features, is given under section 11.

## 5.2 Data Cleaning

Before analysis is done on the datasets, they are first cleaned. As each dataset used is different from one other, the cleaning process may vary.

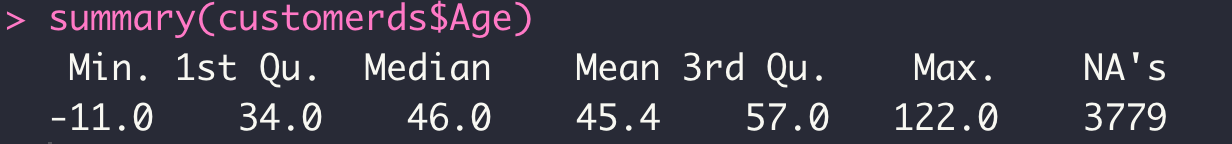
### **5.2.1 Customer Dataset**

The purpose of this dataset is to create customer segmentations within a hotel’s customer base to help the hotelier understand their customer groups. A better understanding of one’s audience could result in better anticipation of guests’ needs and wants which will contribute to better guest management for the hotelier.

1. Removing columns and factoring variables

[ID], [DaysSinceCreation], [DaysSinceLastStay], [NameHash] and [DocIDHash] are irrelevant for our use-case, so they will be dropped. Using str() function, it is observed that categorical variables are not factored accordingly. Hence we use the factor() function to change them where necessary. Using summary() we observed that several categorical SR-prefixed columns occur less than 1% of the time (Appendix 1). Since they do not value-add to our analysis, we will drop them which leave us with [SRHighFloor], [SRCrib], [SRKingSizeBed], [SRTwinBed] and [SRQuietRoom].

1. NA/Invalid/Outlier values



*Figure 1: Summary of Age column*

There are 3779 NAs observed in [Age]. There are also negative values, values less than 18[[1]](#footnote-0) and outliers where age is more than 91. These totals to 10% of the column and could significantly affect data quality. In dealing with this, we opted to not delete these rows as it could result in data loss. As the magnitude of missing data is large, it would not be feasible to individually identify the relevant subgroups, so they are replaced with the median since age can only be an integer and from the histogram (Appendix 3), ages tend to cluster around the 46-54 range.

1. Adding new columns

To assist in our subsequent analysis, we have created the following columns:

[PercOtherRevenue]: shows the percentage of [OtherRevenue] over total revenue. [PercLodgingRevenue]: shows the percentage of [LodgingRevenue] over total revenue. [NumSR]: shows the number of special requests made by a customer.

[CheckInDate]: shows date which customer checked-in by using [DaysSinceFirstStay] and date of data extraction 2018-12-31, we then removed [DaysSinceFirstStay].

[Season]: season of the year which customer checked in. Refer to Appendix 5 for months.

### **5.2.2 Hotel Booking Demand Dataset**

The following modifications were made for the preparation of data exploration. The resulting dataset has 23 features, and 73,756 instances.

1. Factoring variables

sapply() function has been used to find the class of each variable. Some categorical variables have been identified to be not factored accordingly as{stringAsFactors=T} function is used and these variables are stored as numeric/integer values that will not be converted into factors. Thus, factor() function is used to change them when necessary as according to the research paper of this dataset (Appendix 6).

1. Standardising the columns

The [reservation\_status\_date] is in IDate format whereas the arrival date is in a total of four columns mainly [arrival\_date\_year], [arrival\_date\_month], [arrival\_date\_day\_of\_month] and [arrival\_date\_week\_number]. To maintain consistency of the dataset, [arrival\_date\_year], [arrival\_date\_month] and [arrival\_date\_day\_of\_month] have been concatenated to form a new column [arrival\_month] and converted to IDate format (Appendix 7).

1. Removing columns

Some columns that will not be contributable to the insights of the research have been dropped. For example, [company] and [agent] columns will not be relevant as they are encoded with numbers due to data privacy policy. The industry of the company is not given and will not value-add to the conclusion of the datasets. In addition, the original columns [arrival\_date\_year], [arrival\_date\_month] and [arrival\_date\_day\_of\_month] have been removed as well for a cleaner dataset (Appendix 8).

1. Missing/Invalid values/Outliers

To have a visualisation of the dataset, summary() has been used to interpret whether the dataset has missing values, invalid values, or potentially outliers (Appendix 9).

[children] has a total of 4 NA cases. These 4 cases will be removed in the dataset as these NA cases constitute an insignificant percentage in the whole dataset of 0.0003% and replacing the NA values with numeric values will require certain assumptions. Thus, complete.cases() have been used to drop the 4 cases (Appendix 10).

[distribution\_channel] has 1 undefined value. Under [distribution\_channel], there are a total of 2 categorical values, “Travel Agents” and “Tour Operators”. The undefined value is not part of the categorical value and has been dropped since it constitutes a small factor in the dataset of 0.0008% (Appendix 11).

[ADR] is determined as the average daily rate which is calculated by dividing the sum of all lodging transactions by the total number of nights stayed. There are negative lodging transactions (Appendix 12). It is impossible for ADR to be negative as there ought to be some transactions and even if there are refunds, hotels would not provide refunds that are beyond what customers are paid off. Thus, the negative lodging transaction rows have been dropped. In addition, there is an extreme outlier ADR value that will not be dropped as when retrieving the column, the customer has a non refundable transaction with a grade A room and customers had cancelled the reservation. Hence, the charge may be reasonable and the outlier has not been dropped.

[babies] have two extreme outliers in which there are two columns with 9 and 10 babies respectively (Appendix 13). These columns would be dropped as it does not constitute a big part of the dataset (Appendix 14).

There are a total of 180 columns with zero adults, zero children, and zero babies. These columns will be dropped as it is impossible for the hotel room to be booked without having anyone staying. We will not be replacing the values with mean values to retain the authenticity of the dataset and thus these columns will be dropped (Appendix 15).

1. Checking correlation (to be updated)

Pairwise correlation has been down on both categorical and continuous variables to identify if there are any potential collinearities. From Appendix 16 and 17, there is no collinearity between the independent variables. Thus, analysis for the full set of variables can be conducted.

1. Filtering

Given that the context is in Singapore, it would be more applicable to focus on city hotels instead of resorts. Hence, subset() has been used to focus solely on “City Hotel” under [hotel] attribute.

The subsetted dataset is rather balanced where more than 42% of the dataset *is\_cancelled* while the rest is not (Appendix 18). This will ensure the model will be well trained as it helps prevent the model from becoming biassed towards one class.

# **6. D**ata Exploration

## 6.1. Customer Data Set

The objective of this segment is to help the hotelier understand its customer base by examining their nature and needs. Graphs

1. Dashboard 1: Booking analysis

In Dashboard 6.1 we first have a graph depicting the Distribution of Nationality, which shows us where customers are coming from. There is also a pie chart to see which season is the most popular for customers to visit. Next is graph 6.1 which illustrates total bookings and revenue by month, and a bubble graph to see how many bookings were cancelled or no-show.

Booking and Revenue by Months

Zooming in on graph 6.1, we can see that lodging revenue and total booking have a strong relation with each other However, the former is consistently observed to be higher than the latter except for the month of August where revenue surpassed bookings. This could indicate an increased tendency for customers to splurge during their stay, for this part of the year.

1. Dashboard 2: Timing and Requests

From Dashboard 6.2, graphs 6.2 (Average lead time for each guest arrival by months) and 6.3 (Type of Special Request) is an additional graph showcasing the proportion of lodging and non-lodging revenue for each market segment. It is observed that complementary customers contribute more to non-lodging revenue as compared to other market segments.

Average Duration before Guest Arrival

As shown in graph 6.2, customers averagely book their stays more spontaneously for winter seasons, giving the hotel shorter buffer time to prepare for their arrival.This can probably be attributed to the fact that hotel occupancy is lower during winter as shown in the previous plot. As such, customers may not see a need to plan their stays as far ahead due to low anticipated demand. However, the hotelier would need to note that in January and February their lead time can be as little as 25 days.

Frequency of Special Request

When looking at the different types of special requests that customers would make during their stay. We can see from graph 6.3 that requests with regards to bed size are of the highest count, as at least 29,000 of bookings requested for a king size bed while close to 12,000 bookings requested for a twin bed. Therefore, hotels could arrange for more rooms to contain King size and twin beds to meet guests’ demand.

1. Dashboard 3: Market Segment

Dashboard 6.3 sheds light on the hotelier’s customer base, to grant them a better understanding of their target audience.

Market segment share

Market segment shows the specific platforms which the customer used to make their reservations with the hotel. As seen in graph 6.4, a vast majority of customers come from other types of market segment, with travel operators following in second. This information tells us that most customers tend to book through third-party channels such as metasearch sites and online travel agencies like Expedia and Hotel Trivago. Hence the hotel should aim to direct more resources to third parties as part of their distribution strategy.

Age Distribution with Distribution Channels

In graph 6.5, we could see most of the customers that come to the hotel are people between the ages of 46-50 years old. The hotelier’s demographic is more skewed towards the middle-aged population.Middle-aged tourists tend to travel with their families for vacation or on work trips. Hence with this knowledge, the hotelier can pivot their services and marketing strategies to cater to this audience.

Rooms Required

[RoomNights] is right-skewed due to outliers in Graph 6.6a . Outliers will not be removed since it is very possible that a single organiser may book 100+ rooms in a single booking in the event that they are a travel agent or corporate entity. Therefore, we would zoom into where [RoomNights]s is less than 10, for a better visualisation. As shown on graph 6.6b, on average, organisers tend to occupy about 3 rooms throughout their entire stay. 3 rooms could mean: 1 room booked for 3 nights or 3 rooms booked for 1 night each. Furthermore, we could see a sharp drop from 3 to 4 -5 nights stay in the hotel. Therefore, organisers would tend to book only for 1-3 nights.

## 6.2. Hotel Booking Demands Data Set

1. Dashboard 4: Room Assignment

To provide better management of better room allocation, Dashboard 4 allows hotels to understand which rooms are in the highest demand from the pie chart, as there is a bar chart providing us information about rooms they were assigned to when checked in with the collect allocation rate. In the dashboard, will also show No. of children and babies in each transaction, when filtering the rooms booked by the guest could see which rooms are more baby and children popular. As shown we could see that Type A rooms are highest in demand, with the hotel assigning about 91% of the rooms being correctly assigned rooms.

1. Dashboard 5: Cancelation Rates

In dashboard 5, show us what attribute will lead to a higher cancellation rate. From Graph 6.6 we could see what kind of deposits are most often used by different distribution channels. Followed by the average lead time for different customers. By filtering both graphs the hotel can determine the cancellation based on previous transactions.

Distribution Channels for each deposit type

In graph 6.7, the data has been grouped up to the different deposit types and proportion of distribution channels. In the graph, all Deposit types of the highest Distribution Channel are Travel Agency and Operators, followed by direct distribution from the hotel itself. We could see that based on values on the pie chart the most popular deposit type is having no deposit, followed by Non-refundable and refundable deposit types.

Cancelation rate from previous transaction

Based on graph 6.8 we could see that customers that had cancelled only once have the highest chances of cancelling the hotel booking again. With almost 96% of them cancelling again. Hence it would be best for the hotel to anticipate for customers to have repeated cancelation. Whereas, around 40% of customers would cancel their booking if they had never cancelled before.

### **6.2.1. Cancellation Attributes**

Subsequent to data preparation in 5.2.2, data exploration is conducted on hotel booking cancellations. The following findings seek to highlight actionable insights for hotel management.

1. Deposit Type

The majority of bookings comprises customers who did not place a deposit, as compared to customers who made varying amounts of deposit. However, the proportion of cancellations within “Non-Refund” bookings was substantially cancelled (Appendix 20). To determine the reason behind the high proportion of cancellations, the cancellation and deposit policy of the hotel becomes largely relevant in our analysis.

1. Previous Bookings and Cancellations

For first-time customers, there would not be an existing record of previous bookings nor cancellations. We can further our analysis of first-time customers and repeated customers, specifically cancellations of the segmented customers. With a higher number of first-time bookings, the proportion of cancellations among first-time bookings is also higher at 2.66% (Appendix 21). Among repeated customers, there is an evident pattern between proportion of cancellations in previous bookings and the likelihood of cancellations (Appendix 22). Following this observation, hotels can leverage on historical bookings of customers to revise their cancellation policies. Cancellation policies play an important role by impacting the search and booking behaviour of customers.

1. Average Daily Rate (ADR)

Cancellations on average had a lower average daily rate of $104.95 and a lower interquartile range (IQR) of $48.91 as compared to non-cancellations, which had an average daily rate of $108.27 and IQR of $46.15 (Appendix 23, 24). Average Daily Rate is crucial for hotels to leverage on modelling the price sensitivity threshold of customers. Customers may continue to search after they have made a reservation, looking for an even better deal. If a better deal is found after they made their initial booking, customers are inclined to cancel their existing reservation and rebook the better deal (Chen et al, 2011).

1. Lead Time

Cancellations on average had a higher lead time of 150.7 days and higher IQR of 177 days, as compared to non-cancellations, which had an average lead time of 81.9 days and IQR of 109 days (Appendix 25). This is indicative that bookings made further in advance had a higher likelihood of being cancelled. While the data source does not mention the hotel’s cancellation policy, it is likely that there would have been a stipulated late cancellation fee in influencing the trend of lead time on cancellations.

1. Length of Stay (Weekday and Weekend Stays)

The length of stay does not significantly impact the probability of cancellations, with cancellations having a slightly higher average length of stay of 3.02 days than non-cancellations of 2.94 days but the same IQR of 2 days (Appendix 26,27). While the length of stay does not have a definitive impact on cancellations, the factor of cancellation policy based on the length of stay can be explored in mitigating the impacts of cancellations.

# 

# 7. Data Modelling

## 7.1 Customer Segmentation with Clustering

Our team decided to use K-Means clustering to identify specific groups of customers which frequent the hotel. Due to RAM limitations, only 10% of data is randomly selected to be put through clustering. For clustering to work, only numeric variables can be used hence we only include [AverageLeadTime], [LodgingRevenue], [OtherRevenue], [RoomNights], [NumSR] and [DistributionChannel]. [DistributionChannel] was converted into numeric from factor.

### **7.1.1 Optimal Cluster**

There are 3 methods to find the optimal cluster: elbow method, average silhouette method and gap statistics. This section will examine the optimal number of clusters (k) using all 3 methods and use the average k derived to reduce individual model bias.

1. Elbow Method (Appendix 28)

From the plot, the optimal cluster is k=4 as the gradient of slope decreases from 4 onwards which represents the diminishing returns of having more clusters. However the “elbow” is not as obvious which could render the use of another method to confirm the number of clusters.

1. Average Silhouette Method (Appendix 29)

For this method, we pick k where silhouette coefficient is the highest. An ideal silhouette coefficient is 1, which indicates that clusters are clearly distinguished from one another. On the other hand, 0 indicates that clusters are overlapping and -1 implies errors within clusters. From the appendix, optimal k=2 where silhouette width is the highest.

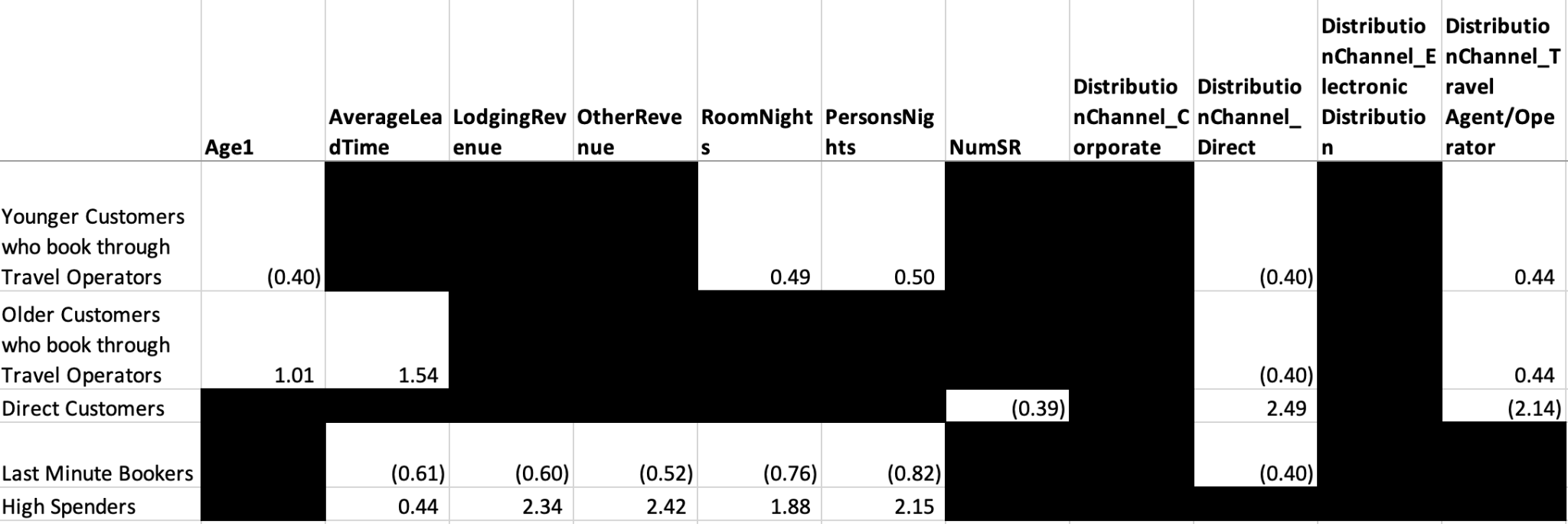
1. Gap Statistics (Appendix 30)

Gap statistics compares the change in within-cluster variation with the uniform distribution. A larger variation results in a larger gap statistic which indicates that more clusters may be needed to represent the data. As shown in the appendix, optimal k=10, where the gap statistic is highest.

With our overall findings, the optimal split is determined to be at k=5.

### **7.1.2 Clustering Results**

Customer segments are shown to be overlapping rather than in relative isolates (Appendix 31). This is expected in our use-case due to the complexity of human behaviour and consumption. To break down each of these segments, cluster centres are extracted into an excel sheet. Cluster centres that are close to 0 indicate that the variable does not differ much from the overall mean and is therefore not a defining variable used in clustering. Conditional formatting is therefore used to filter out values between -0.35 and 0.35 as they are considered to be close to 0. A higher value indicates that it is higher than the mean and the opposite is true for a lower value.



*Figure 2: Customer clustering on Microsoft Excel*

5 prominent customer segments and their characteristics are identified (from top to bottom):

1. **Young customers from travel operators**: these individuals tend to book more rooms for more people, indicating that they are likely to travel in groups.
2. **Older customers that book through travel operators**: these bookings are made way in advance prior to their arrival.
3. **Direct customers**: those that book directly with the hotel tend to have fewer special requests made.
4. **Last minute bookers**: they tend to spend less throughout their stay, require fewer nights of accommodation and usually book for fewer people.
5. **High spenders**: they tend to book in advance for more people and require more rooms throughout their stay. These individuals are not limited to any distribution channel.

With this information, the hotelier can narrow down when a customer group is likely to book (Appendix 32), and explore their tendency to make special requests (Appendix 33).

## 7.2 Cancellation Modelling with Logistic Regression

A 70-30 train-test split was conducted to obtain a logistic regression model with all independent variables in the dataset. However, to mitigate overfitting, only significant variables with a p-value of 5% or less were selected for the second logistic regression model, which underwent feature selection (Appendix 34).

To identify the most important variables contributing to the prediction of cancellations, we used Lasso regression. This was necessary due to difficulties encountered when identifying significant variables using logistic regression on the train set. Lasso regression performs feature selection by shrinking the coefficients of less important predictors to zero, effectively removing them from the model. We used the "auc" measure of model fit and set the number of folds to 10. Based on the output of the Lasso regression, we selected the model with the smallest value of lambda ( Appendix 35).

The final logistic regression model on the train set includes the following 15 features: {"lead\_time", "stays\_in\_weekend\_nights", "stays\_in\_week\_nights", "adults", "meal", "country", "distribution\_channel", "previous\_cancellations", "previous\_bookings\_not\_canceled", “reserved\_room\_type”, "booking\_changes", "deposit\_type", "customer\_type", "adr", "total\_of\_special\_requests"}.

The logistic regression model gave an accuracy of 80.28% on the train set and 80.00% on the test set based on the confusion matrices, with a false negative rate of 16.4% and a false positive rate of 3.6%. (Appendix 36). The ROC-AUC Curve gave a score of 0.85, with a relatively good discriminatory power in predicting booking cancellations (Appendix 37).

## 7.3 Cancellation Modelling with CART

In building the classification tree for the prediction of cancellation, all features were used to grow a maximal tree. The maximal tree is pruned thereafter with an optimal complexity parameter value using the 1 SE rule to obtain the optimal tree with 10 fold cross-validation (Appendix 38). This will give us the final CART model to predict booking cancellation.

The variable importance of features determined by CART model in order of significance are {“deposit\_type”, “lead\_time”, “market\_segment”, “previous\_cancellations”, “total\_of\_special\_requests”, “adr”, “customer\_type”, “days\_in\_waiting\_list”, “distribution\_channel”, “previous\_bookings\_not\_canceled”, “stays\_in\_week\_nights”, “assigned\_room\_type”, “adults”, “stays\_in\_weekend\_nights”, “meal”, “booking\_changes”, “required\_car\_parking\_spaces”, “reserved\_room\_type”, “country”}.

The CART model gave an accuracy of 85.57% on the train set and 84.00% on the test set based on the confusion matrices, with a false negative rate of 10.6% and a false positive rate of 5.4%. (Appendix 39). The ROC-AUC Curve gave a score of 0.89, with a better discriminatory power than logistic regression (Appendix 40).

## 7.4 Cancellation Modelling with Random Forest

Using the parameters of 10-fold cross validation and 500 trees, the initial random forest model gave an accuracy of 84.81% on the train set and 84.90% on the test set based on the confusion matrices, with a false negative rate of 10.8% and a false positive rate of 4.3%. (Appendix 41).

The variable importance of features determined by Random Forest model in order of significance are {“previous\_cancellations”, “lead\_time”, “adr”, “total\_of\_special\_requests”, “market\_segment”, “deposit\_type”, “stays\_in\_week\_nights”, “stays\_in\_weekend\_nights”, “customer\_type”, “booking\_changes”, “adults”, “meal”, “required\_car\_parking\_spaces”, “country”, “assigned \_room\_type”, “distribution\_channel”, “reserved\_room\_type”, “days\_in\_waiting\_list”, “previous\_bookings\_not\_canceled”} (Appendix 42).

Subsequently, to further refine accuracy, an optimal value of ‘mtry’ of 20 was tuned to obtain a train set accuracy of 85.56% and test set accuracy of 85.95% (Appendix 43, 44). The ROC-AUC Curve gave the highest score of 0.91, this discrimination between cancellations is relevant whereby hotels can take proactive measures to classify and retain customers who have a high possibility of cancellation (Appendix 45).

## 7.6 Model Evaluation

The following metrics are used to evaluate the models: Confusion Matrix, Overall Accuracy, Precision, Recall, F1 Score, AUC-ROC

| **Model** | **Evaluation Matrix Percentage/%** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Overall Accuracy | False Negative Rate | False Positive Rate | F1 Score | AUC-ROC |
| Logistic Regression | 80.00 | 16.40 | 3.60 | 72.70 | 0.85 |
| CART | 84.00 | 10.60 | 5.40 | 79.85 | 0.89 |
| Random Forest | 85.95 | 8.40 | 5.60 | 82.85 | 0.91 |

False Positive (Type 1 error) is defined as customers who did not cancel their hotel reservations but have been predicted to have been cancelled whereas False Negative (Type 2 error) is defined as customers who have cancelled their hotel reservations have been predicted as not cancelled. Our team believes that False Negative is crucial in our evaluation in terms of hotel cancellation prediction due to the time sensitivity in filling the vacant slot, where hotels may not be able to fill the cancelled rooms in time and leads to a loss of revenue (Appendix 46)

Based on the above metrics, Random Forest is the best-performing model with the highest overall accuracy, lowest false negative rate and F1 Score. The team had previously determined that false negative negative rate is crucial in our decision making to determine which model to use. Since the random forest model has a false negative rate of at least 2% to the next best alternative, CART model. The team believes that random forest is the best in determining cancellation rate and can be implemented in hotels to help resolve the problem with regards to cancellation rate and suggest possible mitigation measures.

## 7.6 Model Implementation

The advantages and limitations of the models should also be considered in its implementation.

### **7.6.1 Logistic Regression**

Logistic Regression is a simple model for hotels to model cancellation behaviour of their customers, and can be further extended to multinomial regression where relevant to the business model of the hotel. Regularisation techniques implemented such as Lasso Regression can also be used to overcome the issues of overfitting in high-dimensional datasets.

However, a major limitation is the assumption of linearity between the dependent variable and the independent variables. Linearly separable data is rarely found in real-world scenarios, and this may reduce the effectiveness of Logistic Regression for certain datasets. When the relationship between the outcome variable and predictors is non-linear, other models such as decision trees, random forests, or neural networks may perform better.

Logistic regression also requires moderate or no multicollinearity between independent variables. As a result, logistic regression is not flexible with the addition of new data (Grover, n.d.). Hence, when additional features are added to the dataset, it is essential to check for multicollinearity before fitting a Logistic Regression model.

### **7.6.2 CART**

As a tree algorithm, CART can automatically select the most important features for prediction by choosing the best split based on a measure of purity or information gain. This is especially useful when dealing with datasets with many features, as it reduces the computational burden and can improve the accuracy of the model. Additionally, CART computes feature importance scores based on the frequency with which a feature is used for splitting across all decision trees in an ensemble. This score can provide additional insights into the relative importance of features and help identify potential redundancies in the dataset.

Overall, the ability of CART to automatically select important features for prediction can be very useful for handling large and complex datasets, as it reduces the need for manual feature selection and can improve the accuracy and interpretability of the model.

The issue of CART being prone to overfitting can be mitigated through pruning and cross-validation. This helps to reduce the complexity of the model and improve its generalisation performance. Furthermore, ensemble methods such as bagging offered by Random Forest also address the limitations of a CART model.

### **7.6.3 Random Forest**

Using bagging, Random Forest can reduce the overfitting problem that CART faces. By constructing multiple decision trees and taking an average of their outputs, random forest can produce a more generalised model that is less prone to overfitting. For large and complex datasets, Random Forest can handle interactions between variables and capture non-linear relationships that may be missed by CART. Furthermore, Random forest is less sensitive to outliers than CART. Outliers have less impact on the performance of random forest since the model is constructed based on a majority vote of many trees, and outliers are more likely to be balanced out across the different trees.

However, while Random Forest can provide insight into the relative importance of features, it can be difficult to interpret the actual decision-making process of the model, as it involves an ensemble of decision trees. Additionally, the use of randomised feature subsets and bootstrapped samples can make it harder to trace the contributions of individual features or data points to the final prediction. This can limit the interpretability of the model, which can be a concern in certain applications where transparency and accountability are important to determine the cancellation prediction of a customer.

# 

# 8. **Solutions**

With the business problems in the hotel industry identified, we propose the following solutions and its feasible implementations in addition to existing initiatives.

## 8.1 Overbooking

Overbooking is the practice of accepting more reservations than there is inventory available, based on the expectations that some reservations will result in no-shows or cancellations (Localle, 2021). Operating at maximum occupancy would allow the hotel to produce above usual profit margins.

However, there are 2 main drawbacks to overbooking:

1. Loss of reputation and negative guest experience
   1. Reputation is important in the hospitality industry, and can be tarnished if customers are turned away when they have previously made reservation
   2. A common mitigation measure used is through overbooking partnerships with same-starred hotels and training staff to handle overbookings professionally
2. Cost of compensation
   1. Appropriate compensation would be owed to affected customers
   2. As long as the cost of compensating affected customers is lower than the revenue brought about by the increase in operating occupancy, net profit will increase.

The ideal number of overbookings is calculated by the probability of cancellation minus a buffer. Traditionally, this probability is taken from historical rates of cancellation from previous months. However, with the implementation of analytical techniques such as Random Forest, hotels would be able to have an overbooking number that varies according to certain features.

Some limitations with our model currently exist for the immediate implementation of overbooking. Firstly, the dataset we use has missing data that would be crucial in calculating the overbooking number. E.g. no. of *walk-in bookings* and *overstays.* If there is a sizable or reliable stream of walk-in bookings, the ideal number to overbook by would decrease, as last minute cancellations can simply be replaced by impromptu walk-ins instead of risking payment of compensation fees. Secondly, the compensation policy as well as the cost of the overbooking partnership is integral to the calculations of the ideal number for overbooking.

## 8.2 Targeted Marketing Strategies

Understanding Customer Needs

While hotels aim to maximise their revenue, they face operational limits during peak periods where demand for bookings is especially high from multiple distribution channels. On the contrary during non-peak periods, hotels face the challenge of greater overhead cost allocation with low occupancy rate. In addition to hotels’ current implementation of off-peak and peak pricings, hotels can leverage on targeted marketing strategies to optimise daily bookings. By understanding the needs and preferences of different customer segments, hotels can tailor their marketing efforts to attract and retain these guests. For example, business travellers may value convenience and fast Wi-Fi, while families may prioritise spacious rooms and on-site amenities for children. Hotels can also utilise different distribution channels to target these customer segments. For example, online travel agencies may be more effective for last-minute bookings, while direct bookings through the hotel's website or mobile app can offer more personalised and flexible options.

Optimising Distribution Channels

With the prevalence of online bookings, it is estimated that 74% of total revenue in the Travel & Tourism Sector will be generated through online sales by 2027 (Statista. 2022). This brings about the need for hotels to leverage on different online distribution channels to optimise revenue. Online travel agencies, which are third party websites, are commonly used as they aid hotels in increasing their visibility and bookings, especially for last minute bookings. However, the benefit received by hotels from these travel agencies varies due to the cost of commissions charged. On the other hand, transient bookings, which are bookings made directly with the hotel, can be more profitable for hotels in some cases. Additionally, hotels may be able to offer special promotions or packages to guests who book directly with them, which can help increase revenue.

As there is no particular best distribution channel, the desired approach by most hotels should be to strike a balance between various distribution channels through targeted marketing strategies which influence customers’ behaviour. One approach is to offer special promotions or packages for guests who book directly with the hotel, such as free breakfast or a discount on their next stay. This can incentivize guests to book directly and increase the number of transient bookings.

## 8.3 Pricing Strategies and Policies

With the aforementioned predictive models for customer cancellations, they can be deployed by integrating into the central reservation system used by hotels, in particular their customer records database. Given the time-sensitivity of bookings and its impact on attributes, this requires the models to be updated and evaluated on a daily basis.

Leveraging Discriminatory Pricing

To capture greater revenue, hotels can charge higher prices based on customer and booking details. Variable importance of features enable hotels to price based on features with high significance. By using discriminatory pricing through predictive models, hotels can optimise their revenue by charging higher prices to customers who are willing to pay more, while still attracting price-sensitive customers with lower prices.

Discriminatory pricing helps hotels to increase their revenue and profitability by tailoring their pricing strategy to specific customer segments and demand patterns, while still remaining competitive in the market.

Cancellation Policies

Certain hotels are more reluctant to overbook over concerns of customer relation concerns, but instead, they are more inclined to charge and impose cancellation fees to recover lost income due to last minute cancellations and no-shows. Cancellation policies and deposits are integral for hotels to recover opportunity cost. Based on the prediction of cancellations by customers, hotels can use data insights to offer more flexible cancellation policies to customers who are less likely to cancel or to incentivize customers to keep their reservations. By analysing historical booking data and market trends, hotels can adjust their pricing strategies to maximise their revenue and occupancy rates. More importantly, the prediction of customer cancellations enables hotels to reduce losses and maximise the recovery of lost opportunity cost through filling up vacant rooms as soon as possible at the best price.

However, it is important for hotels to balance their cancellation policies with customer satisfaction. Overly strict or unfair policies can lead to negative reviews and deter potential customers. As hotels face high competition within the vicinity, it is crucial for them to communicate their cancellation policies clearly and provide customers with flexibility to improve the overall customer experience, which in turn contributes to the hotel's business reputation.

## 8.4 Staff Rostering

Employee retention is a significant issue in the hospitality industry, which experiences peak and off-peak seasons with varying manpower requirements. While analytics cannot solve the root cause of seasonal employment, it can help hotels predict their peak seasons and begin recruitment efforts earlier, reducing their manpower deficit. By utilising seasonal demand data, hotels can roster part-time staff shifts to ensure steady employment throughout the year, rather than offering only periodic employment. Consequently, this helps to reduce staff turnover rate.

Optimising staff rostering also enables hotels to manage its overhead cost and maximise operational efficiency. By analysing historical data on occupancy rates, booking patterns, and employee productivity, hotels can optimise their staff rostering to reduce unnecessary labour costs and ensure adequate staffing levels during peak periods. For example, hotels can forecast demand for different types of staff, such as front desk personnel or housekeeping staff, and adjust their staffing levels accordingly. This can help reduce overstaffing during low-demand periods and understaffing during peak periods.

# 9. **References**

*Advanced analytics in hospitality*. (2017, October 3). McKinsey. Retrieved March 31, 2023, from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/advanced-analytics-in-hospitality

Antonio, N., de Almeida, A., & Nunes, L. (n.d.). *A hotel's customers dataset*. Kaggle. Retrieved March 31, 2023, from https://www.kaggle.com/datasets/nantonio/a-hotels-customers-dataset

Antonio, N., de Almeida, A., & Nunes, L. (2020, December 16). *A hotel's customers personal, behavioral, demographic, and geographic dataset from Lisbon, Portugal (2015–2018)*. ScienceDirect. Retrieved March 31, 2023, from https://www.sciencedirect.com/science/article/pii/S2352340920314645?via%3Dihub

Chen, C.-C., Schwartz, Z., & Vargas, P. (2011, March 16). *The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers*. ScienceDirect. Retrieved March 31, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S0278431910000320

Falk, M., & Vieru, M. (2018, October 8). *Modelling the cancellation behaviour of hotel guests*. CORE. Retrieved March 31, 2023, from <https://core.ac.uk/download/pdf/302225593.pdf>

Grover, K. (n.d.). *Advantages and Disadvantages of Logistic Regression*. OpenGenus Foundation. https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/

Hansen, J. (2023, March 28). *Employee Turnover Rates by Industry (2019, 2020, 2021)*. Awardco. Retrieved March 31, 2023, from https://www.award.co/blog/employee-turnover-rates

*Hotel and resort industry market size worldwide 2022*. (2023, January 30). Statista. Retrieved March 31, 2023, from https://www.statista.com/statistics/1186201/hotel-and-resort-industry-market-size-global/

*How AI and Data Science Improve Hospitality Industry*. (2018, August 10). AltexSoft. Retrieved March 31, 2023, from https://www.altexsoft.com/blog/datascience/how-the-hospitality-industry-uses-performance-enhancing-artificial-intelligence-and-data-science/

Lacalle, E. (2021, February 1). *What is overbooking in hotels and what are its advantages?* Mews. Retrieved April 2, 2023, from https://www.mews.com/en/blog/hotel-overbooking-strategy

Mostipak, J. (2020, February 11). *Hotel booking demand*. Kaggle. Retrieved March 31, 2023, from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Thomas, L. (2022, July 28). *Hospitality operating costs rise to 55.2% of turnover*. The Drinks Business. Retrieved March 31, 2023, from https://www.thedrinksbusiness.com/2022/07/hospitality-operating-costs-rise-to-55-2-of-turnover/

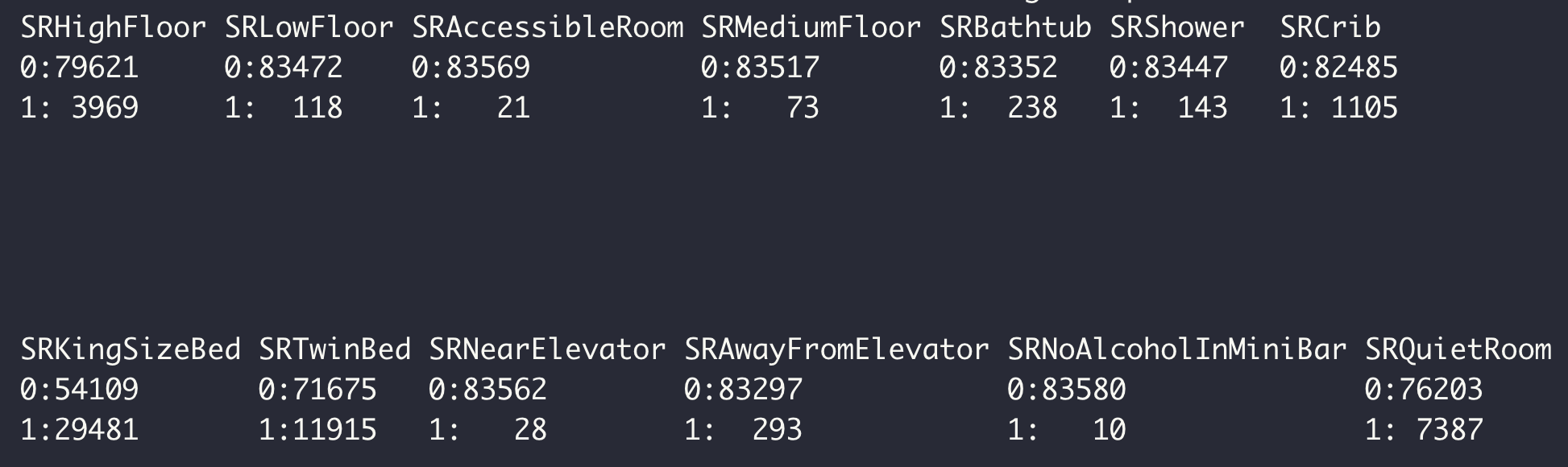
*Travel & Tourism - Worldwide*. (n.d.). Statista. Retrieved April 2, 2023, from https://www.statista.com/outlook/mmo/travel-tourism/worldwide

*What You Need To Know About Hotel Booking Cancellations & Big Data*. (n.d.). Iunera. Retrieved March 31, 2023, from https://www.iunera.com/kraken/interviews/what-you-need-to-know-about-hotel-booking-cancellations-big-data/

# 

# 10. **Appendices**

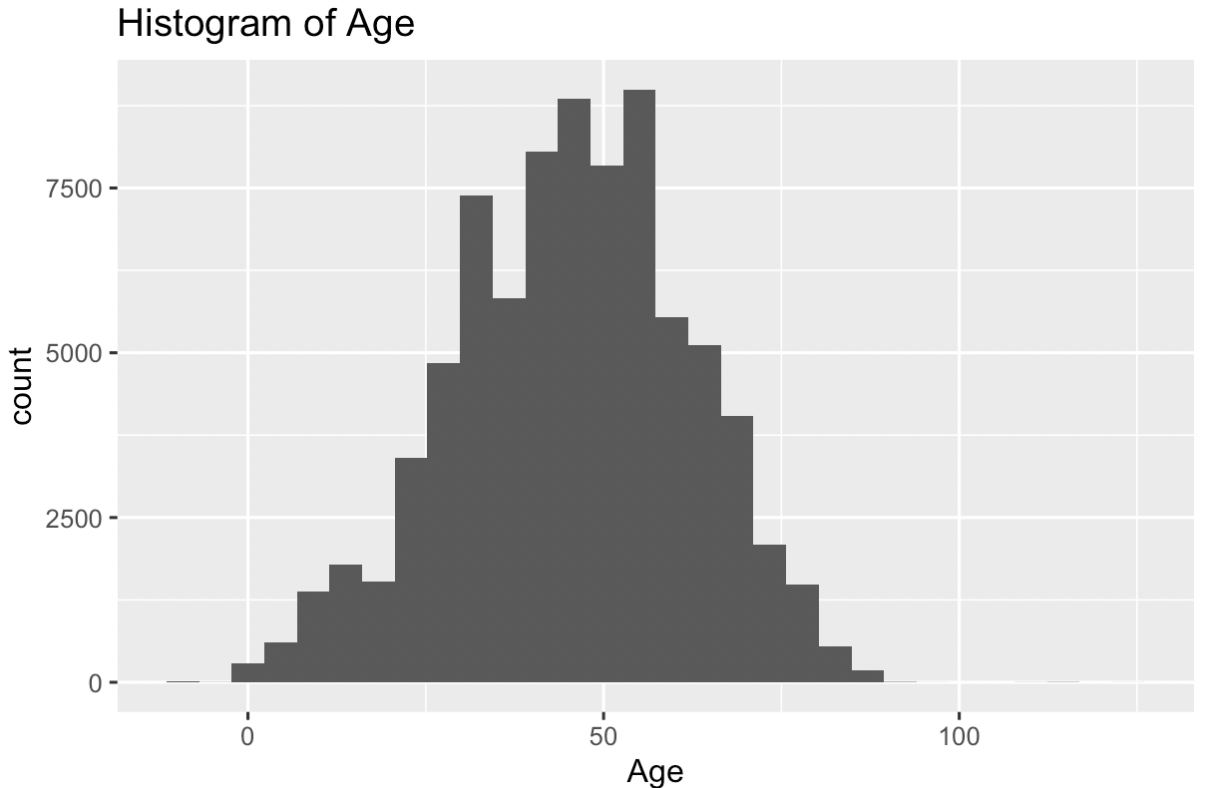
Appendix 1: Identifying insignificant SR-prefixed columns



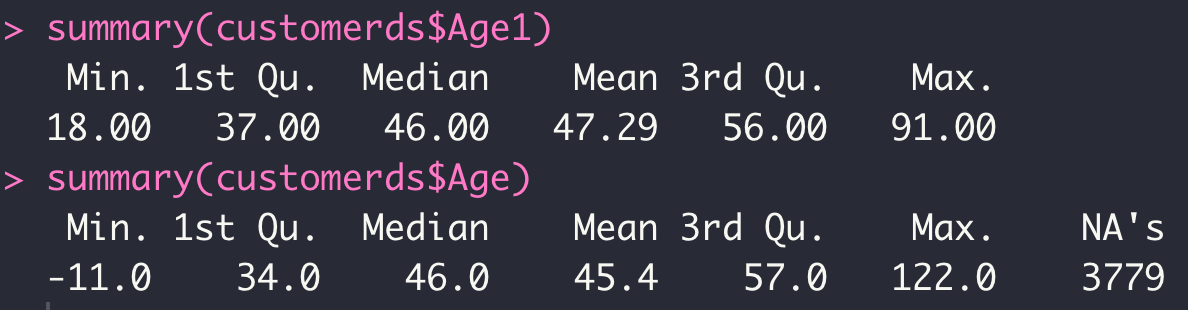
Appendix 2: Boxplot to identify outlier in Age



Appendix 3: Histogram of Age column to observe where data tends to cluster around



Appendix 4: Comparison between original and cleaned Age columns



Appendix 5: Seasons Classifying conditions

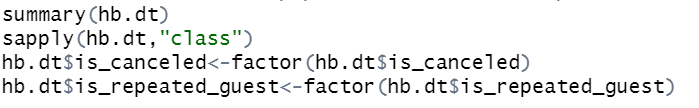
Spring: March, April, May

Summer: June, July, August

Autumn: September, October, November

Winter: December, January, February

Appendix 6: Checking class of variables and factorising variables



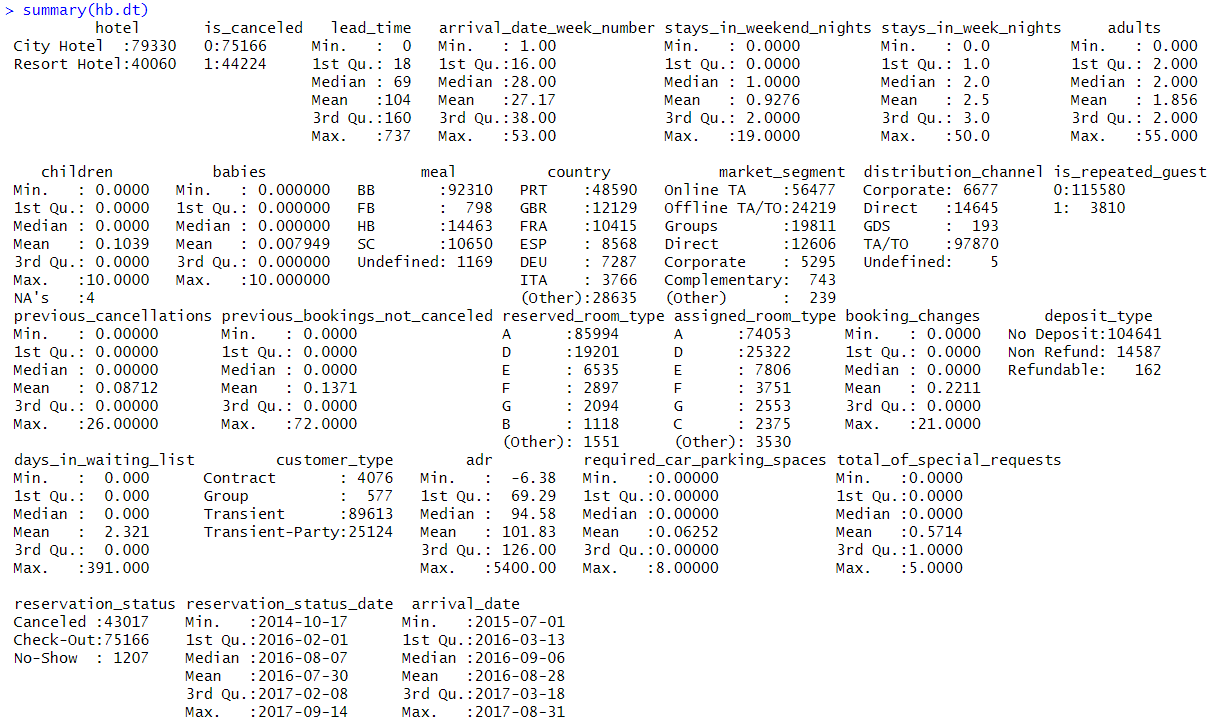
Appendix 7: Standardising columns



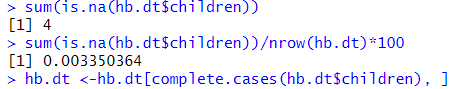
Appendix 8: Removal of insignificant variables



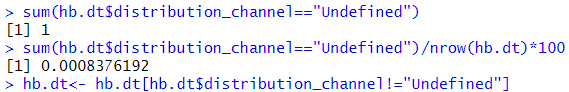
Appendix 9: Summary of dataset before further cleaning



Appendix 10: Missing values in [children]



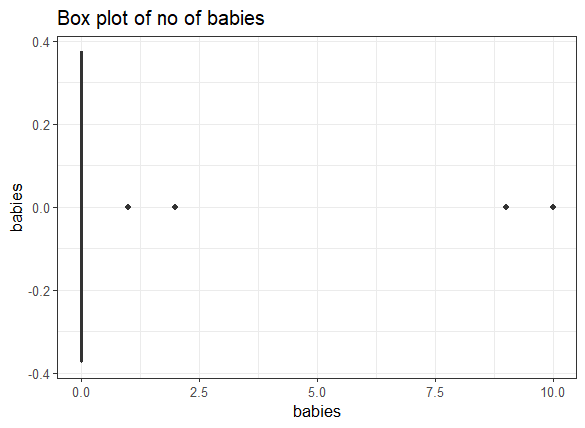
Appendix 11: Invalid values in [distribution\_channel]



Appendix 12: Presence of negative value in [ADR]



Appendix 13: Boxplot to visualise extreme outliers in [babies]



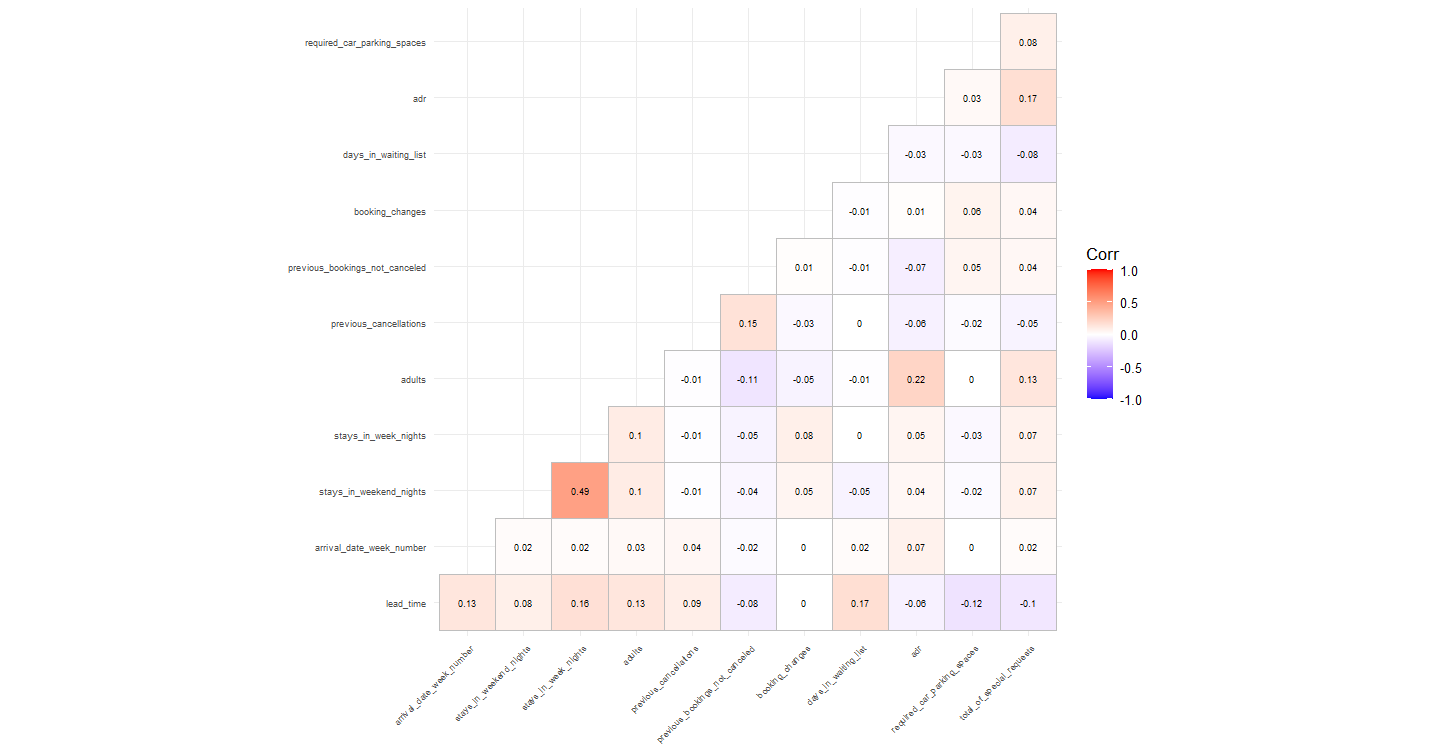
Appendix 14: Removal of extreme outliers in [babies]



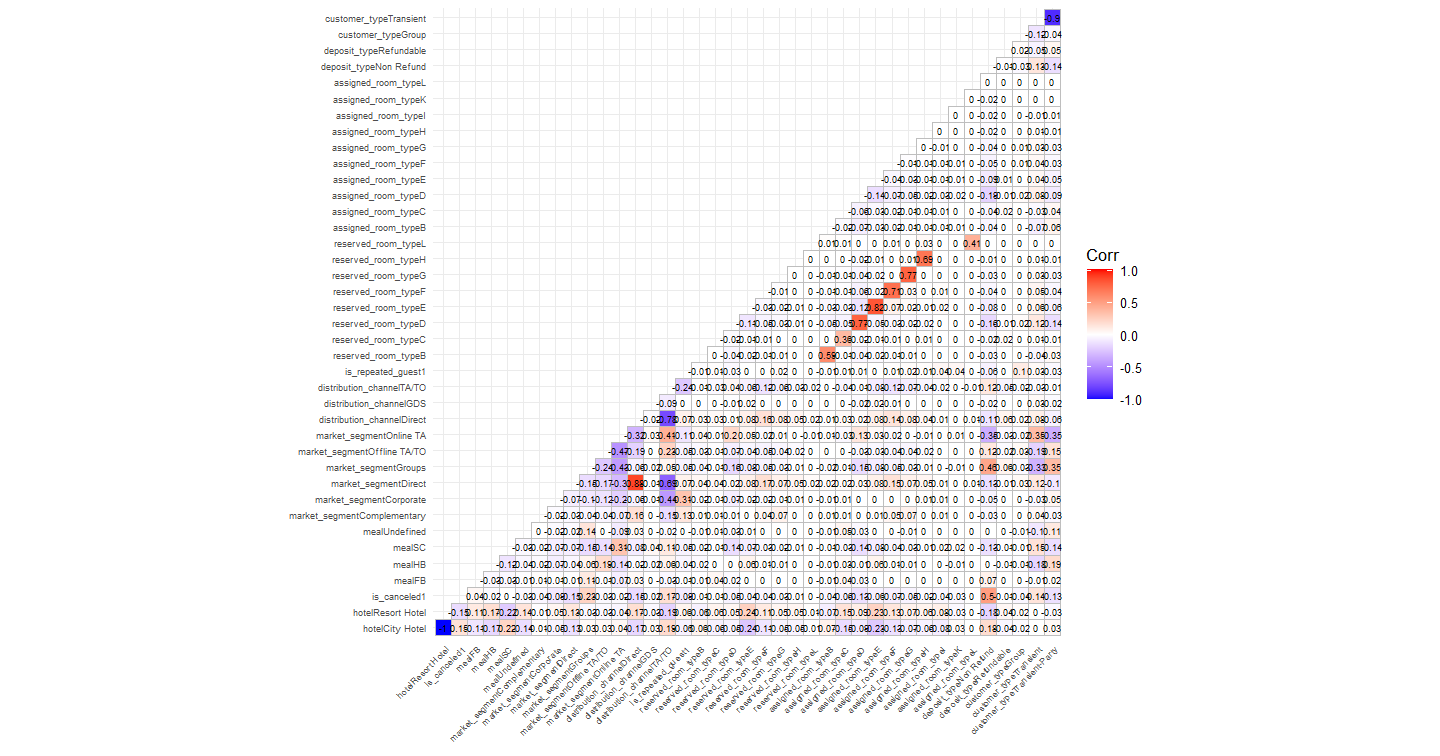
Appendix 15: Removal of columns with no presence of people



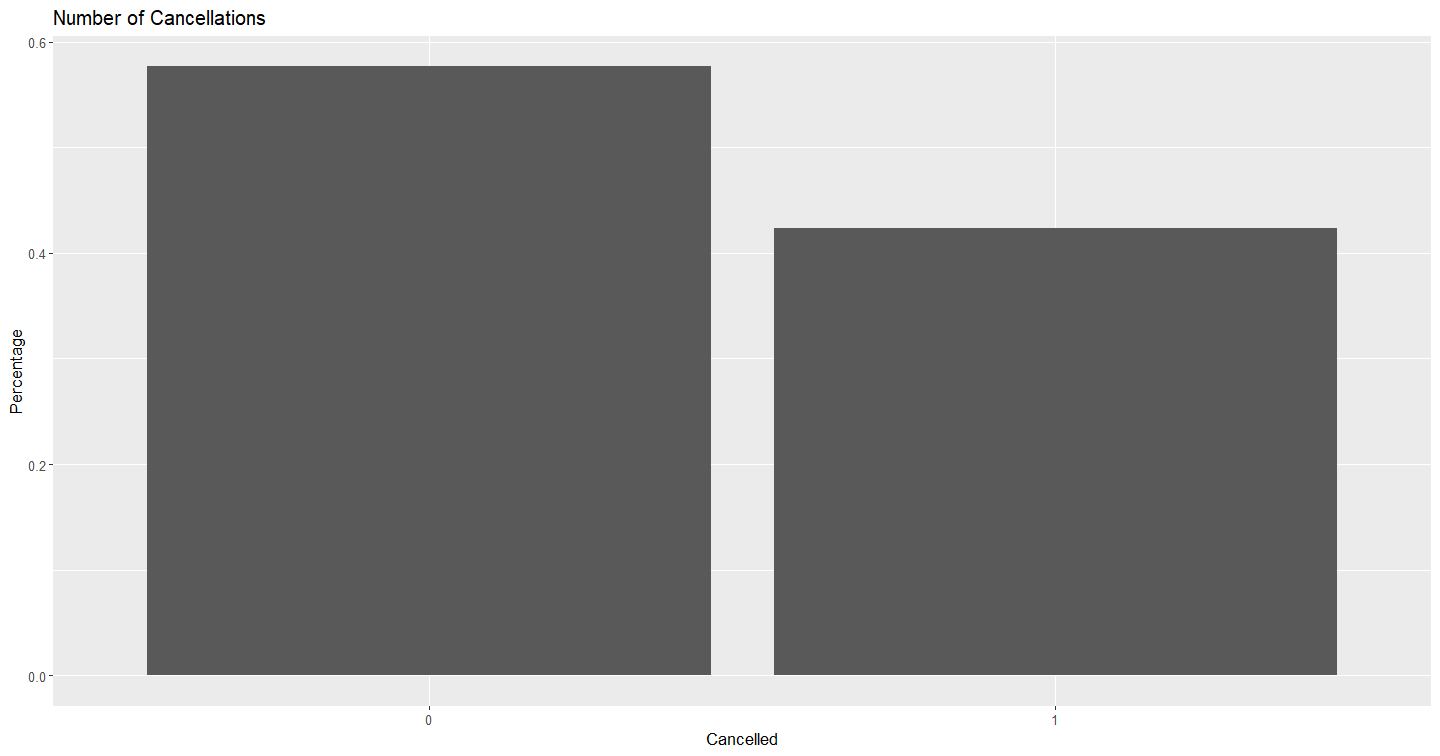
Appendix 16: Correlation between independent continuous X variables



Appendix 17: Correlation between independent categorical X variables

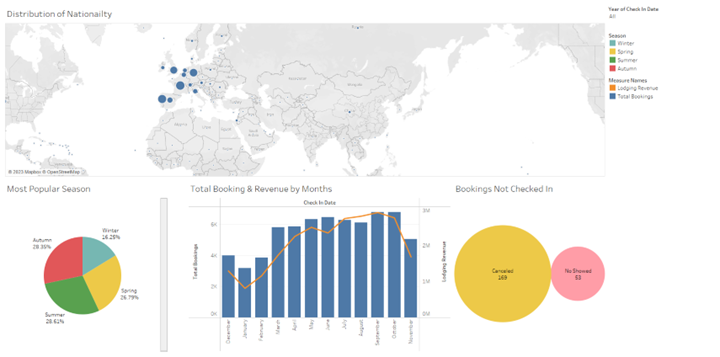


Appendix 18:

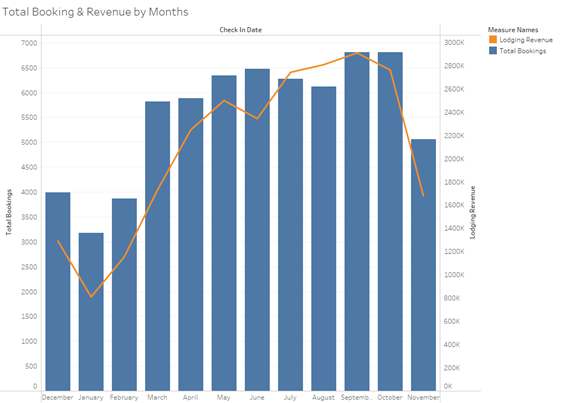


Appendix 19: Data Exploration on Dashboard

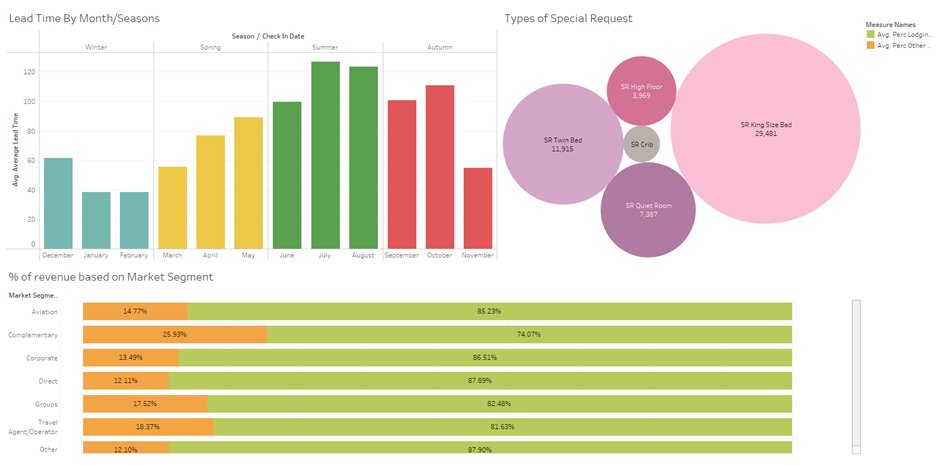
Dashboard 6.1:



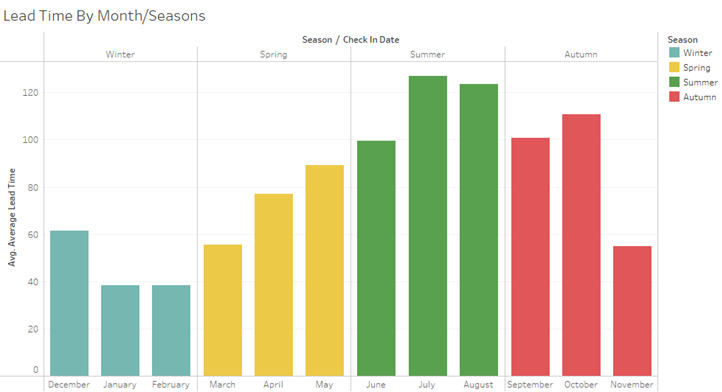
Graph 6.1:



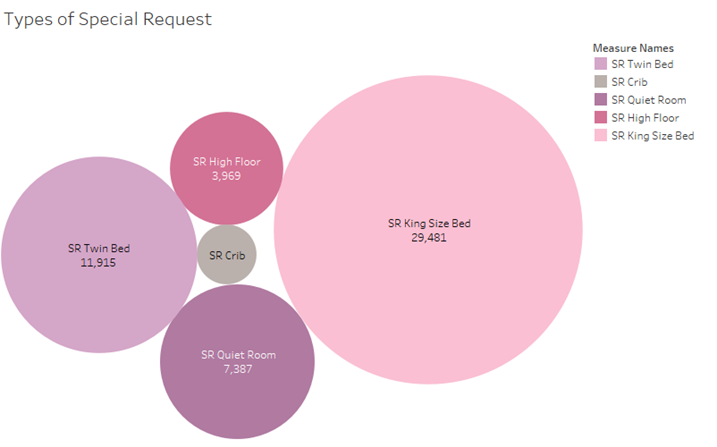
Dashboard 6.2 :



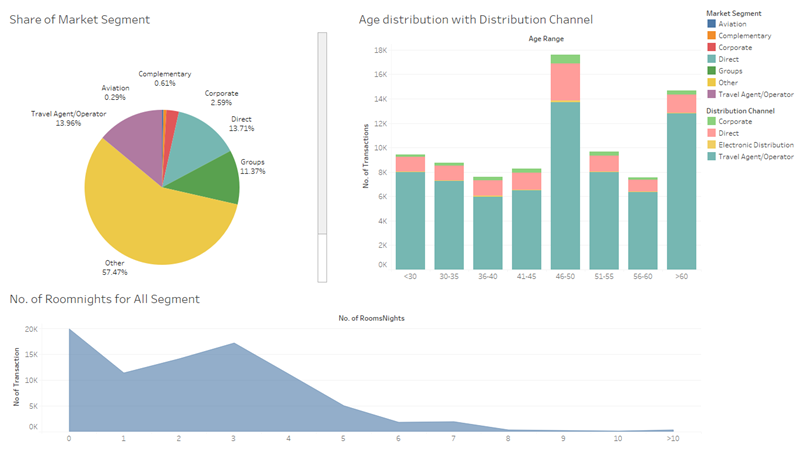
Graph 6.2:



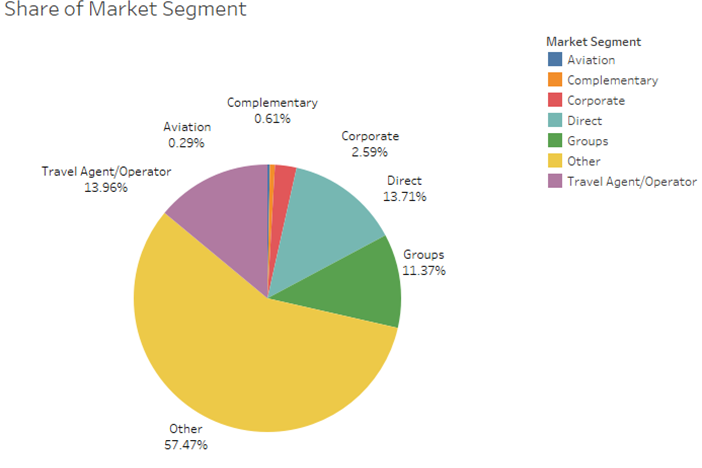
Graph 6.3:



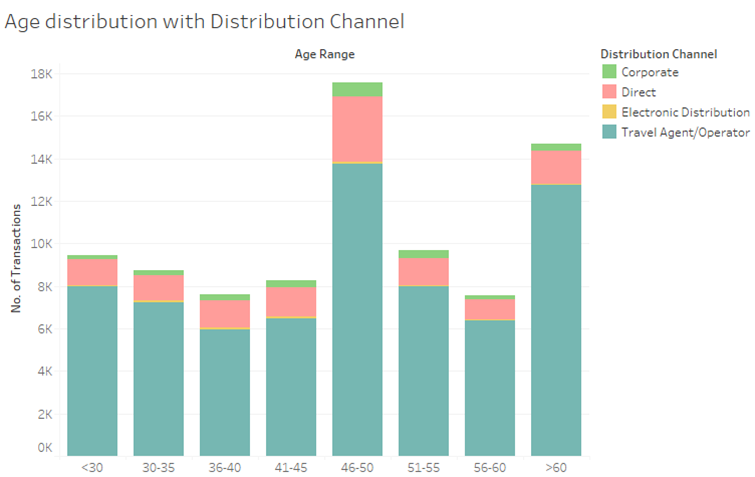
Dashboard 6.3:



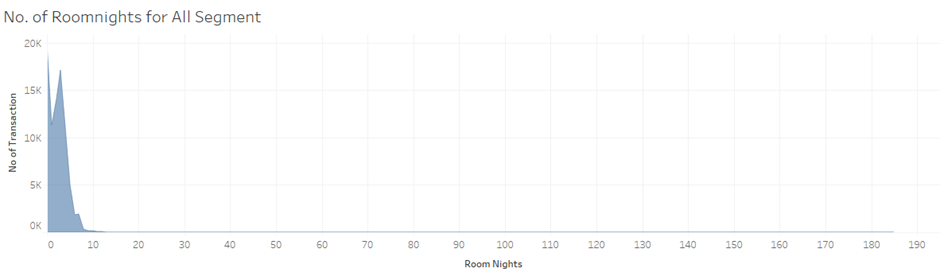
Graph 6.4:



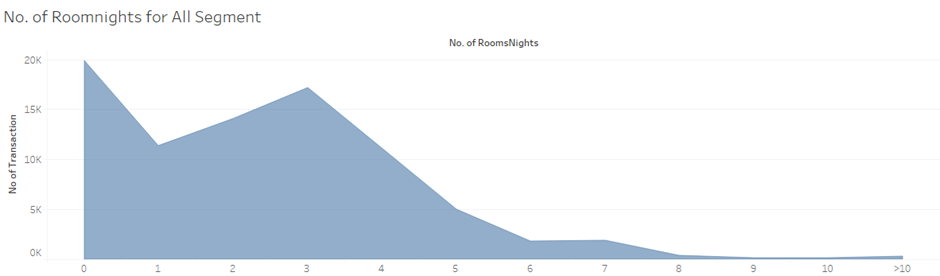
Graph 6.5:



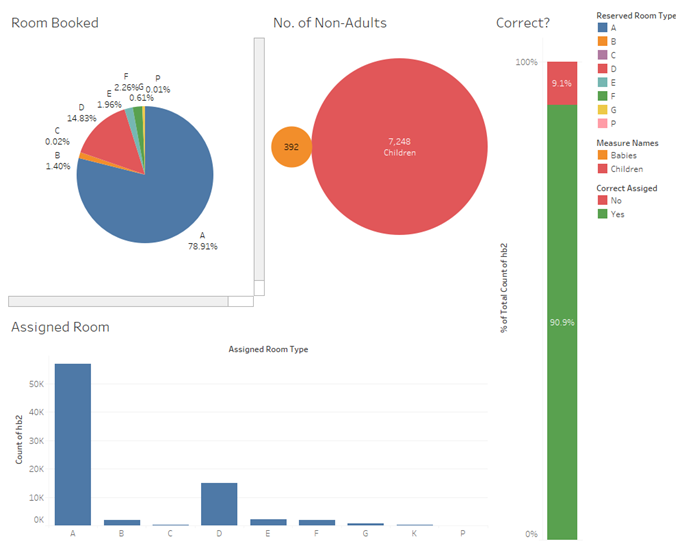
Graph 6.6a:



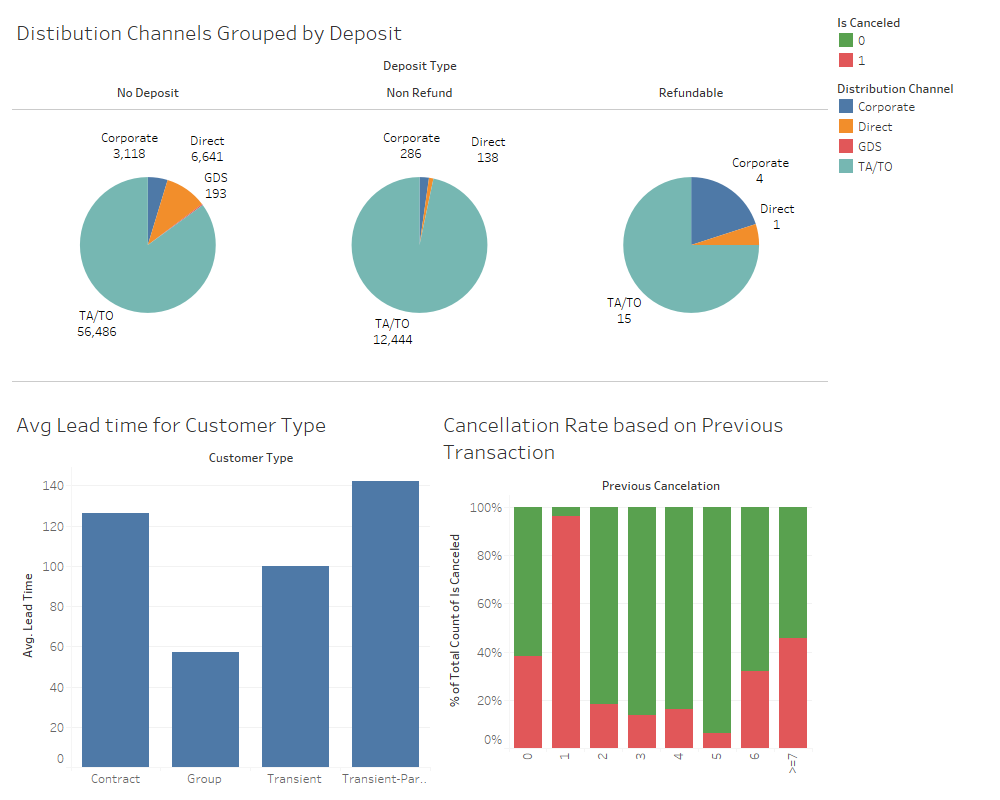
Graph 6.6b:



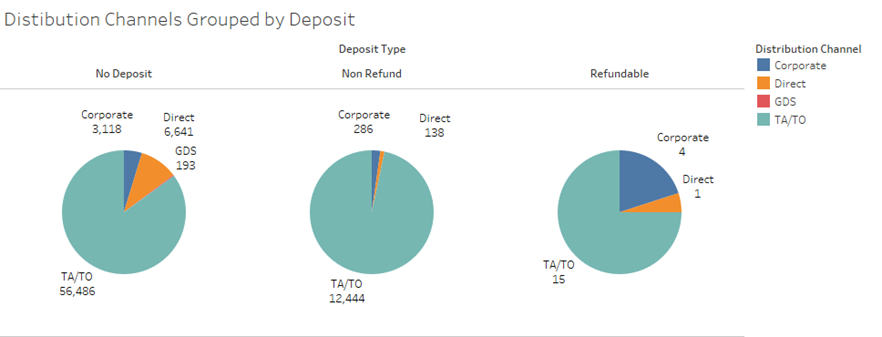
Dashboard 6.4



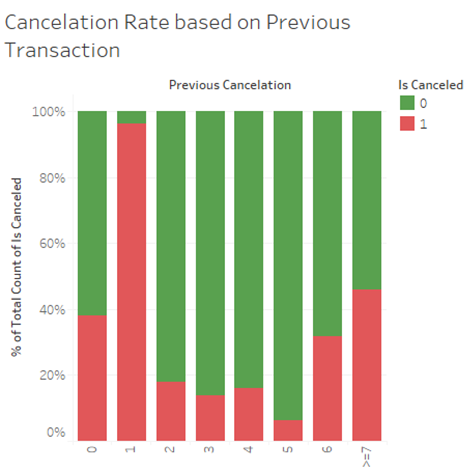
Dashboard 6.5



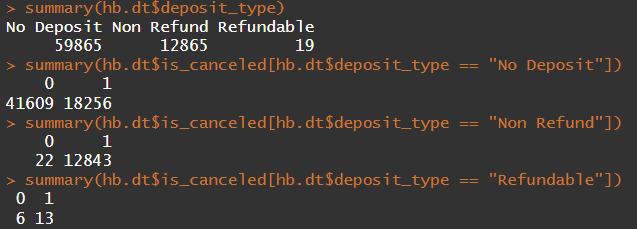
Graph 6.7



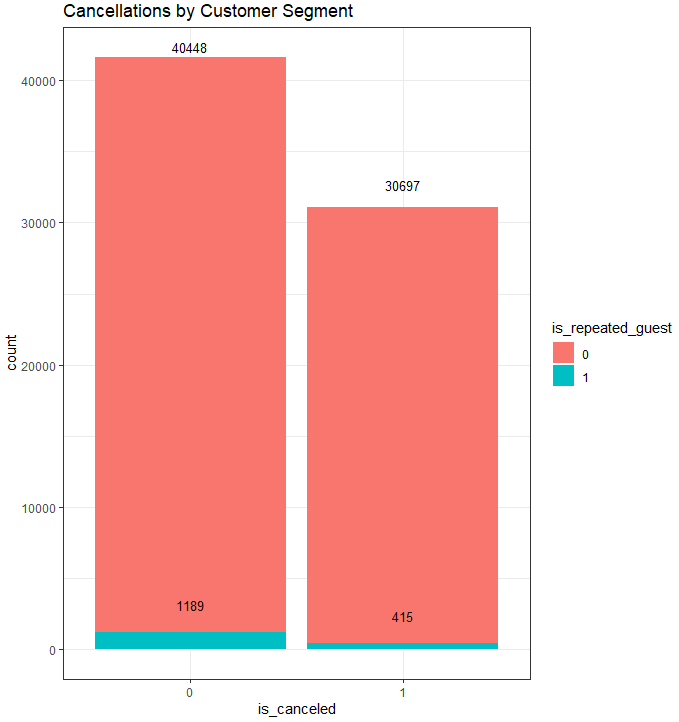
Graph 6.8

****

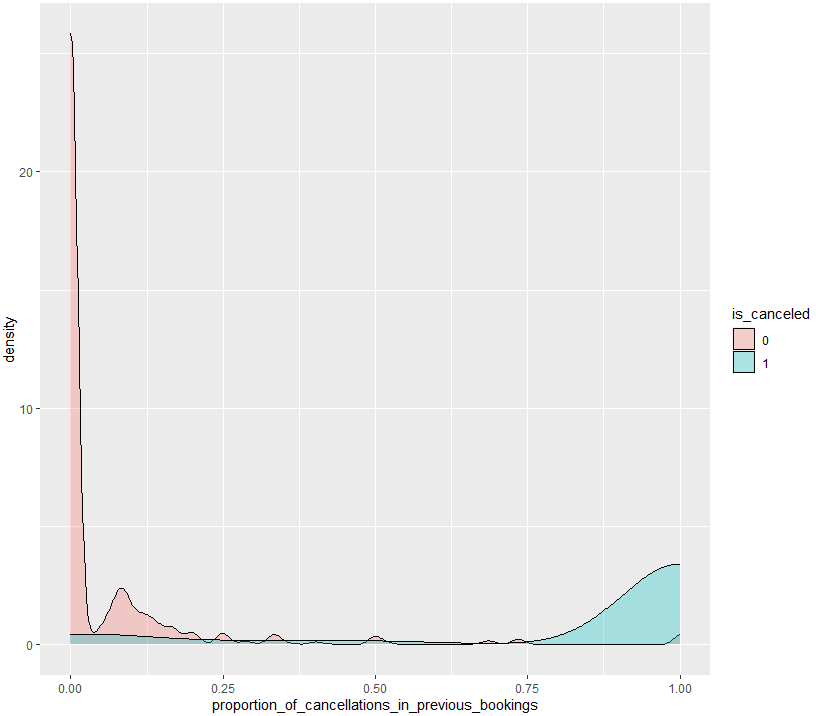
Appendix 20: Proportion of Cancellations in Deposit Types



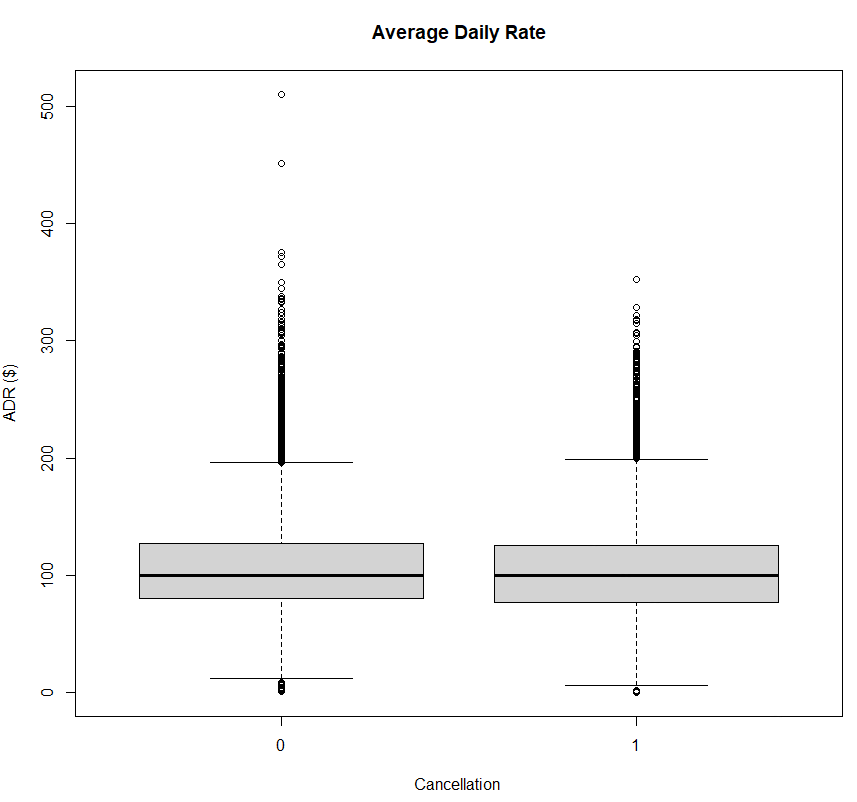
Appendix 21: Cancellation Count of Segmented Customers



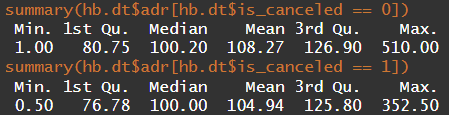
Appendix 22: Booking Patterns of Repeated Customers



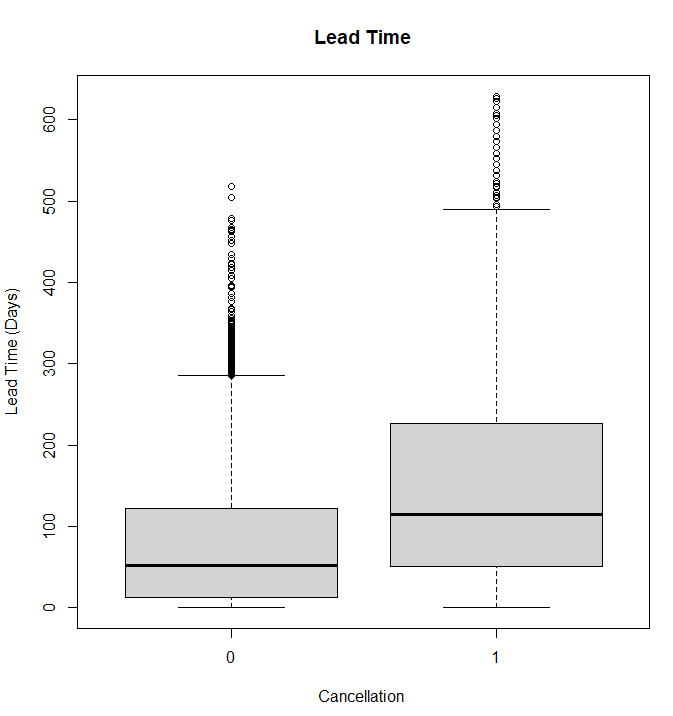
Appendix 23: Box Plot of Average Daily Rate (ADR) by Cancellation



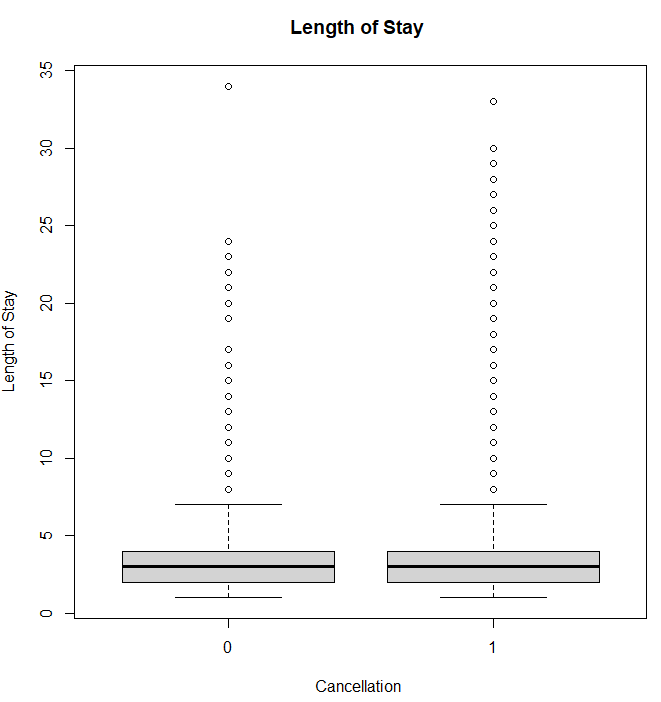
Appendix 24: Average Daily Rate (ADR) by Cancellation

****

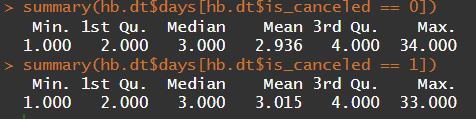
**A**Appendix 25: Box Plot of Lead Time by Cancellation



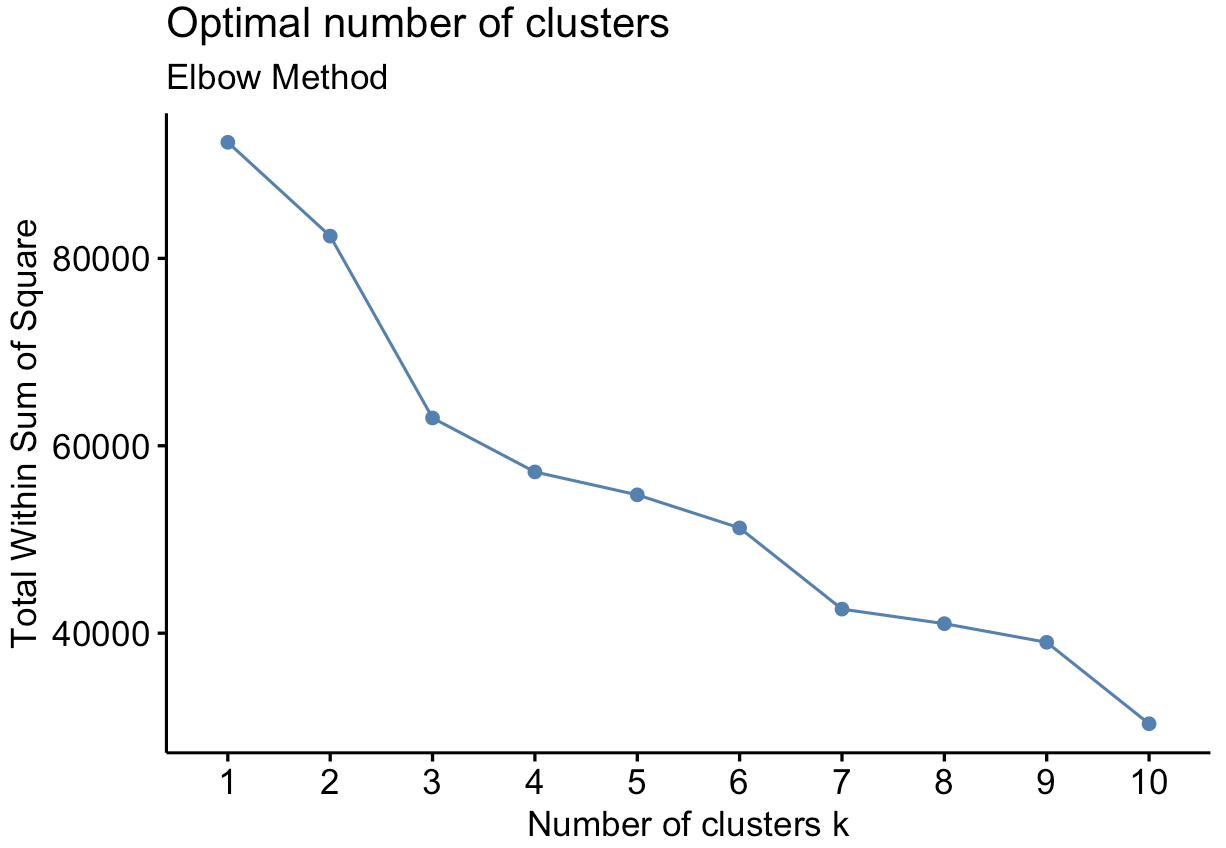
Appendix 26 : Box Plot of Length of Stay by Cancellation



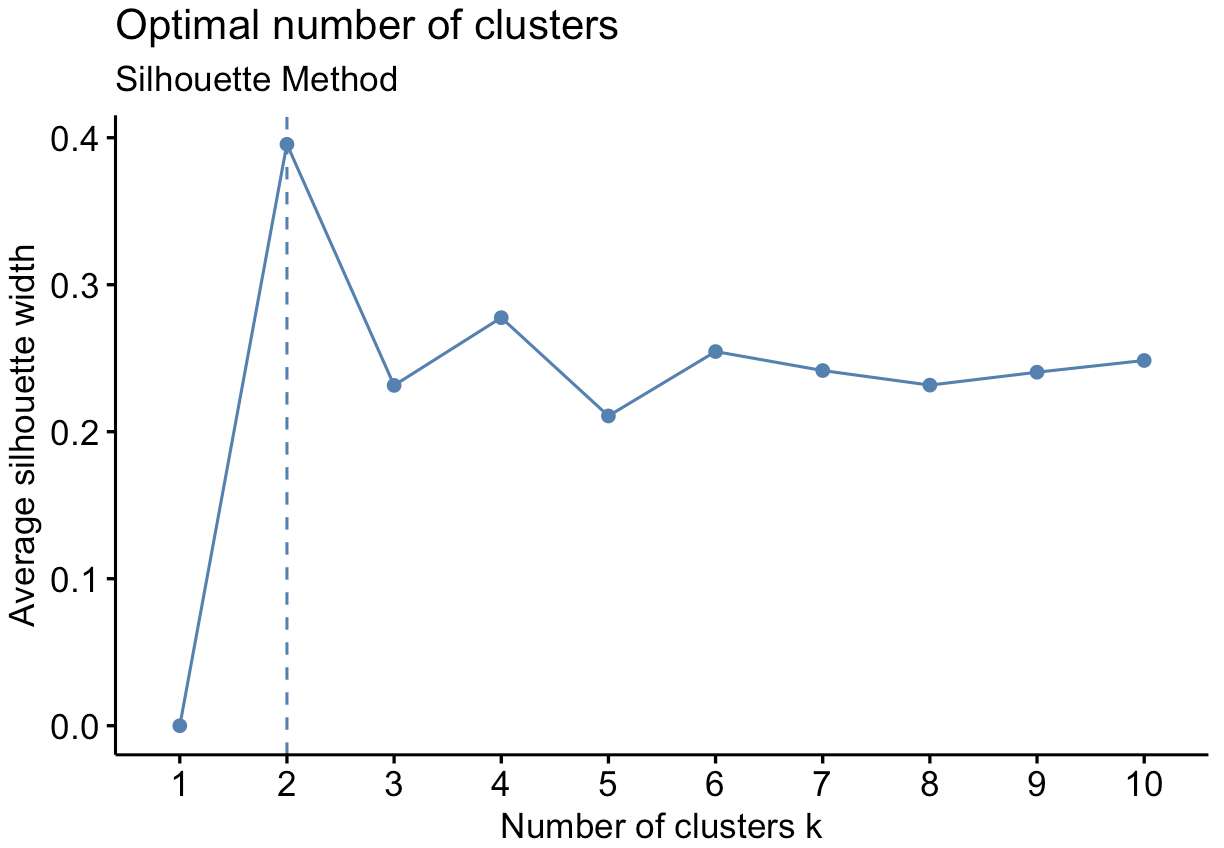
Appendix 27:



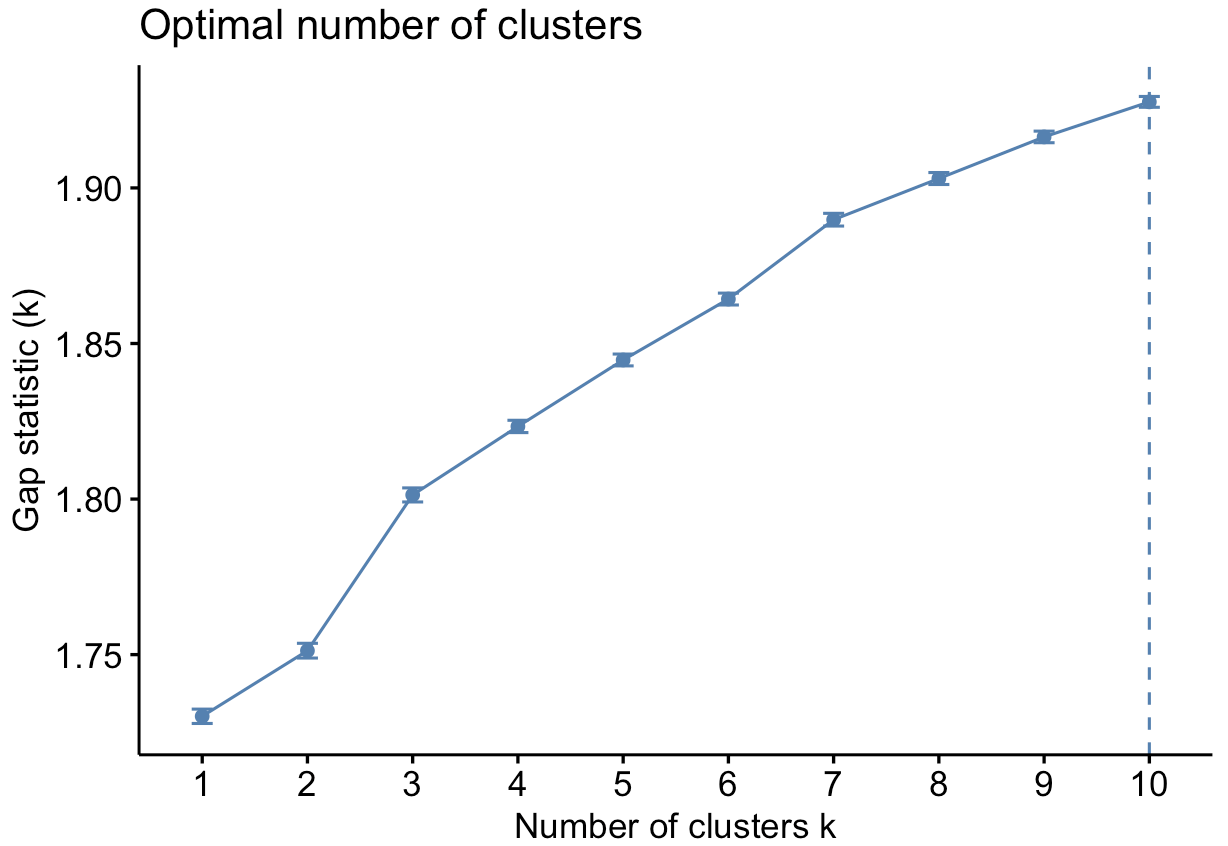
Appendix 28:



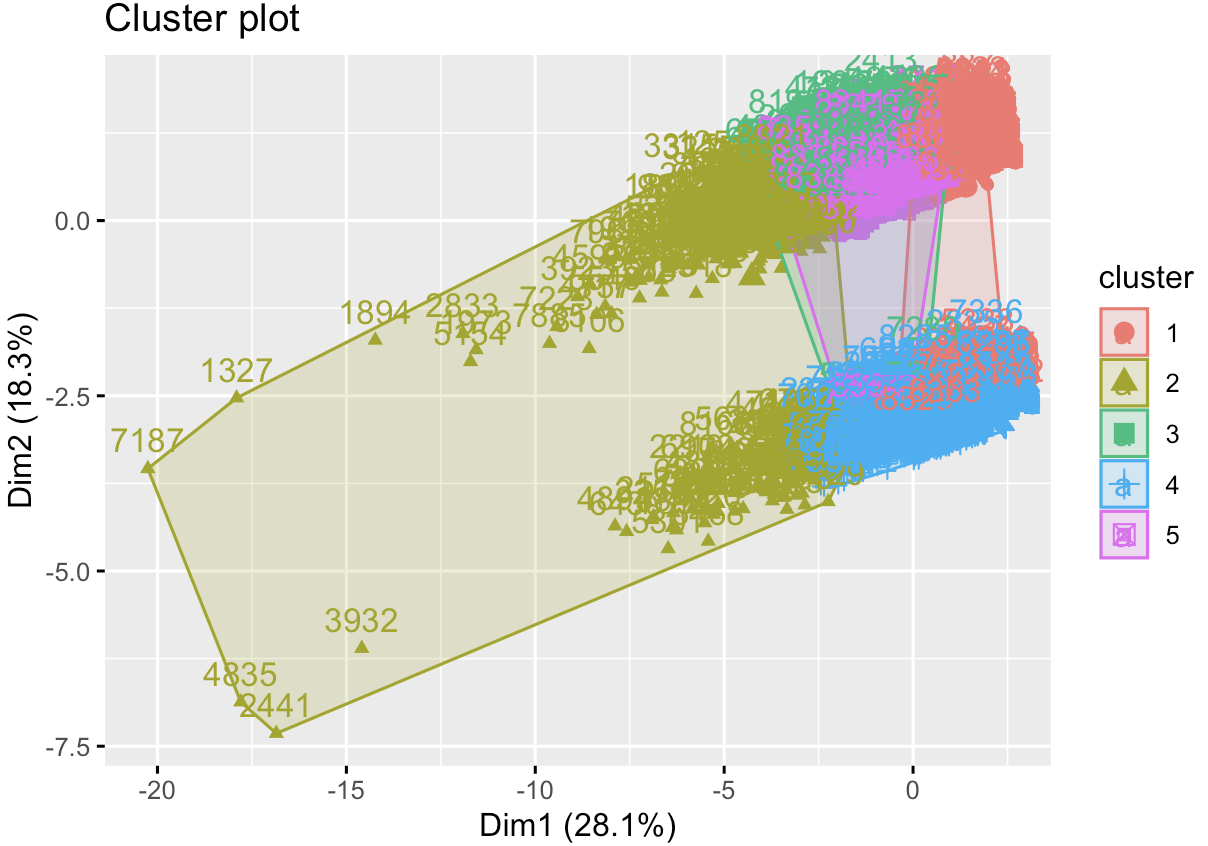
Appendix 29:



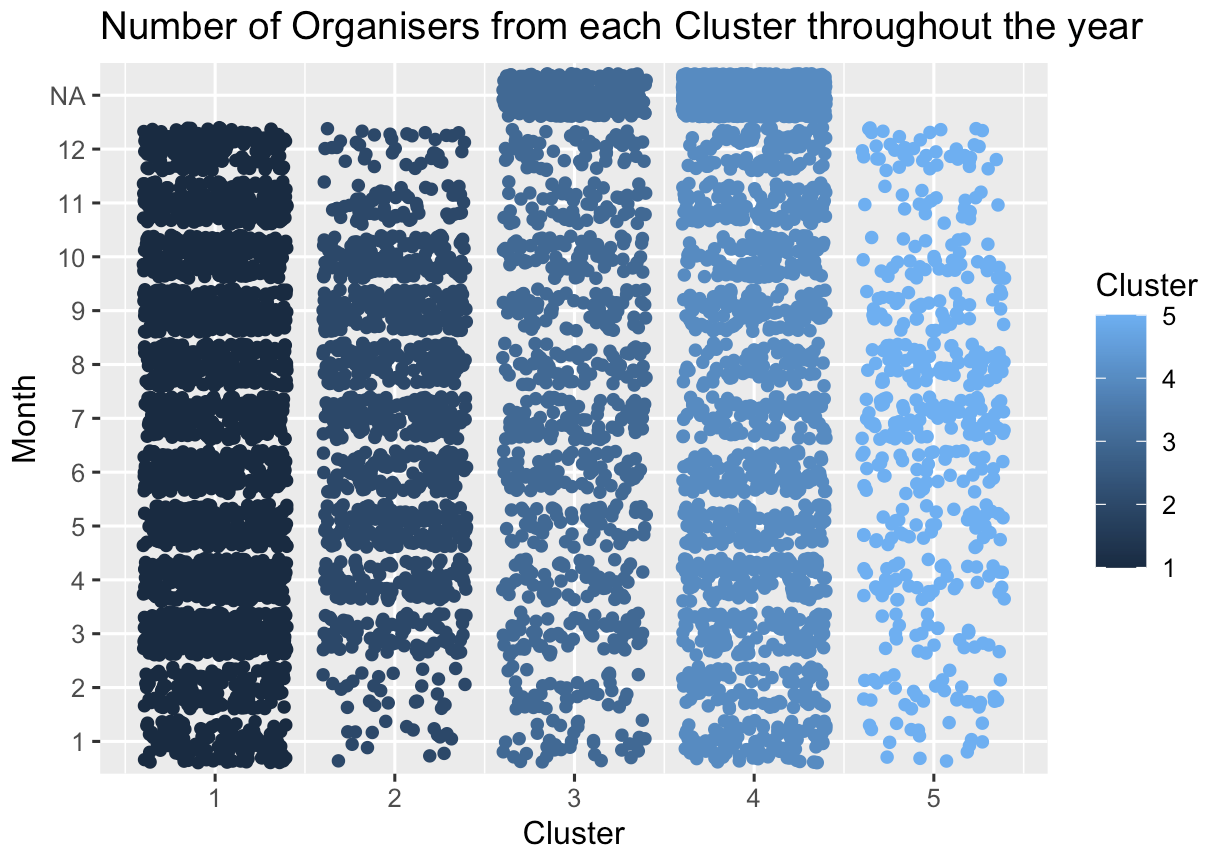
Appendix 30:



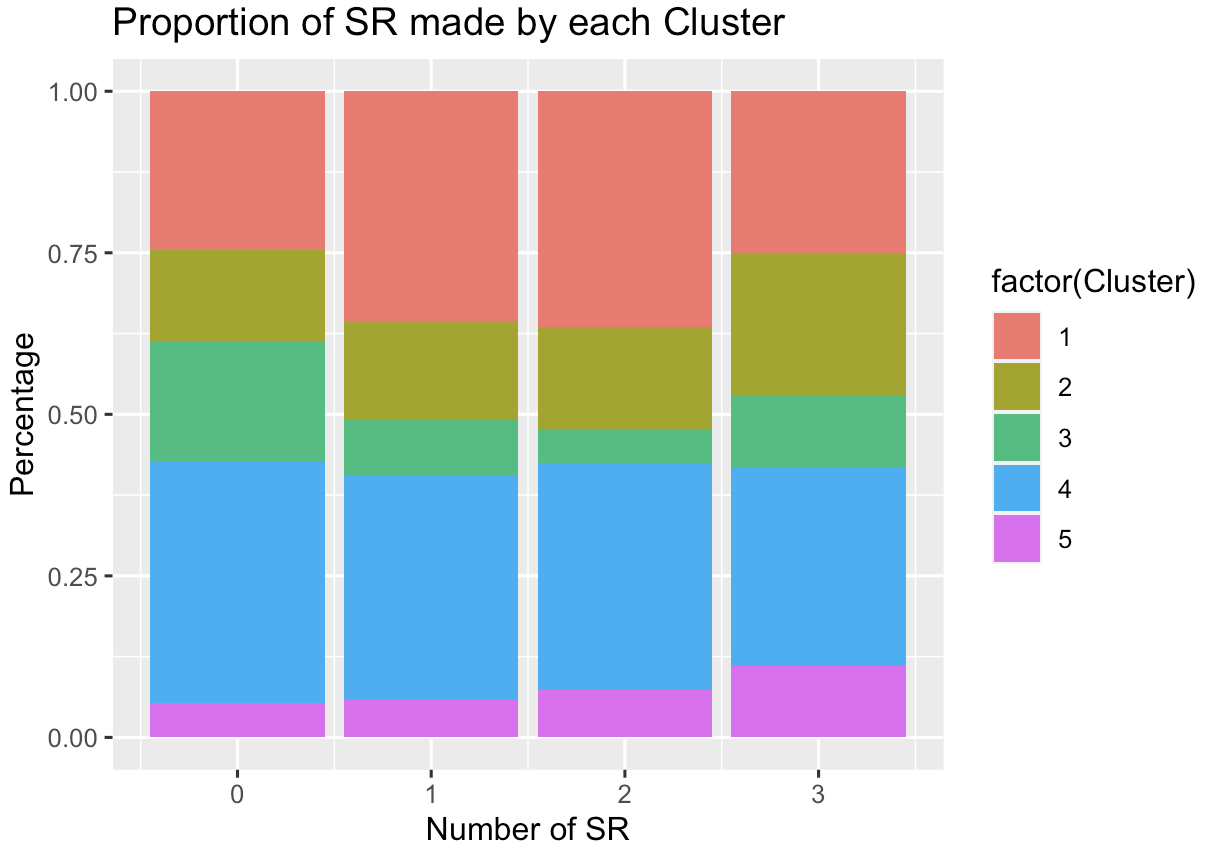
Appendix 31:



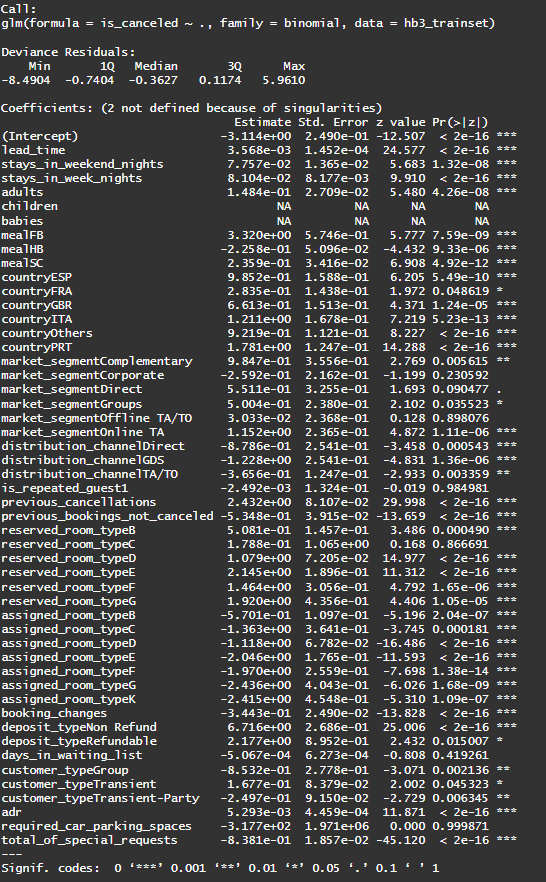
Appendix 32:



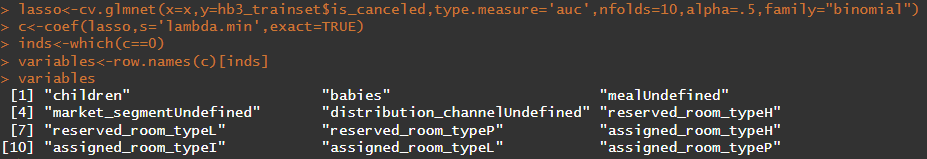
Appendix 33:



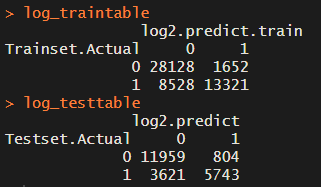
Appendix 34: Logistic Regression Model



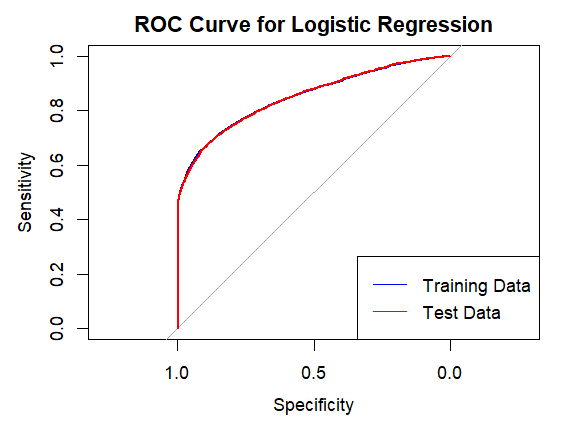
Appendix 35: Lasso Regression for Feature Selection



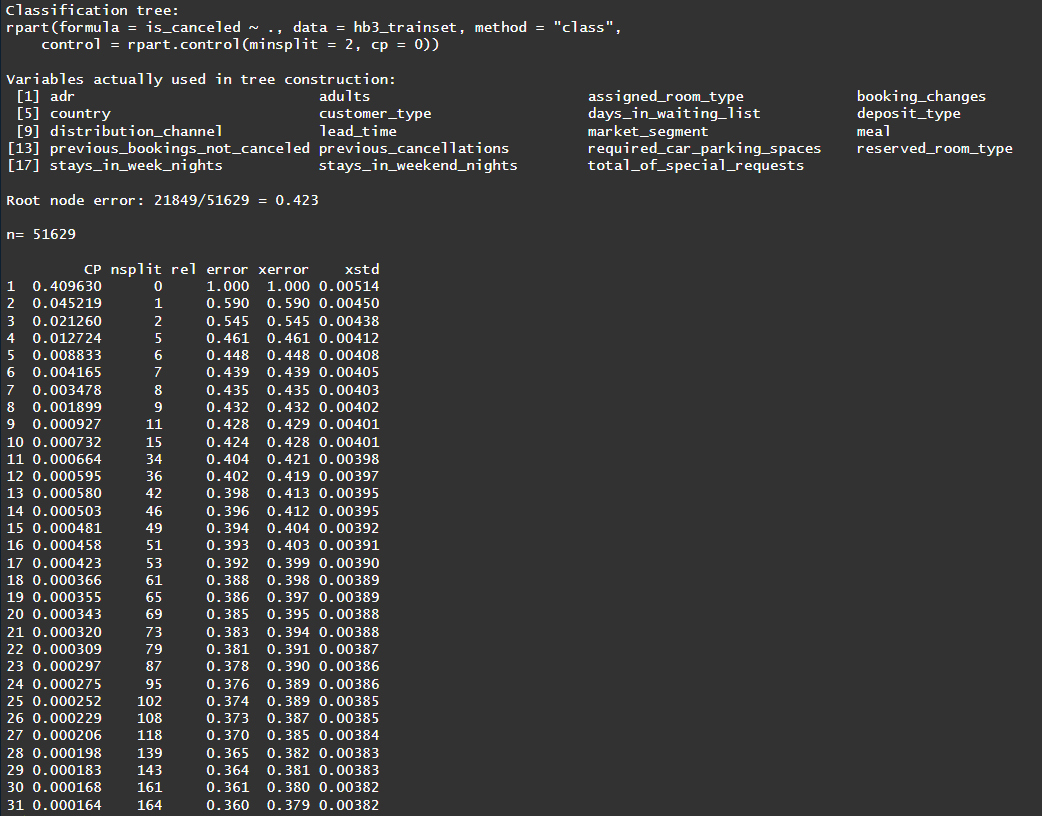
Appendix 36: Confusion Matrix for Logistic Regression Train Set and Test Set



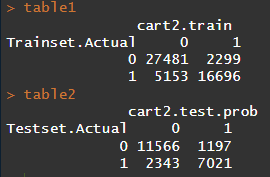
Appendix 37: ROC-AUC Curve for Logistic Regression



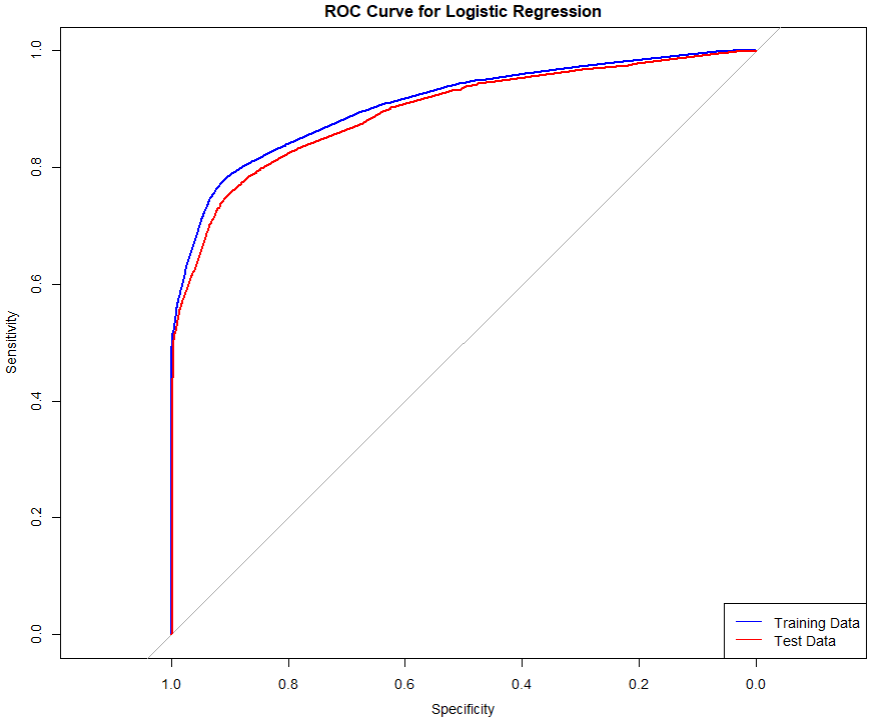
Appendix 38: Pruned CART Model



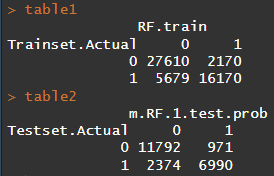
Appendix 39: Confusion Matrices of Final CART Model



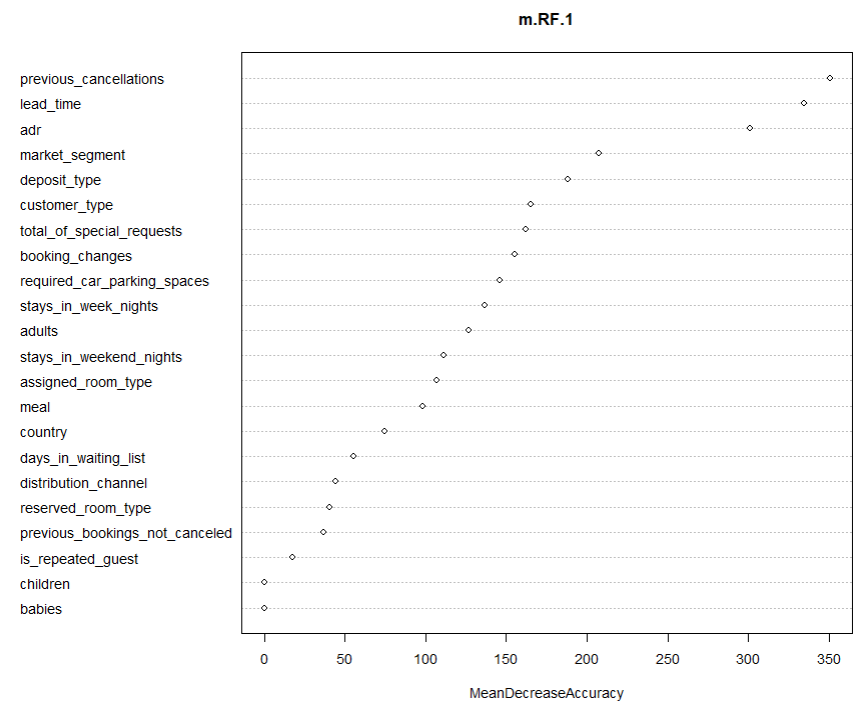
Appendix 40: ROC-AUC Curve for Final CART Model



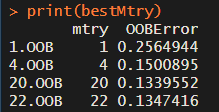
Appendix 41: Random Forest Model



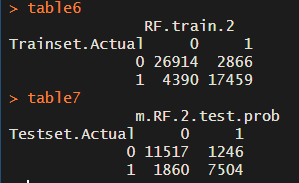
Appendix 42: Random Forest Variable Importance



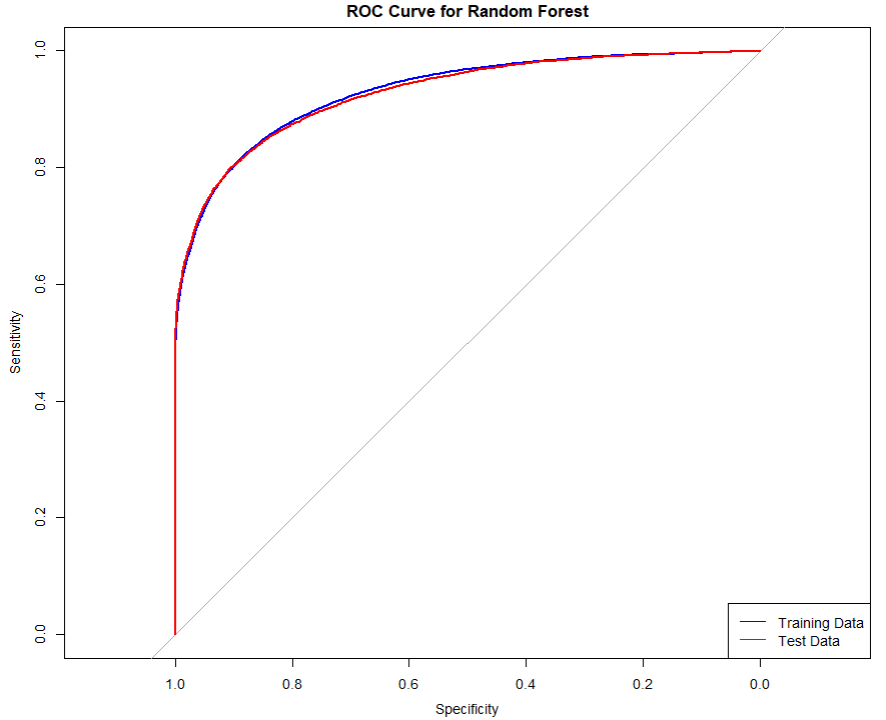
Appendix 43: Random Forest Tuning by OOBE



Appendix 44: Confusion Matrices of Tuned Random Forest



Appendix 45: AUC-ROC Curve of Tuned Random Forest



Appendix 46: Confusion Matrix Interpretation on Hotel Operations

| **True Positive** | * Impact on Operations to Charge Higher Rates or Deposit * High True Positive Rate Required for Optimal Vacancy Management |
| --- | --- |
| **True Negative** | * No Impact on Sales and Operations * High True Negative Rate Required for Optimal Revenue Management |
| **False Positive** | * Effect of “Overpredicting” Cancellations * Cancellation of customer’s confirmed booking can lead to negative customer experience and loss of business reputation |
| **False Negative** | * Greatest Impact on Sales and Operations * Time sensitivity in filling vacant rooms due to cancellations predicted wrongly |

# 11. Data Dictionary

## 11.1 Hotel Bookings Demand Dataset

**Target Variable**

| **Status** | **Data Value Description** | **Target Variable** |
| --- | --- | --- |
| ***Is\_Canceled*** | Categorical | Value indicating if the booking was cancelled (1) or not (0) |

**Booking Features**

| **Variable** | **Type** | **Description** | **Source/Engineering** |
| --- | --- | --- | --- |
| ***ADR*** | Numeric | Average Daily Rate as defined by [[5]](https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib5) | BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights |
| ***Adults*** | Integer | Number of adults | BO and BL |
| ***Agent*** | Categorical | ID of the travel agency that made the booking | BO and BL |
| ***ArrivalDateDayOfMonth*** | Integer | Day of the month of the arrival date | BO and BL |
| ***ArrivalDateMonth*** | Categorical | Month of arrival date with 12 categories: “January” to “December” | BO and BL |
| ***ArrivalDateWeekNumber*** | Integer | Week number of the arrival date | BO and BL |
| ***ArrivalDateYear*** | Integer | Year of arrival date | BO and BL |
| ***AssignedRoomType*** | Categorical | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***Babies*** | Integer | Number of babies | BO and BL |
| ***BookingChanges*** | Integer | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation | BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal |
| ***Children*** | Integer | Number of children | BO and BL/Sum of both payable and non-payable children |
| ***Company*** | Categorical | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons | BO and BL. |
| ***Country*** | Categorical | Country of origin. Categories are represented in the ISO 3155–3:2013 format | BO, BL and NT |
| ***CustomerType*** | Categorical | Type of booking, assuming one of four categories: | BO and BL |
| Contract - when the booking has an allotment or other type of contract associated to it; |
| Group – when the booking is associated to a group; |
| Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; |
| Transient-party – when the booking is transient, but is associated to at least other transient booking |
| ***DaysInWaitingList*** | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer | BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS |
| ***DepositType*** | Categorical | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: | BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking׳s arrival or cancellation date. |
| No Deposit – no deposit was made; |
| In case no payments were found the value is “No Deposit”. |
| If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”. |
| Non Refund – a deposit was made in the value of the total stay cost; |
| Otherwise the value is set as “Refundable” |
| Refundable – a deposit was made with a value under the total cost of stay. |
| ***DistributionChannel*** | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL and DC |
| ***IsRepeatedGuest*** | Categorical | Value indicating if the booking name was from a repeated guest (1) or not (0) | BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest |
| ***LeadTime*** | Integer | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date | BO and BL/ Subtraction of the entering date from the arrival date |
| ***MarketSegment*** | Categorical | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL and MS |
| ***Meal*** | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages: | BO, BL and ML |
| Undefined/SC – no meal package; |
| BB – Bed & Breakfast; |
| HB – Half board (breakfast and one other meal – usually dinner); |
| FB – Full board (breakfast, lunch and dinner) |
| ***PreviousBookingsNotCanceled*** | Integer | Number of previous bookings not cancelled by the customer prior to the current booking | BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled. |
| ***PreviousCancellations*** | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking | BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled. |
| ***RequiredCardParkingSpaces*** | Integer | Number of car parking spaces required by the customer | BO and BL |
| ***ReservationStatus*** | Categorical | Reservation last status, assuming one of three categories: | BO |
| Canceled – booking was canceled by the customer; |
| Check-Out – customer has checked in but already departed; |
| No-Show – customer did not check-in and did inform the hotel of the reason why |
| ***ReservationStatusDate*** | Date | Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel | BO |
| ***ReservedRoomType*** | Categorical | Code of room type reserved. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***StaysInWeekendNights*** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel | BO and BL/ Calculated by counting the number of weekend nights from the total number of nights |
| ***StaysInWeekNights*** | Integer | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel | BO and BL/Calculated by counting the number of week nights from the total number of nights |
| ***TotalOfSpecialRequests*** | Integer | Number of special requests made by the customer (e.g. twin bed or high floor) | BO and BL/Sum of all special requests |

## 11.2 Customer Booking Dataset

| **Variable** | **Type** | **Description** |
| --- | --- | --- |
| *ID* | Numeric | Customer ID |
| *Nationality* | Categorical | Country of origin. Categories are represented in the ISO 3155–3:2013 format [[1]](https://www.sciencedirect.com/science/article/pii/S2352340920314645?via%3Dihub#bib0001) |
| *Age* | Numeric | Customer's age (in years) at the last day of the extraction period. |
| *DaysSinceCreation* | Numerical | Number of days since the customer record was created (number of days elapsed between the creation date and the last day of the extraction period) |
| *NameHash* | Categorical | Name of the customer's SHA2–256 hash string. A hash-string is the string resulting from a mathematical function that maps a string of arbitrary length to fixed-length [[2]](https://www.sciencedirect.com/science/article/pii/S2352340920314645?via%3Dihub#bib0002). Hash functions are used for different purposes. In this case, to allow customer's anonymization. |
| *DocIDHash* | Categorical | SHA2–256 hash-string of the identification document number the customer provided at check-in (passport number, national ID card number, or other) |
| *AverageLeadTime* | Numeric | The average number of days elapsed between the customer's booking date and arrival date. In other words, this variable is calculated by dividing the sum of the number of days elapsed between the moment each booking was made and its arrival date, by the total of bookings made by the customer |
| *LodgingRevenue* | Numeric | Total amount spent on lodging expenses by the customer (in Euros). This value includes room, crib, and other related lodging expenses |
| *OtherRevenue* | Numeric | Total amount spent on other expenses by the customer (in Euros). This value includes food, beverage, spa, and other expenses |
| *BookingsCanceled* | Numeric | Number of bookings the customer made but subsequently cancelled (the customer informed the hotel he/she would not come to stay) |
| *BookingsNoShowed* | Numeric | Number of bookings the customer made but subsequently made a “no-show” (did not cancel, but did not check-in to stay at the hotel) |
| *BookingsCheckedIn* | Numeric | Number of bookings the customer made, and which end up with a staying |
| *PersonsNights* | Numeric | The total number of persons\*nights that the customer stayed at the hotel. This value is calculated by summing all customers checked-in bookings’ persons/nights. PersonNights of each booking is the result of the multiplication of the number of staying nights by the sum of adults and children |
| *RoomNights* | Numeric | Total of room\*nights the customer stayed at the hotel (checked-in bookings). RoomNights are the multiplication of the number of rooms of each booking by the number of nights of the booking |
| *DaysSinceLastStay* | Numeric | The number of days elapsed between the last day of the extraction and the customer's last arrival date (of a checked-in booking). The last day of the extraction period is December 31, 2018.  A value of −1 indicates the customer never stayed at the hotel |
| *DaysSinceFirstStay* | Numeric | The number of days elapsed between the last day of the extraction and the customer's first arrival date (of a checked-in booking). The last day of the extraction period is December 31, 2018.  A value of −1 indicates the customer never stayed at the hotel |
| *DistributionChannel* | Categorical | Distribution channel usually used by the customer to make bookings at the hotel |
| *MarketSegment* | Categorical | Current market segment of the customer |
| *SRHighFloor* | Boolean | Indication if the customer usually asks for a room on a higher floor (0: No, 1: Yes) |
| *SRLowFloor* | Boolean | Indication if the customer usually asks for a room on a lower floor (0: No, 1: Yes) |
| *SRAccessibleRoom* | Boolean | Indication if the customer usually asks for an accessible room (0: No, 1: Yes) |
| *SRMediumFloor* | Boolean | Indication if the customer usually asks for a room on a middle floor (0: No, 1: Yes) |
| *SRBathtub* | Boolean | Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes) |
| *SRShower* | Boolean | Indication if the customer usually asks for a room with a shower (0: No, 1: Yes) |
| *SRCrib* | Boolean | Indication if the customer usually asks for a crib (0: No, 1: Yes) |
| *SRKingSizeBed* | Boolean | Indication if the customer usually asks for a room with a king-size bed (0: No, 1: Yes) |
| *SRTwinBed* | Boolean | Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes) |
| *SRNearElevator* | Boolean | Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes) |
| *SRAwayFromElevator* | Boolean | Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes) |
| *SRNoAlcoholInMiniBar* | Boolean | Indication if the customer usually asks for a room with no alcohol in the mini-bar (0: No, 1: Yes) |
| *SRQuietRoom* | Boolean | Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes) |

1. Most hotels’ booking policies require the person booking to be at least 18 years of age. [↑](#footnote-ref-0)