R Term Project: Hotel Booking Analysis

Ghosha Shah

16 December, 2022

(1) Introduction

References:

- I have used following references for describing the background of this analysis.
  Saputro and Nanang (2021)
  António (2019)
  Tsai, Lin, and Lee (2022)
  Novakovic and Turina (n.d.)
  Martin-Fuentes and Mellinas (2018)
  Abrate and Viglia (2016)
  Saputro and Nanang (2021)
  Antonio, De Almeida, and Nunes (2017)
  Chalupa and Petricek (2022)
  Zheng and Forgacs (2017)
  Vives, Jacob, and Payeras (2018)
  Lee (2018)

1.1 Abstract:

- When visitors spend money in hotels, restaurants, and entertainment venues, hospitality directly contributes to local economies. The hospitality sector is expanding as more and more individuals use their money to travel and engage in leisure activities. Nevertheless, the demand for hotel rooms is not divided evenly throughout the year because people usually check into a hotel during the holidays or a special event. The hotel sector is highly unpredictable, and bookings depend on a wide range of variables, including hotel type, seasonality, day of the week, and many more. To assist the hotels in making better plans, it is now even more important to analyse the patterns present in the historical data. In order to maximize revenue, hotels can run multiple analysis using past data.

- The study seeks to learn details about how visitors behave while making hotel reservations. The hotel management frequently uses a pricing plan, one of which involves increasing the room rate when demand was strong and implementing a promotion when demand was low. As a result, it is crucial for the pricing strategy to be able to predict future demand accurately. Demand for various client segments may vary, making forecasting more complex because various segments require different models. This information can help hotels modify their consumer strategies and prepare for unforeseen circumstances.

- In a hotel, thousands of reservations are made every single month. The hotels have all the data regarding the reservation such as number of people, type of meal selected, arrival date, stay duration, type of room selected, Reservation Status, etc.

- Therefore, it can be considered that the hotels have almost all the details required to perform data analysis of the Booking System. This implies that using these features, many useful insights can be gained such as Booking time pattern, category of clients in terms of country, age, Market Segment, Distribution Channel, etc. An extremely crucial outcome which can be determined here is which customers had showed up at the arrival or updated date and which ones had cancelled their reservation.

1.2 Literature Review:

1.2.1
Stepan Chalupa and Martin Petricek, in their research paper **Understanding customer's online booking intentions using hotel big data analysis** had performed clustering to identify different types of hotel customers. The research focuses on consumer classification and its implementation towards demand planning. They adopted the Two-Step cluster analysis to determine market segments. The seven variables were used in the analysis (length of stay, average room rate, distribution channel, reservation day, day of arrival, lead time and payment conditions). "Corporates", "Early Bird Bookers", "Last Minute Bookers", "Product Seekers", "Values Seekers", and "Last Minute Bookers" were identified as segments clusters. It was found that most reservations were in cluster six on Tuesday, with the arrival day being Monday.

Findings of the study:

"Last Minute Bookers" stay in the hotel for one night With a low average lead time of 9, 27 days before arrival, "Value Seekers" search the internet for the greatest possible rates and deals. The lower average accommodation rates are mainly the result of expensive and cheap nights (with the MLOS and even NR discounts).
"Product Seekers" are seeking for a certain product whose availability is limited, primarily due to recent sales. These clients are willing to pay higher average lodging rates.
"Early Bird Bookers" plan their stays and make an effort to enjoy the benefits of the accommodation facility's low demand when they book. A cheaper pricing is associated with advanced booking and non-refundable offers. This segment is likely to have a high cancellation rate.
"Long-time Stayers" stay in the hotel for many nights and are subject to MLOS limits and, in most cases, off-season stays in the leisure segment.
"Corporates" were identified separately since their behaviour varies during the booking and stay periods. They accept the higher rates in the same time frame as leisure clients.

1.2.2
In the study **Determinants of Customer Satisfaction in the Hotel Industry**, Huiying Li, Qiang Ye and Rob Law conducted a research to to show how internet hotel reviews may be utilised to systematically identify the factors of consumer satisfaction in the hospitality domain that visitors discuss. They investigated 42,866 relevant internet reviews in the Beijing hotel sector to determine important variables of consumer satisfaction that were most apparent in rating hotels. After the characteristics influencing consumer satisfaction were identified, they did accommodation sector comparisons for budget versus luxury hotels. Although the primary objective of this study was to demonstrate how online hotel reviews can be used to identify determinants of traveller satisfaction through content analysis, the lodging segment comparisons were undertaken to provide hotel managers with guidance on how to best improve visitor satisfaction.

Findings of the study:

The most major finding of the study was the identification of 15 aspects influencing consumer satisfaction from online reviews, which were then categorised into six categories. The factors of room, transportation convenience, convenience to tourist places, and value for money were at the top of the list in terms of attracting mass customers, while parking, air conditioning, and TV were at the bottom. Customers reserving luxury hotels prioritised the room factor, whereas those booking affordable hotels prioritised the budget.

1.2.3
In a research **Application of online booking data to hotel revenue management** authors provided a model that helps hotels to obtain an optimal room charge and overbooking level for maximizing the

expected sales. Firstly, the overbooking concept for online booking systems is explained. Second, they used a discrete choice model to estimate client decision behaviors using online booking data. Third, they investigated the predicted sales maximization problem by merging the estimated choice model with the theoretical overbooking model and numerically solving the optimal overbooking level and room charge. Finally, they demonstrated numerical examples of the best overbooking tactics and room prices based on online booking data from two big luxury hotels in Tokyo's Shinjuku area.

Findings of the study:

The results, in particular, suggest concrete levels for the room charge and overbooking. For example, if the over-sale cost per room is large and the cancellation rate is low, the result demonstrates the explicit trade-off between the over-sale cost and the reimbursement for the sales loss by overbooking. In this situation, the model plays a significant role in determining the overbooking amount and the room charge.
Furthermore, we explored the gradual shift of the room charges when the hotels repeatedly optimise them in turns in the numerical case of the equilibrium room charges of the two hotels. We have seen that the prices fall to lower levels compared to the original hotel costs and the predicted sales ultimately decrease.

1.2.4
Stepan Chalupa and Martin Petricek had conducted research called **Using Technology and Customer Behaviour Characteristics to Improve Hotel Sales Performance**. The employment of this system is closely related to intense data gathering and processing. These systems' ineffective utilization may result in financial loss. In order to assess the sales performance of the chosen luxury hotel, this study connects the requirement for using customer-oriented information systems with the fundamentals of revenue management. In this study, booking time variables (the time interval between the date of booking and check.in) are used to determine customer booking behavior. This study's primary objective was to demonstrate that, in order to enhance their overall performance, hotels should make better use of their existing resources and their outputs. The use of data and the ability to process it are related to both general knowledge and professional expertise.

Findings of the study:

The number of reservations dropped down immediately as a result of the price increase. Regarding the timing of their booking decisions, clients' behaviour does not directly differ, there might be another reason behind change in customer behavior. The decline in hotel performance needs to be looked at from a different angle since there is no evidence of a change in consumer behavior. A "top-down" pricing strategy, which is frequently used to reduce demand over the long period prior to arrival, is being used by the hotel revenue manager. When the demand starts to increase specifically, that is, when the biggest demand is being seen, this technique should be adjusted. Two type of customers were classified as below:
Individual traveller: These consumers generate roughly 75% of the business and shop seven days on average before their arrival. These customers, as suggested in the literature study, are concerned with getting the best deal, while the hotel should profit from the lack of information. Customers that purchased a product package make up the majority of the second group of clients (the expenses for other services are excluded from the selling price).

1.2.5
In the study **Strategic and tactical price decisions in hotel revenue management**, Graziano Abrate and Giampaolo Viglia studied online price changing in hospitality. The tourism and hospitality sectors have seen tremendous change over the past two decades, increasing the risk for business owners looking to gain an advantage over competitors. In fact, the widespread use of new technology and the intense competition in the business have compelled operators to look for ways to enhance both their long-term goals and their practical solutions. With the introduction of new methods to readily update prices in real time over the internet, one of the major shifts has taken place in the domain of pricing. In the tourist and hospitality industries, 22% of operators have previously employed various types of price intelligence software, increasing their profitability and gross margins by 25% and 10%, respectively.

Findings of the study:

The study demonstrates the relationship between strategic and tactical aspects in dynamic pricing, which is based on three categories of variables: tangible, reputational, and contextual. Online consumer reviews,

which are more volatile than other types of reputation when compared to traditional forms of reputation, are becoming more and more important in determining prices. There are several managerial consequences. First, the value of tangible features is still significant and can be leveraged to enhance revenue managers' tactical solutions. For instance, a hotel with nearly all of its basic rooms booked many days prior to the check-in date can choose to eliminate the availability of the room, only allowing guests to book premium accommodations during that time. Other comparable sectors that have varying degrees of tangible qualities can use the same method (e.g. seats in business vs. economy class in travel companies). In addition, internet reviews are becoming more significant in the context of reputation.

1.3 Data Description:

The selected dataset for this project can be found through Kaggle: *https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand*

- The given dataset contains almost all the information regarding a particular Hotel Booking having arrival date between the 1st of July of 2015 and the 31st of August 2017. As this is the real data, it can be highly significant for data mining and revenue management.
- The data consists of around 119,390 booking transactions from 2 hotels: an anonymous city hotel from Lisbon and a resort hotel from Algarve, each having 31 attributes. Every observation represents a single hotel booking.
- The data mainly comprises of the information such as:

1. Type of hotel (Hotel/ Resort).
2. Number of days between Booking and Arrival.
3. Arrival Date including Date, Week Number, Month and Year.
4. Country of customers' origin.
5. Type of Meal selected.
6. Number of Stays in Weekend Nights as well as Week Nights.
7. Count of adults, children, and babies in the stay.
8. Market Segment as well as Distribution Channel of the guests.
9. Is the customer repeated? If yes, how many cancellations or approvals have they accounted for?
10. Reserved and Assigned room type.
11. Deposit Type.
12. Reservation Status (Canceled/ Check-out).

- Description of all the attributes along with the data type can be explained as follows:

Variable	Type	Description
Hotel	Categorical	Type of hotel (Hotel/Resort)
isCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)
leadTime	Integer	Number of days between the entering date of the booking into the PMS and the arrival date
arrivalDateYear	Integer	Year of arrival date
arrivalDateMonth	Categorical	Month of arrival date (January to December)
arrivalDateweekNumber	Integer	Week number of the arrival date
arrivalDateDayOfMonth	Integer	Day of the month of the arrival date
staysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
staysInWeekNights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Integer	Number of adults
children	Integer	Number of children
babies	Integer	Number of babies
meal	Categorical	Type of meal booked Undefined/SC, BB, HB , FB
country	Categorical	Country of origin
marketSegment	Categorical	Market segment designation. In categories, the term TA means Travel Agents and TO means TourOperators
distributionChannel	Categorical	Booking distribution channel. The term TA means Travel Agents and TO means Tour Operators
isRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
previousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking
previousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking
reservedRoomType	Categorical	Code of room type reserved
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

Variable	Type	Description
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
depositType	Categorical	Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
agent	Integer	ID of the travel agency that made the booking
company	Integer	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
daysInWaitingList	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
customerType	Categorical	Type of booking, assuming one of four categories: 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
requiredCarParkingSpaces	Integer	Number of car parking spaces required by the customer
totalOfSpecialRequests	Integer	Number of special requests made by the customer
reservationStatus	Categorical	Reservation last status, assuming one of three categories: 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

- *Vitalness of the data:*
  - Descriptive analysis can be used to understand patterns, trends, and anomalies in data.
  - Variety of factors can be predicted such as seasonality, customer satisfaction, type of meal, booking cancellation prediction, categories of clients, etc.
  - Researchers in machine learning can utilise the dataset to compare how well various algorithms perform when approaching a particular task (such as classification, segmentation)
  - Educators can use the datasets to solve data mining or classification problems using machine learning.

- Hereby, I will be using this dataset to predict the following results:

1. Do adults contribute more to the Number of stays in weekends at the Hotel?
2. Analysis of Booking Repetition by month of Arrival.
3. What was the Cancellation rate of Bookings from distinct Market Segments?

(2) Method

- We will perform Empirical Analysis on the dataset using various mentioned factors to observe whether they affect the booking or cancellation rate of customers.
- As this is a massive dataset, there are several approaches through which we can perform analysis on the data. Few instances of dataset is shown below:

```
##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0       342              2015               July
## 2 Resort Hotel           0       737              2015               July
## 3 Resort Hotel           0         7              2015               July
## 4 Resort Hotel           0        13              2015               July
## 5 Resort Hotel           0        14              2015               July
## 6 Resort Hotel           0        14              2015               July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                       0
## 2                       27                         1                       0
## 3                       27                         1                       0
## 4                       27                         1                       0
## 5                       27                         1                       0
## 6                       27                         1                       0
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2        0      0   BB     PRT         Direct
## 2                    0      2        0      0   BB     PRT         Direct
## 3                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                     0
## 2               Direct                 0                     0
## 3               Direct                 0                     0
## 4            Corporate                 0                     0
```

```
## 5               TA/TO               0                     0
## 6               TA/TO               0                     0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                 C                 C
## 2                              0                 C                 C
## 3                              0                 A                 C
## 4                              0                 A                 A
## 5                              0                 A                 A
## 6                              0                 A                 A
##   booking_changes deposit_type agent company days_in_waiting_list customer_type
## 1               3   No Deposit  NULL    NULL                    0     Transient
## 2               4   No Deposit  NULL    NULL                    0     Transient
## 3               0   No Deposit  NULL    NULL                    0     Transient
## 4               0   No Deposit   304    NULL                    0     Transient
## 5               0   No Deposit   240    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   adr required_car_parking_spaces total_of_special_requests reservation_status
## 1   0                            0                         0          Check-Out
## 2   0                            0                         0          Check-Out
## 3  75                            0                         0          Check-Out
## 4  75                            0                         0          Check-Out
## 5  98                            0                         1          Check-Out
## 6  98                            0                         1          Check-Out
##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
## 3              2015-07-02
## 4              2015-07-02
## 5              2015-07-03
## 6              2015-07-03
```

2.1 Method 1

We will find out if there is association between the features "adults" and "stays_in_week_nights". Correlation gives the measure of association between two variables and it ranges from -1 (perfect negative correlation) and 1 (perfect positive correlation).

Null Hypothesis : There is no association between number of adults and total stays in week nights.
Alternatove Hypothesis : There is a significant association between number of adults and total stays in week nights.

We have performed correlation test using Pearson's method.

```
# correlation between number of weekend stays and number of adults
hotel_bookings$stays_in_week_nights <- as.numeric(as.character(hotel_bookings$stays_in_week_nights))
hotel_bookings$adults <- as.numeric(as.character(hotel_bookings$adults))
result1 <- cor.test(hotel_bookings$stays_in_weekend_nights,hotel_bookings$adults,method = "pearson")
```

2.2 Method 2

One way ANOVA test to determine if the month of Arrival affects the repetition of guests.

```
#one way anova between independent variable arrival_date_month and dependent variable is_repeated_guest
Ind.ANOVA2 <- aov(is_repeated_guest ~ arrival_date_month, hotel_bookings)
result2 <- Anova(Ind.ANOVA2, type = "III")
```

ANOVA, which stands for Analysis of Variance, is a parametric test used to analyze the mean value of samples to check whether they are significantly different. In addition. this test uses only one independent variable. This test compares means of groups (usually 3 to 4) to analyze the variance.

We will perform one-way ANOVA test between is_repeated_guest and arrival_date_month variable to find out if guests are repeated in different proportions during distinct Arrival months.

Null Hypothesis: There is no difference in guest repetition for all Arrival months.
Alternate Hypothesis: There is a significant difference in guest repetition for all Arrival months.

After conducting the test, Post hoc test was performed to find differences between the groups.

Moreover, following assumptions for one way ANOVA test were checked:

1. Normality: Residuals should form normally distributed data.
   A histogram is plotted for residuals in order to check normality.
2. Independence of data: It is difficult to check independence of data. Most of the times, observations are taken randomly and are not related to each other. Therefore, ANOVA assumes that the observations are independent of each other.
3. Homogeneity of variance : Equal variance between independent groups.
   Levene Test of Variance was performed for checking this assumption.

From all these steps, it can be discovered whether Arrival month affects repetition of guests.

2.3 Method 3
One way ANOVA test to determine if the Market Segment of customers affects the cancellation rate.

ANOVA, which stands for Analysis of Variance, is a parametric test used to analyze the mean value of samples to check whether they are significantly different. In addition. this test uses only one independent variable. This test compares means of groups (usually 3 to 4) to analyze the variance.

We will perform one-way ANOVA test between is_canceled and market_segment variable to find out if guests of one particular category of market_segment has significantly accounted for more cancellations than other categories.

Null Hypothesis: There is no difference in cancellation rate of all the Market Segments.
Alternate Hypothesis: There is a difference in cancellation rate of all the Market Segments.

```
Ind.ANOVA3 <- aov(is_canceled ~ market_segment, hotel_bookings)
result3 <- Anova(Ind.ANOVA3, type = "III")
```

After conducting the test, Post hoc test was performed to find differences between the groups.

Moreover, following assumptions for one way ANOVA test were checked:

1. Normality: Residuals should form normally distributed data.
   A histogram is plotted for residuals in order to check normality.
2. Independence of data: It is difficult to check independence of data. Most of the times, observations are taken randomly and are not related to each other. Therefore, ANOVA assumes that the observations are independent of each other.
3. Homogeneity of variance : Equal variance between independent groups.
   Levene Test of Variance was performed for checking this assumption.

From all these steps, it can be discovered whether Market Segment of the clients affects cancellation rate.

(3) Results

3.1 Result 1

The correlation function is implemented to find the correlation between Number of stays in weekend and Number of adults.

```
##
##  Pearson's product-moment correlation
##
## data:  hotel_bookings$stays_in_weekend_nights and hotel_bookings$adults
## t = 31.879, df = 119388, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08624359 0.09749259
## sample estimates:
##       cor
## 0.09187102
```

Results of the Correlation Test are as follows:
T = 31.879, degree of freedom (df) = 119388, $p - value < 2.2e - 16 < 0.05$. The value of correlation is obtained which is equal to 0.0919.

3.2 Result 2

Results of one way ANOVA between is_repeated_guest and arrival_date_month were as under:

```
## Anova Table (Type III tests)
##
## Response: is_repeated_guest
##                    Sum Sq     Df F value    Pr(>F)
## (Intercept)           7.2      1 235.087 < 2.2e-16 ***
## arrival_date_month   20.9     11  61.719 < 2.2e-16 ***
## Residuals          3667.6 119378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post hoc test was performed to determine the difference between groups. The results of this test are shown here.

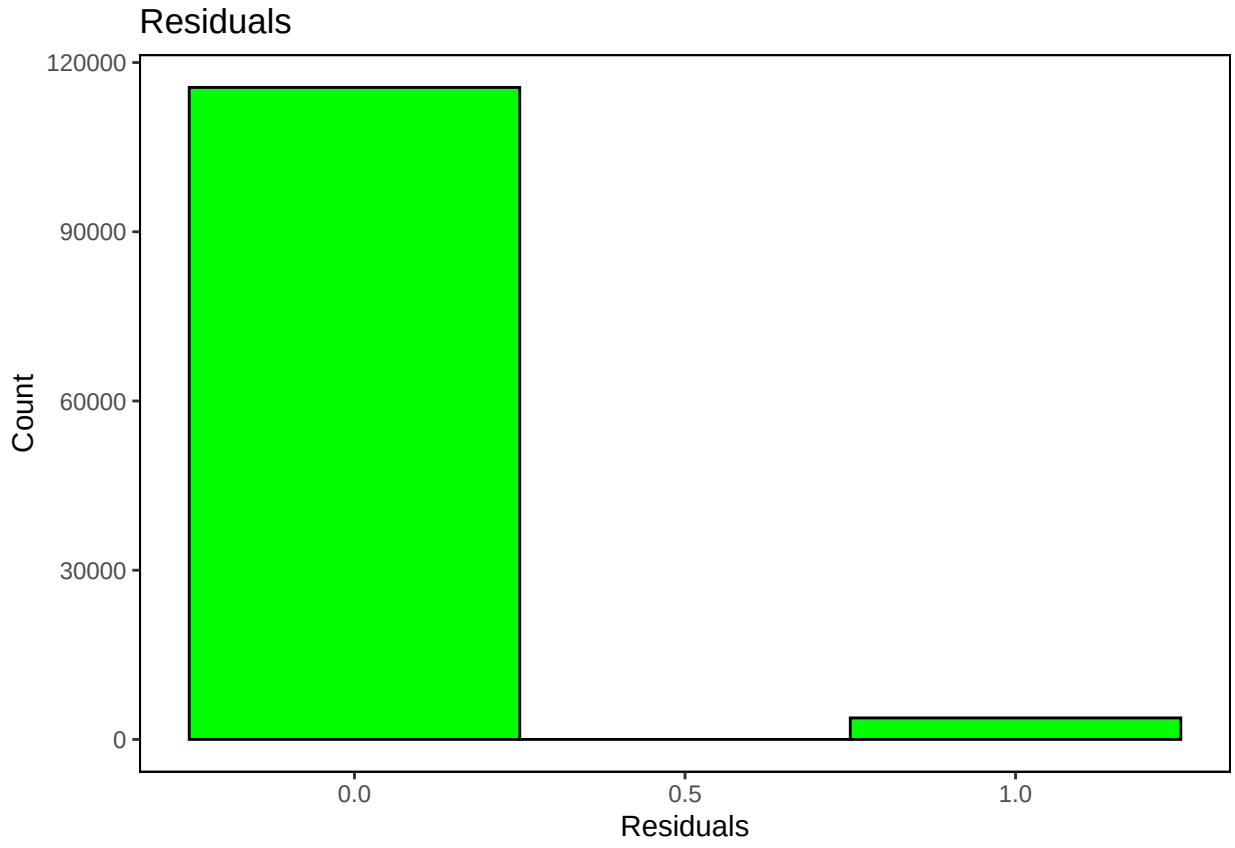
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = is_repeated_guest ~ arrival_date_month, data = hotel_bookings)
##
## $arrival_date_month
##                              diff           lwr          upr     p adj
## August-April       -0.0094510307 -1.674711e-02 -0.002154955 0.0013843
## December-April      0.0207918980  1.196114e-02  0.029622655 0.0000000
## February-April      0.0254212067  1.703929e-02  0.033803126 0.0000000
## January-April       0.0414382286  3.222257e-02  0.050653883 0.0000000
## July-April         -0.0017469928 -9.197065e-03  0.005703079 0.9998154
## June-April          0.0002585353 -7.460471e-03  0.007977541 1.0000000
## March-April         0.0179752316  1.003233e-02  0.025918131 0.0000000
```

```
## May-April           0.0055198378 -2.057477e-03  0.013097153 0.4192301
## November-April      0.0176055015  8.780393e-03  0.026430610 0.0000000
## October-April       0.0025258086 -5.154626e-03  0.010206243 0.9957082
## September-April    -0.0063925032 -1.419079e-02  0.001405786 0.2366699
## December-August     0.0302429287  2.175543e-02  0.038730429 0.0000000
## February-August     0.0348722373  2.685276e-02  0.042891712 0.0000000
## January-August      0.0508892593  4.200198e-02  0.059776536 0.0000000
## July-August         0.0077040378  6.642246e-04  0.014743851 0.0181003
## June-August         0.0097095660  2.385738e-03  0.017033394 0.0008976
## March-August        0.0274262623  1.986683e-02  0.034985697 0.0000000
## May-August          0.0149708685  7.796532e-03  0.022145205 0.0000000
## November-August     0.0270565322  1.857491e-02  0.035538156 0.0000000
## October-August      0.0119768393  4.693675e-03  0.019260003 0.0000050
## September-August    0.0030585275 -4.348814e-03  0.010465869 0.9725004
## February-December   0.0046293087 -4.807932e-03  0.014066549 0.9084643
## January-December    0.0206463306  1.046138e-02  0.030831282 0.0000000
## July-December      -0.0225388908 -3.115913e-02 -0.013918652 0.0000000
## June-December      -0.0205333627 -2.938706e-02 -0.011679663 0.0000000
## March-December     -0.0028166664 -1.186623e-02  0.006232897 0.9973591
## May-December       -0.0152720602 -2.400250e-02 -0.006541616 0.0000007
## November-December  -0.0031863965 -1.301938e-02  0.006646583 0.9962268
## October-December   -0.0182660894 -2.708618e-02 -0.009445997 0.0000000
## September-December -0.0271844012 -3.610731e-02 -0.018261495 0.0000000
## January-February    0.0160170220  6.218679e-03  0.025815365 0.0000060
## July-February      -0.0271681995 -3.532803e-02 -0.019008370 0.0000000
## June-February      -0.0251626714 -3.356876e-02 -0.016756584 0.0000000
## March-February     -0.0074459751 -1.605811e-02  0.001166161 0.1691206
## May-February       -0.0199013689 -2.817754e-02 -0.011625201 0.0000000
## November-February  -0.0078157052 -1.724766e-02  0.001616251 0.2220136
## October-February   -0.0228953981 -3.126608e-02 -0.014524716 0.0000000
## September-February -0.0318137098 -4.029266e-02 -0.023334762 0.0000000
## July-January       -0.0431852215 -5.219935e-02 -0.034171092 0.0000000
## June-January       -0.0411796933 -5.041733e-02 -0.031942052 0.0000000
## March-January      -0.0234629970 -3.288853e-02 -0.014037467 0.0000000
## May-January        -0.0359183908 -4.503797e-02 -0.026798815 0.0000000
## November-January   -0.0238327271 -3.401278e-02 -0.013652672 0.0000000
## October-January    -0.0389124200 -4.811786e-02 -0.029706984 0.0000000
## September-January  -0.0478307318 -5.713472e-02 -0.038526739 0.0000000
## June-July           0.0020055282 -5.471724e-03  0.009482780 0.9993277
## March-July          0.0197222244  1.201405e-02  0.027430395 0.0000000
## May-July            0.0072668307 -6.405948e-05  0.014597721 0.0545682
## November-July       0.0193524944  1.073804e-02  0.027966947 0.0000000
## October-July        0.0042728014 -3.164626e-03  0.011710229 0.7735319
## September-July     -0.0046455103 -1.220458e-02  0.002913561 0.6878652
## March-June          0.0177166963  9.748298e-03  0.025685095 0.0000000
## May-June            0.0052613025 -2.342738e-03  0.012865343 0.5046907
## November-June       0.0173469662  8.498900e-03  0.026195032 0.0000000
## October-June        0.0022672733 -5.439529e-03  0.009974076 0.9984131
## September-June     -0.0066510385 -1.447530e-02  0.001173222 0.1893069
## May-March          -0.0124553938 -2.028662e-02 -0.004624172 0.0000131
## November-March     -0.0003697301 -9.413782e-03  0.008674322 1.0000000
## October-March      -0.0154494230 -2.338046e-02 -0.007518383 0.0000000
## September-March    -0.0243677348 -3.241296e-02 -0.016322510 0.0000000
## November-May        0.0120856637  3.360933e-03  0.020810394 0.0003724
```

```
## October-May         -0.0029940292 -1.055891e-02  0.004570854 0.9801299
## September-May       -0.0119123410 -1.959685e-02 -0.004227831 0.0000262
## October-November    -0.0150796929 -2.389413e-02 -0.006265256 0.0000015
## September-November -0.0239980047 -3.291532e-02 -0.015080688 0.0000000
## September-October   -0.0089183118 -1.670452e-02 -0.001132102 0.0099100
```

Assumptions:

Histogram of residuals can be shown as below:



On performing Levene Test for Variance, we got results as:

```
## # A tibble: 1 x 4
##     df1    df2 statistic         p
##   <int>  <int>     <dbl>     <dbl>
## 1    11 119378      61.7 4.55e-138
```

3.3 Result 3

Results of one way ANOVA between is_canceled and market_segment were as under:

```
## Anova Table (Type III tests)
##
## Response: is_canceled
##                 Sum Sq     Df  F value    Pr(>F)
## (Intercept)       11.4      1   52.668 3.973e-13 ***
## market_segment  1981.6      7 1306.817 < 2.2e-16 ***
```

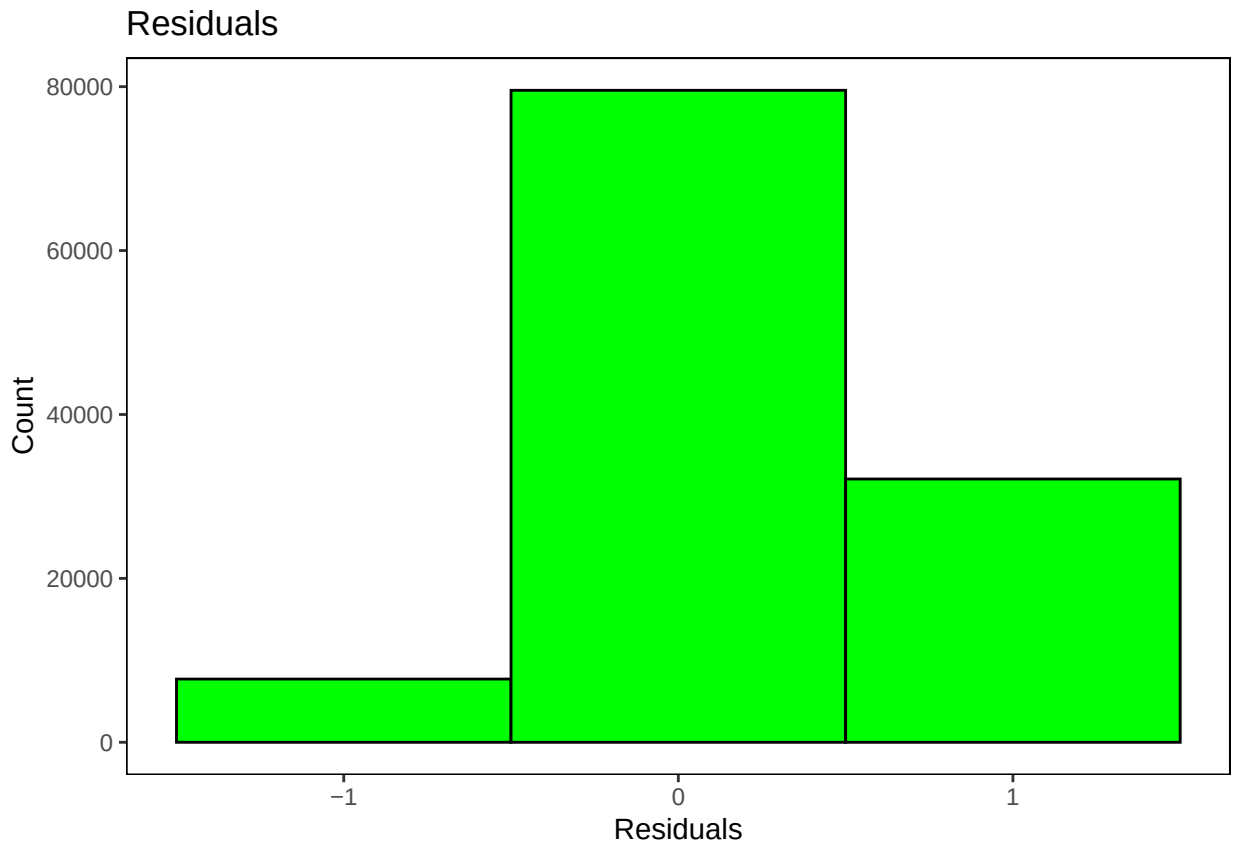
```
## Residuals      25861.1 119382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post hoc test was performed to determine the difference between groups. The results of this test are shown here.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = is_canceled ~ market_segment, data = hotel_bookings)
##
## $market_segment
##                                    diff          lwr         upr     p adj
## Complementary-Aviation      -0.08885747 -0.194094105  0.01637917 0.1710920
## Corporate-Aviation          -0.03206273 -0.125723223  0.06159776 0.9687598
## Direct-Aviation             -0.06599028 -0.158479882  0.02649933 0.3747430
## Groups-Aviation              0.39121108  0.299032361  0.48338980 0.0000000
## Offline TA/TO-Aviation       0.12375105  0.031671547  0.21583054 0.0012029
## Online TA-Aviation           0.14780215  0.055977840  0.23962646 0.0000293
## Undefined-Aviation           0.78059072 -0.221096866  1.78227830 0.2603236
## Corporate-Complementary      0.05679474  0.001530818  0.11205865 0.0389137
## Direct-Complementary         0.02286719 -0.030388231  0.07612261 0.8986741
## Groups-Complementary         0.48006855  0.427354896  0.53278219 0.0000000
## Offline TA/TO-Complementary  0.21260851  0.160068558  0.26514847 0.0000000
## Online TA-Complementary      0.23665961  0.184568191  0.28875104 0.0000000
## Undefined-Complementary      0.86944818 -0.129381046  1.86827741 0.1422672
## Direct-Corporate            -0.03392755 -0.057029012 -0.01082608 0.0002292
## Groups-Corporate             0.42327381  0.401450274  0.44509734 0.0000000
## Offline TA/TO-Corporate      0.15581378  0.134413199  0.17721435 0.0000000
## Online TA-Corporate          0.17986488  0.159590406  0.20013935 0.0000000
## Undefined-Corporate          0.81265345 -0.185022536  1.81032943 0.2088587
## Groups-Direct                0.45720136  0.441129461  0.47327325 0.0000000
## Offline TA/TO-Direct         0.18974132  0.174248621  0.20523402 0.0000000
## Online TA-Direct             0.21379242  0.199896622  0.22768823 0.0000000
## Undefined-Direct             0.84658099 -0.150985749  1.84414774 0.1660900
## Offline TA/TO-Groups        -0.26746003 -0.280973462 -0.25394661 0.0000000
## Online TA-Groups            -0.24340893 -0.255057206 -0.23176066 0.0000000
## Undefined-Groups             0.38937964 -0.608158329  1.38691760 0.9369569
## Online TA-Offline TA/TO      0.02405110  0.013215961  0.03488624 0.0000000
## Undefined-Offline TA/TO      0.65683967 -0.340689131  1.65436847 0.4846844
## Undefined-Online TA          0.63278857 -0.364716710  1.63029385 0.5350861
```

Assumptions:

Histogram of residuals can be shown as below:



On performing Levene Test for Variance, we got results as:

```
## # A tibble: 1 x 4
##     df1    df2 statistic     p
##   <int>  <int>     <dbl> <dbl>
## 1     7 119382      451.     0
```

(4) Discussion

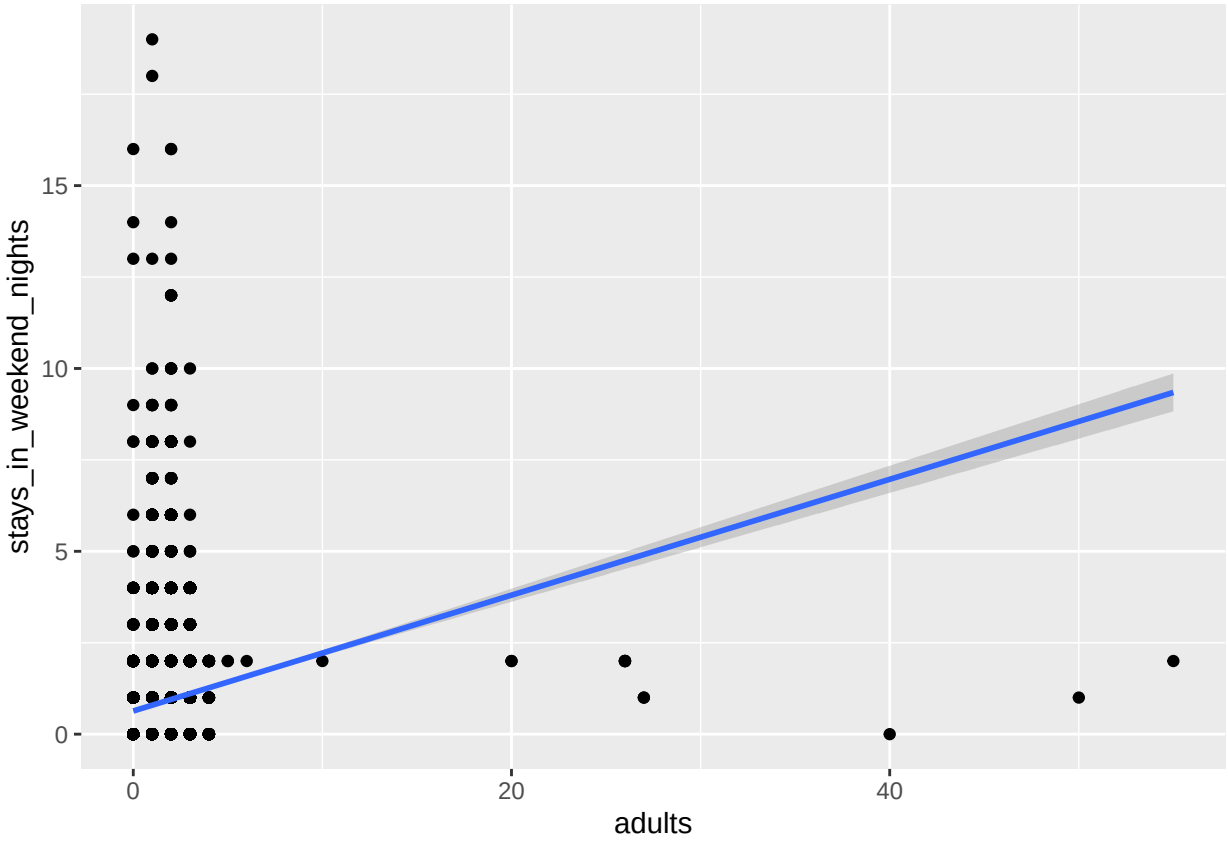
In order to move towards the problem's solution, I have performed certain analysis between different factors.

4.1 Discussion 1

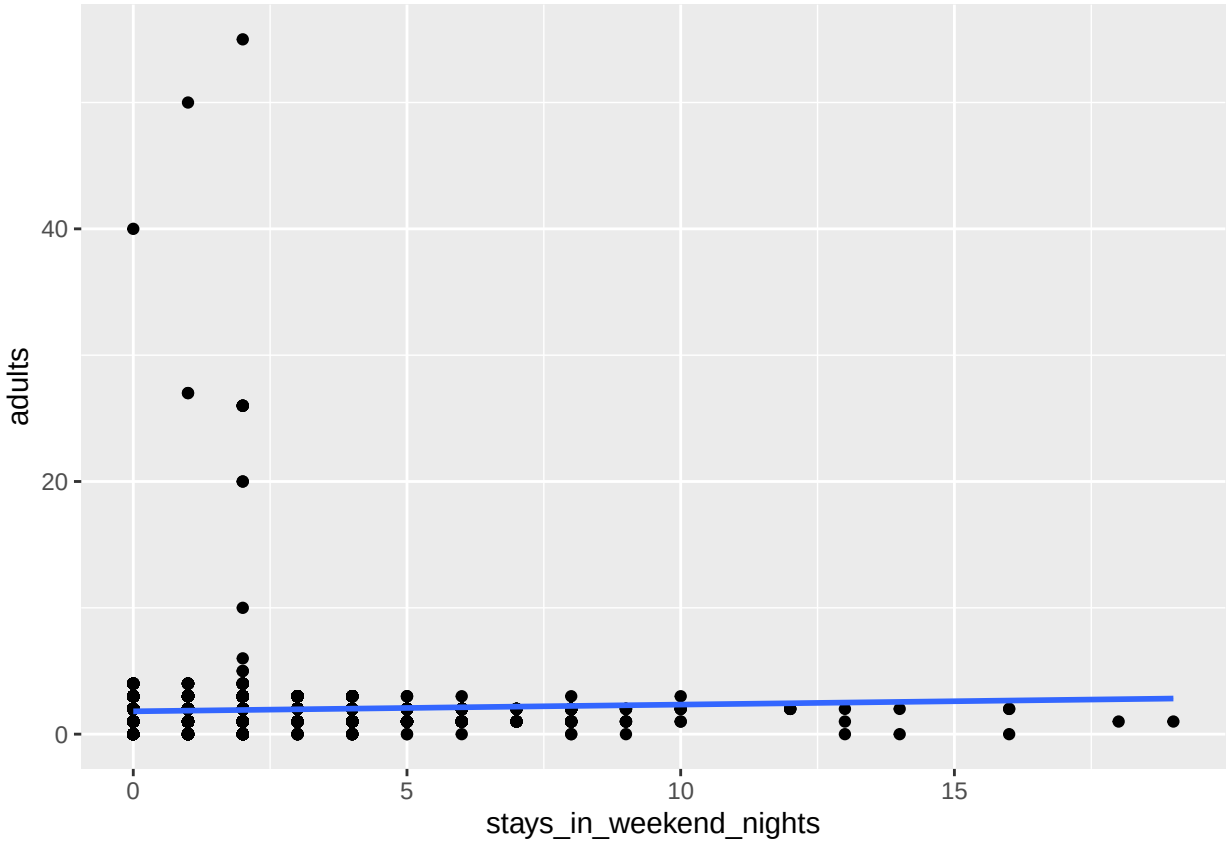
Correlation test was conducted on Number of weekend stays and number of adults in order to find out if there was any correlation present between these two factors.

The output of test was "cor = 0.0919", which illustrated that the correlation between Number of weekend stays and Number of adults is 0.0919. As per Cohen(1992) in (Marshal et al), if the value of cor < 0.1 ,then that correlation is considered to be negligible correlation. Hence, there is a negligible correlation between the Number of weekend stays and Number of adults.

A scatter plot between adults and stays_in_weekend_nights is shown.



A scatter plot between stays_in_weekend_stays and adults is shown.



Thus, from these scatter plots also, it is apparent that there is no linear correlation between the variables stays_in_weekend_nights and adults.

4.2 Discussion 2

4.2.1

One way ANOVA test was performed between is_repeated_guest and arrival_date_month.

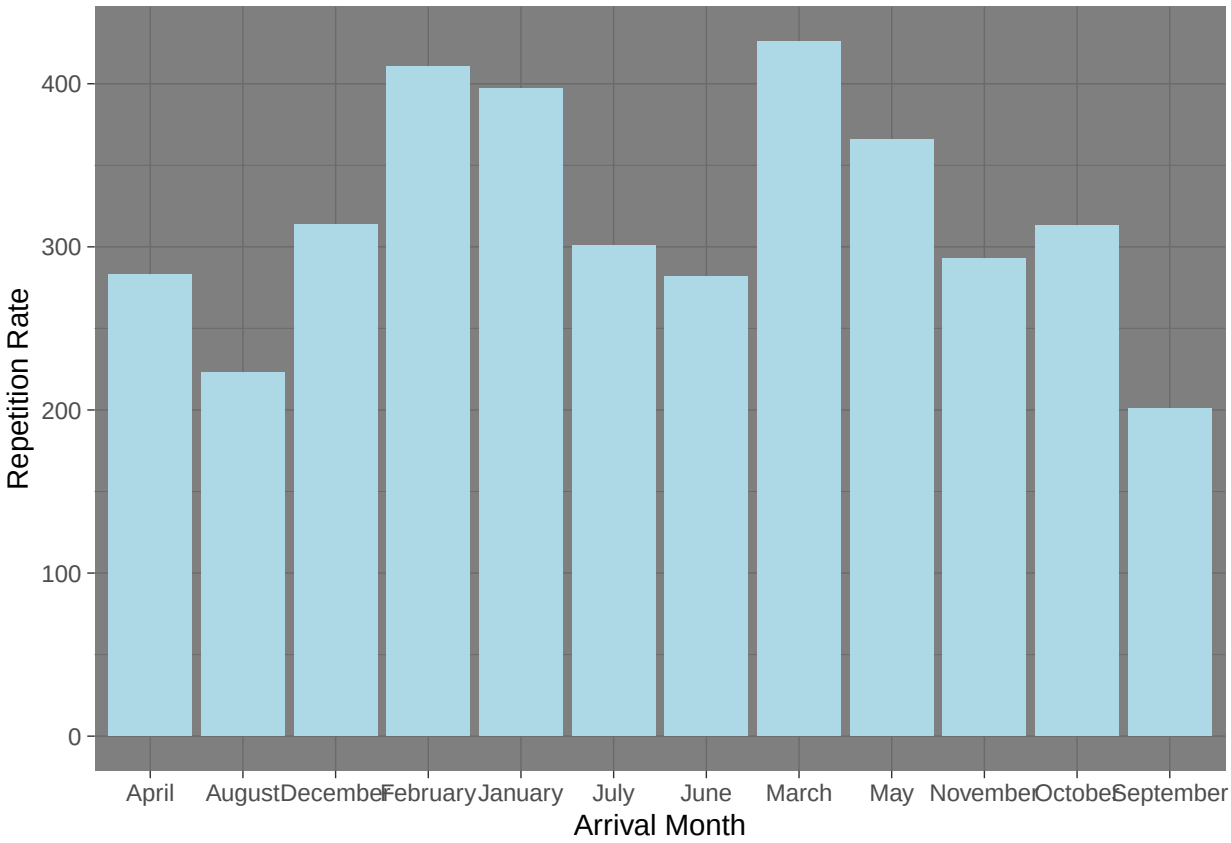
Here, is_repeated is a numerical variable having 2 values. Whereas, arrival_date_month is a categorical variable having 12 values. There was a statistically significant difference between groups as determined by one-way ANOVA. Hence, Post hoc comparisons using the Tukey test were carried out.

There was a significant difference between the Repetition of travelers from following Arrival Months: January-April, February-August, January-August, July-January, June-January and few more.

Thus,these results prove that Arrival Month of a client affects the repetition rate.

4.2.2

To view number of Booking Repetitions in regard to Arrival Month, we will visualize a Bar Chart here. The chart shows proportion of Repetitions in each Month.



From the Bar Chart, it can be stated that the highest number of Repetitions were seen for month of March followed by February and January. Next, month of May had some repeated visitors. Repeated guests were rarely observed in August and September. Rest other months had almost similar proportion of repetitions.

4.3 Discussion 3

4.3.1

One way ANOVA test was performed between is_canceled and market_segment.

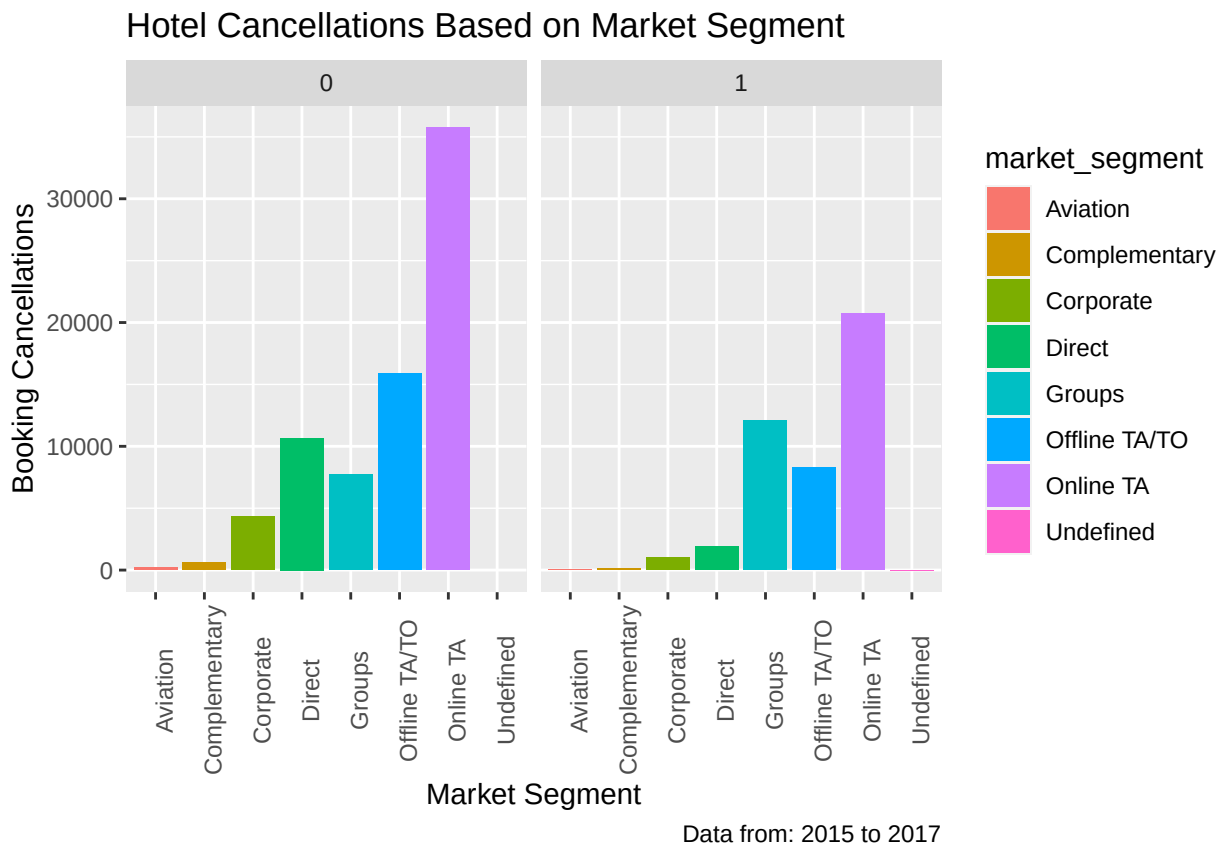
Here, is_repeated is a numerical variable having 2 values. Whereas, market_segment is a categorical variable having 7 values. There was a statistically significant difference between groups as determined by one-way ANOVA. Hence, Post hoc comparisons using the Tukey test were carried out.

There was a significant difference between the cancellation of visitors from following Market Segments: Groups-Aviation (0.3912),Groups-Complementary (0.4800), Groups-Corporate (0.4233), Groups-Direct (0.4572), Offline TA/TO-Complementary (0.2126), Online TA-Complementary (0.2367) and few more.

Thus,these results prove that Market Segment of a client affects the cancellation rate.

4.3.2

A Grouped Bar Chart will be employed for this purpose. Counts of Cancellations by Market Segment is shown. Note:
is_canceled = 1 : Canceled
is_canceled = 0 : Not canceled



The bar graph depicts that most cancellations were from Online Travel Agents as opposed to that of Aviation. Next, Groups accounted for the second highest number of cancellations.

Furthermore, Offline Travel Agents made midway number of cancellations followed by Direct and Corporate Market Segment.

Finally, customers of both Complementary and Aviation were least likely to cancel a booking.

Future Implications

- The goal is of this project is to increase the revenue of the hotels using the insights describing maximum number of Bookings and Cancellations in a particular given condition. We can use several factors to identify maximum number of customers in terms of Arrival Date, Stays in Weekend and Weekdays, Age, Type of Meal, Market Segment, Distribution Channel, Reservation Status.
- Our main objective was to analyze bookings of the two hotels namely, an anonymous city hotel from Lisbon and a resort hotel from Algarve. Data for Arrival Date between 1st of July of 2015 and the 31st of August 2017 were selected for the analysis. In future, dataset for more hotels and/or duration can be taken.
- We have used the following features for gaining our insights:
  Arrival Date, Age, Stays in Weekends and Week days, Repetition, Cancellation, Market Segment, Reservation Status.

- We have not taken these features into account for our analysis in this project, which can be considered in the future:
  Assigned and reserved room type, Meal, Country, Agent, Distribution Channel, Company, Required car parking slots, Total special requests.

Abrate, Graziano, and Giampaolo Viglia. 2016. "Strategic and Tactical Price Decisions in Hotel Revenue Management." *Tourism Management* 55: 123–32.

Antonio, Nuno, Ana De Almeida, and Luis Nunes. 2017. "Predicting Hotel Booking Cancellations to Decrease Uncertainty and Increase Revenue." *Tourism & Management Studies* 13 (2): 25–39.

António, Nuno. 2019. "Predictive Models of Hotel Booking Cancellation: A Semi-Automated Analysis of the Literature." *Tourism & Management Studies* 15 (1): 7–21.

Chalupa, Stepan, and Martin Petricek. 2022. "Understanding Customer's Online Booking Intentions Using Hotel Big Data Analysis." *Journal of Vacation Marketing*, 13567667221122107.

Lee, Misuk. 2018. "Modeling and Forecasting Hotel Room Demand Based on Advance Booking Information." *Tourism Management* 66: 62–71.

Martin-Fuentes, Eva, and Juan Pedro Mellinas. 2018. "Hotels That Most Rely on Booking. Com–Online Travel Agencies (OTAs) and Hotel Distribution Channels." *Tourism Review*.

Novakovic, Jasmina, and Snezana Turina. n.d. "Hotel Reservation Cancellations: Analysis and Prediction Using Machine Learning Algorithms." *ACADEMIC JOURNAL* 4.

Saputro, Pujo Hari, and Herlino Nanang. 2021. "Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets." *Journal of Applied Data Sciences* 2 (1): 40–56.

Tsai, Yao-Hsu, Chien-Cheng Lin, and Min-Hsien Lee. 2022. "Analysis of Application Data Mining to Capture Consumer Review Data on Booking Websites." *Mobile Information Systems* 2022.

Vives, Aldric, Marta Jacob, and Marga Payeras. 2018. "Revenue Management and Price Optimization Techniques in the Hotel Sector: A Critical Literature Review." *Tourism Economics* 24 (6): 720–52.

Zheng, Chuyi, and Gabor Forgacs. 2017. "The Emerging Trend of Hotel Total Revenue Management." *Journal of Revenue and Pricing Management* 16 (3): 238–45.