IST 687 Final Report: Hotel Industry Analysis

By

Shengjie Yao, Jianghui Li, Qian Zhang, and Dong Hyeok Lee

Syracuse University

May 2021

Introduction

Project Overview

This comparative analysis project was commissioned by the European Hotels Group. The overall objective of this report is to assess the performance of the two European Hotel Group hotels located in the city and the resort during 2015-2017.

The performance of each property underwent a series of structured assessments. Overall performance of the two hotels was compared and reported, performance-related factors were identified, future revenue predictions were made as well as recommendations on both hotels were suggested based on current data.

Dataset and Methodology

Dataset

The datasets used for this project were provided by the European Hotels Group. One dataset is from the resort hotel and the other one is from the city hotel. Datasets contain bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. The overall booking recorded in the resort dataset was 40,060(40,059 after data cleaning) and the overall booking recorded in the city dataset was 79,330(40,059 after data cleaning).

Each dataset reported the following data:

Variable name	Type	Description
ADR	Numeric	Average Daily Rate
Adults	Integer	Number of Adults
Agent	Categorical	ID of the travel agency that made the booking

Arrival Date	Date	Date customer was scheduled to arrive
Assigned Room Type	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
Babies	Integer	Number of babies
Children	Integer	Number of 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
Country	Categorical	Country of origin. Categories are represented in the ISO 3155-3:2013 format
Customer Type	Categorical	Type of booking, assuming one of the 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
Days in Wating List	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
Deposit Type	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit - no deposit was made. 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.
Distribution Channel	Categorical	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Is Cancelled	Categorical	Value indicating if the booking was canceled or not
Is repeated Guest	Categorical	Value indicating if the booking name was from a repeated guest or not
Lead Time	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
Market Segment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: 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)
Previous Booking Not Canceled	Integer	Number of previous bookings not canceled by the customer prior to the current booking
Previous Cancellations	Integer	Number of previous bookings that were canceled by the customer prior to the current booking
Required Card Parking Spaces	Integer	Number of car parking spaces required by the customer
Reservation Status	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
Reservation status Date	Date	Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel
Reserved Room Type	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
Stay In Weekend Nights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
Stay In Week Nights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
Total of Special Requests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)

Table 1: Data description

Additional data attributes were created for data analysis purposes:

Variable name	**Type**	**Description**
Arrival month	Categorical	Month information based on arrival date

Table 2: Additional data attributes

Analysis results were computed based on the above dataset and attributes. Please refer to the descriptions of the variables for detailed information.

Data cleaning rules

Upon initial data screening, significant outliers were screened out. Additionally, NA dates were omitted in the City dataset. The final city hotel dataset contained 40,059 sets of data.

Methodology

The project consisted of the following analysis sections: Performance analysis, Performance-related factors identification, Future performance prediction, and Recommendations.

Performance analysis was performed based on descriptive analysis. The descriptive analysis looked at both hotel-related variables(e.g. ADR) and customer-related variables (e.g. Cancellation). A customer segment map was also included. The performance of the two hotels was directly compared.

Performance-related factors identification was performed using Association Rules Mining. Factors that connect visitor types to the cancellation were identified.

Future performance prediction was performed using several linear models to predict ADR and Cancelation rates as well as several SVM models to predict who is likely to cancel for both hotels. Factors that predict future performances (i.e. ADR and Cancelation rates) were identified.

Last but not least, recommendations were formed based on all the analyses discussed in the project.

Performance Analysis

Overall description of the data

Resort

Variable	Range	Median	Mean
Is Canceled	0-1	0	0.2776
Lead Time	0-737	57	92.68
Arrival Date*	2015/07/01-2017/08/31	NA	NA
Reservation Status Date	2014/11/18-2017/09/14	NA	NA
Reservation Status	NA	NA	NA
Stays In Weekend Nights	0-19	1	1.19
Stays In Week Nights	0-50	3	3.129
Adults	0-55	2	1.867
Children	0-10	0	0.1287
Babies	0-2	0	0.0139
Meal	NA	NA	NA
Country	NA	NA	NA
Market Segment	NA	NA	NA
Distribution Channel	NA	NA	NA
Is Repeated Guest	NA	NA	NA
Previous Cancellations	0-26	0	0.1017
Previous Bookings Not Canceled	0-30	0	0.1465
Reserved Room Type	NA	NA	NA
Assigned Room Type	NA	NA	NA
Booking Changes	0-17	0	0.288
Deposit Type	NA	NA	NA
Agent	NA	NA	NA

Company	NA	NA	NA
Days in Waiting List	0-185	0	0.5278
Customer Type	NA	NA	NA
ADR	-6.38-508	75	94.95
Required Parking Spaces	0-8	0	0.1381
Total Special Requests	0-5	0	0.6197

City

Variable	**Range**	**Median**	**Mean**
Is Canceled	0-1	0	0.7884
Lead Time	0-629	74	135.2
Arrival Date*	2015/07/01-2017/08/31	NA	NA
Reservation Status Date	2014/10/17-2017/08/29	NA	NA
Reservation Status	NA	NA	NA
Stays In Weekend Nights	0-14	1	0.7886
Stays In Week Nights	0-34	2	2.235
Adults	0-4	2	1.856
Children	0-3	0	0.09
Babies	0-10	0	0.00312
Meal	NA	NA	NA
Country	NA	NA	NA
Market Segment	NA	NA	NA
Distribution Channel	NA	NA	NA
Is Repeated Guest	NA	NA	NA
Previous Cancellations	0-21	0	0.1147
Previous Bookings Not Canceled	0-72	0	0.1555
Reserved Room Type	NA	NA	NA
Assigned Room Type	NA	NA	NA

Booking Changes	0-20	0	0.1216
Deposit Type	NA	NA	NA
Agent	NA	NA	NA
Company	NA	NA	NA
Days in Waiting List	0-391	0	5.217
Customer Type	NA	NA	NA
ADR	0-352.5	96	101.7
Required Parking Spaces	0-1	0	0.00699
Total Special Requests	0-5	0	0.3236

*missing data = 39270

Detailed descriptive analysis

Reserved Room Type

Different room types were reserved in the two hotels.

Resort

In the resort hotel, A,D,E,G,F,C, H,L,B,P, and L types of rooms were reserved by customers. The most reserved room is A type of 23,399 and the least reserved room is P type of 2.

A	D	E	G	F	C	H	L	B	P
23399	7432	4982	1610	1106	918	601	6	3	2

City

In the city hotel, A,D,F,B,E,G,P and C types of rooms were reserved by customers. The most reserved room is A type of 33,158 and the least reserved room is C type of 6.

A	D	F	B	E	G	P	C
33158	4765	848	589	544	139	10	6

Assigned Room Type

Different room types are assigned in the two hotels.

Resort

In the resort hotel, A,D,E,C,G,F,H,I,B,P, and L types of rooms were assigned to customers. The most assigned room is A type of 17,046 and the least assigned room is L type of 1.

A	D	E	C	G	F	H	I	B	P	L
17046	10339	5637	2214	1853	1733	712	363	159	2	1

City

In the City hotel, A,D,B,F,E,G,K,C and P types of rooms were assigned to customers. The most assigned room is A type of 31,132 and the least assigned room is P type of 10.

A	D	B	F	E	G	K	C	P
31132	6007	990	919	756	188	36	21	10

The reserved room types and the actually assigned rooms are different for both hotels.

Customer Type

There are four types of customers that have stayed in both hotels.

Resort

In the resort hotel, the most stayed customer types are transient of 30,208 and the least stayed customer types are group customers of 284.

Transient	Transient-Party	Contract	Group
30208	7791	1776	284

City

In the city hotel, the most stayed customer types are transient of 29,232 and the least stayed customer types are group customers of 120.

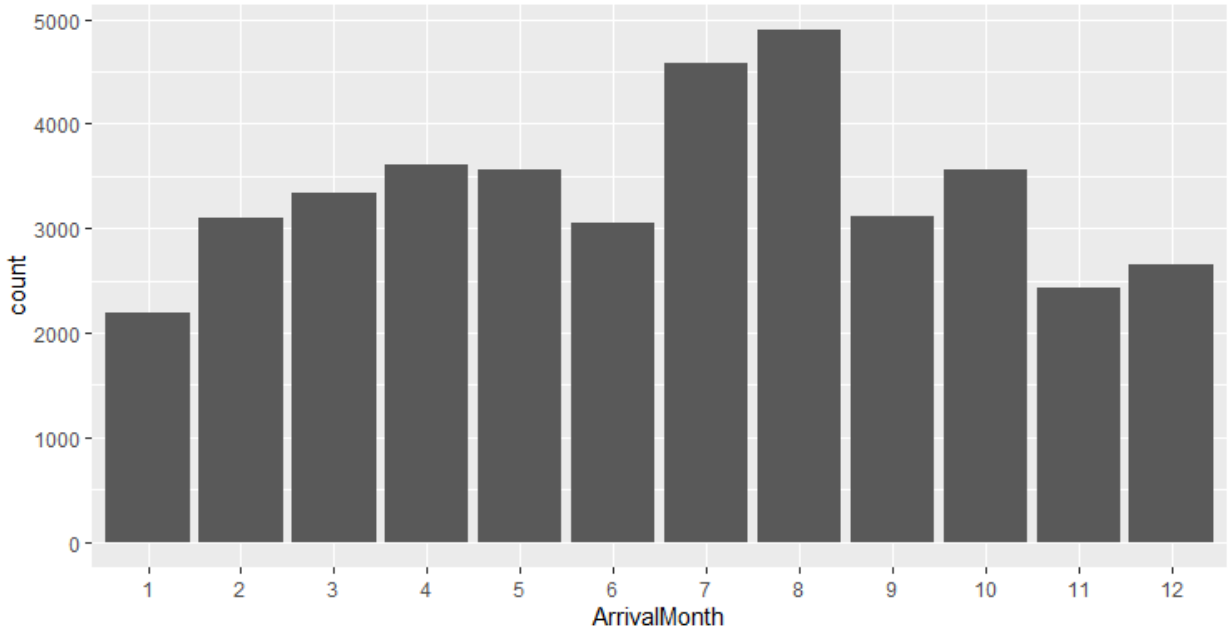
```
       Transient Transient-Party        Contract           Group
           29232            8766            1941             120
```

Arrival Month

The arrival month of the customers was calculated and shown in bar charts.

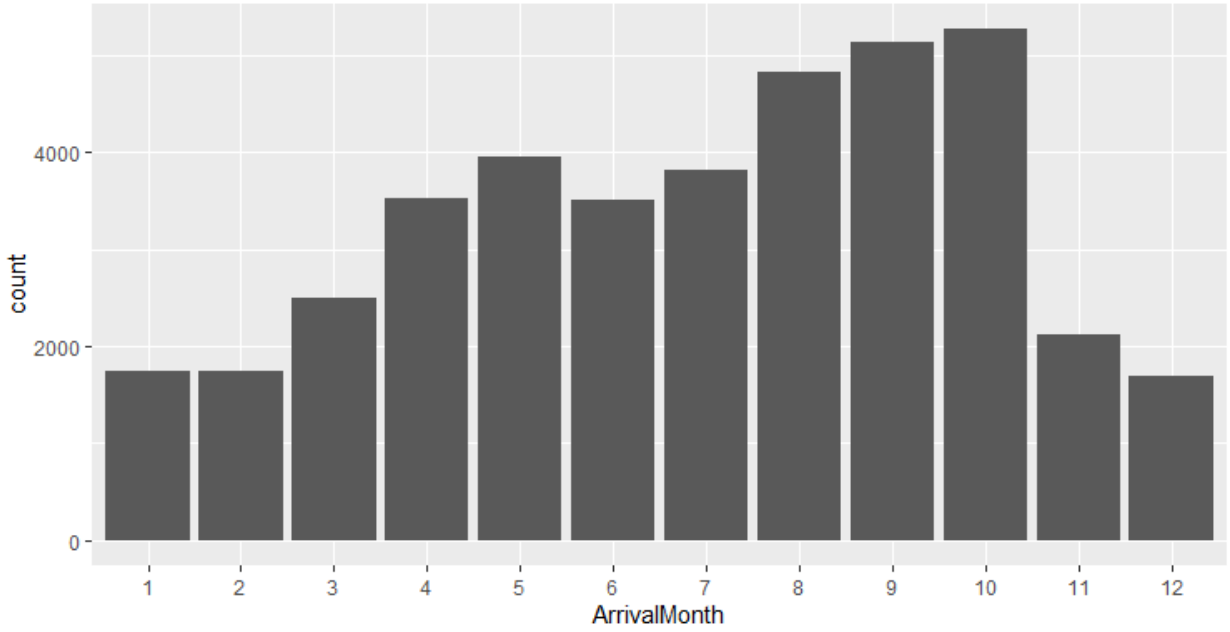
Resort

The arrival month of the customers for the resort hotel was negatively skewed. The most popular months are July and August while the least popular month is January. One possible explanation could be July and August are students' summer vacations.



City

The arrival month of the customers for the city hotel was negatively skewed. The most popular months are September and October while the least popular month is December. One possible explanation could be there may be more business meetings between September and October.



Deposit Type

For customer payment, three payment types are included.

Resort

In the resort hotel, no deposit was paid 38,198 times, non-refund 1,719 times, and refundable 142 times.

```
No Deposit Non Refund Refundable
     38198       1719        142
```

City

In the city hotel, no deposit was paid 28,275 times, non-refund 11,767 times, and refundable 17 times.

```
No Deposit Non Refund Refundable
     28275      11767         17
```

Distribution Channel

There were four distribution channels for the resort hotel while there were five for the city hotel.

Resort

The most used distribution channel for the resort hotel is TA/TO with 28,924 with the least being Corporate with 3,269. There is one undefined distribution channel.

TA/TO	Direct	Corporate	Undefined
28924	7865	3269	1

City

The most used distribution channel for the city hotel is also TA/TO with 36,726 with the least being GDS with 38. There are four undefined distribution channels.

TA/TO	Direct	Corporate	GDS	Undefined
36726	2045	1246	38	4

Cancellation

The total cancellation of each hotel was calculated. 1 being canceled and 0 is not canceled. A **cancellation rate** was calculated by dividing the total number of canceled stays by the total number of booked stays.

Resort

The total cancellation of the resort hotel is 11,122. The cancelation rate is 11,122/(28,937+11,122)=**0.2776**

	0	1
28937	11122	

City

The total cancellation of the city hotel is 33,102. The cancelation rate is 31,582/(31,582+8477)=**0.7884**

0	1
8477	31582

The cancellation rate of the city hotel is much higher than the cancellation rate of the resort hotel.

Repeated Guests

The total number of repeated guests was calculated. 1 being repeated guests and 0 being non-repeated guests. A **repeat guest rate** was calculated by dividing the total number of repeated guests by the total number of guests.

Resort

The total number of repeated guests in the resort hotel is 1,778. The repeat guest rate was 1,778/(38,281+1,778)=**0.0444**

0	1
38281	1778

City

The total number of repeated guests in the city hotel is 2,032. The repeat guest rate was 845/(39214+845)=**0.0210**

0	1
39214	845

The repeat guest rate of the resort hotel is higher than the repeat guest rate of the city hotel.

Market Segment

The market segments were calculated. There were six market segments for the resort hotel with seven market segments (plus undefined).

Resort

The most used market segment for the resort hotel is online TA with 17,728 while the least is complementary with 201.

Online TA	Offline TA/TO	Direct	Groups	Corporate	Complementary
17728	7472	6513	5836	2309	201

City

The most used market segment for the city hotel is also online TA with 17,417 while the least excluding undefined is aviation with 41.

Online TA	Groups	Offline TA/TO	Direct	Corporate	Complementary
17417	11022	8539	1736	1090	212
Aviation	Undefined				
41	2				

Customer Meal Choices

Four meal choices are given to customers within both hotels.

Resort

The most chosen meal option in the resort hotel is BB with 30,004 with the least being SC with 86. There are 1,169 undefined meal choices.

BB	HB	Undefined	FB	SC
30004	8046	1169	754	86

City

The most chosen meal option in the city hotel is also BB with 32,370 with the least being FB with 36.

BB	SC	HB	FB
32370	4435	3218	36

Reservation Status

Reservation status was counted for both hotels including checked out, canceled, or no-show.

Resort

28,937 customers checked out, 10,831 canceled and 291 did not show in the resort hotel.

Check-Out	Canceled	No-Show
28937	10831	291

City

8,477 customers checked out, 30,732 canceled and 850 did not show in the city hotel.

Canceled	Check-Out	No-Show
30732	8477	850

There are way more canceled customers in the city hotel than the resort hotel.

ADR analysis

ADR, or the average daily rate is often considered one important factor in reflecting a hotels performance in the hospitality industry to measure the strength of revenues generated. It is calculated by summing all of the revenues generated by all the occupied rooms and dividing that sum by the total number of occupied rooms over a given time period. It is a simple average that shows the revenues generated per occupied room.

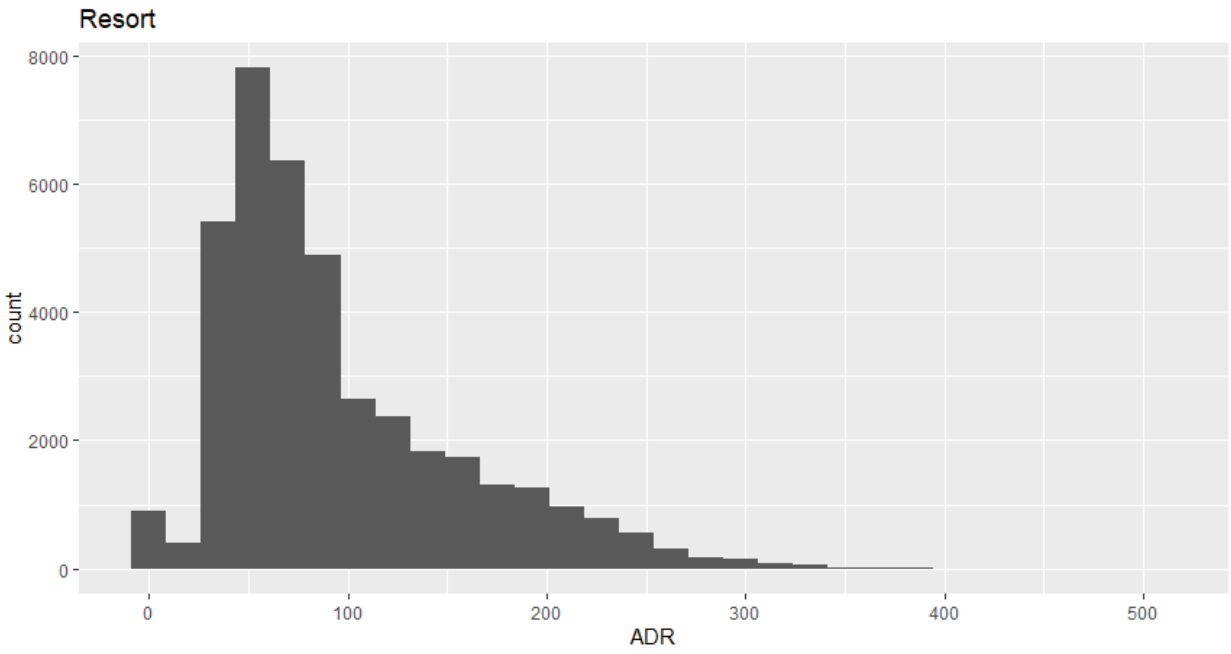
In this part, ADR of the both hotels was used along with other indicating factors to indicate the performance of each hotel.

ADR

The ADRs for both hotels were calculated and the data were shown in histograms.

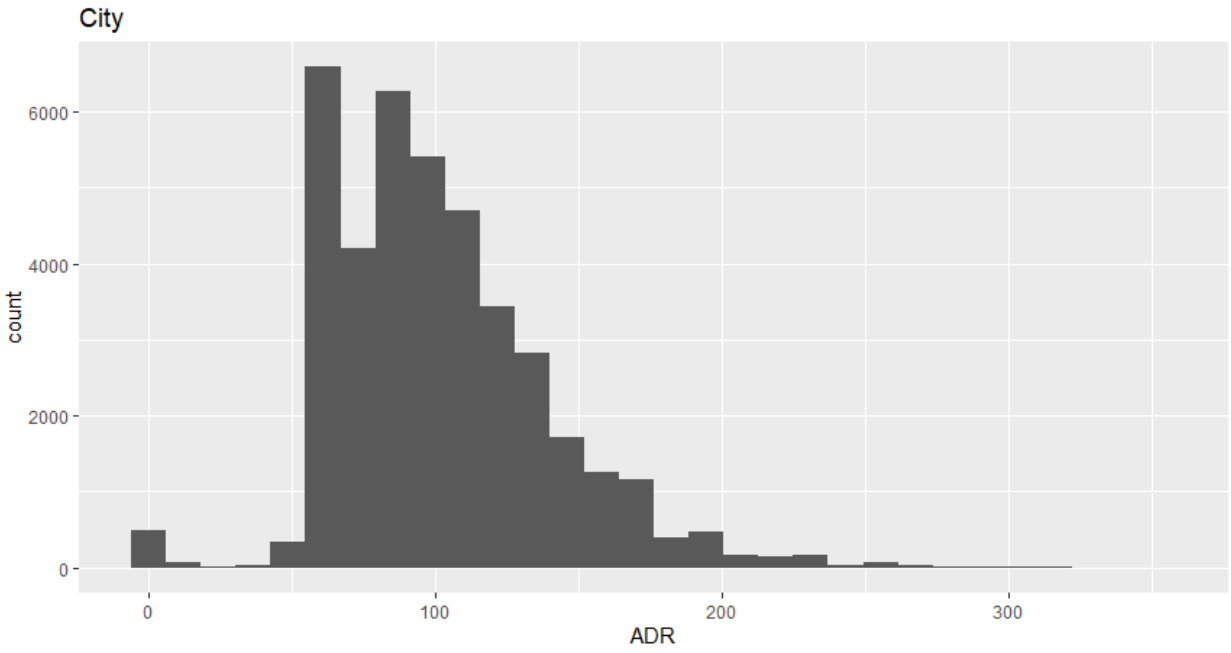
Resort

The most frequent average daily rate for the resort hotel lies between 0-200 with several negative revenues. The overall distribution of the data is positively skewed.



City

The most frequent average daily rate for the city hotel lies between 0-200 with 0 negative revenues. The overall distribution of the data is positively skewed.

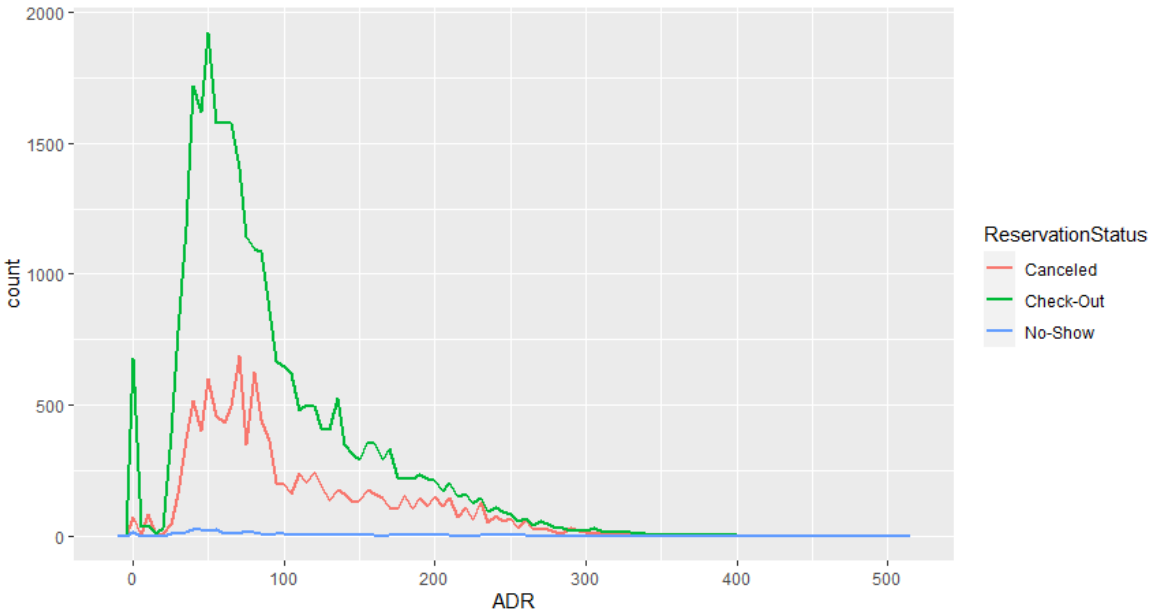


There are more higher ADRs for the resort hotel than the city hotel.

ADR - Cancellation

Cancellation could be impacting ADR. When putting ADR with Cancellation, clear patterns for each hotel are shown below.

Resort



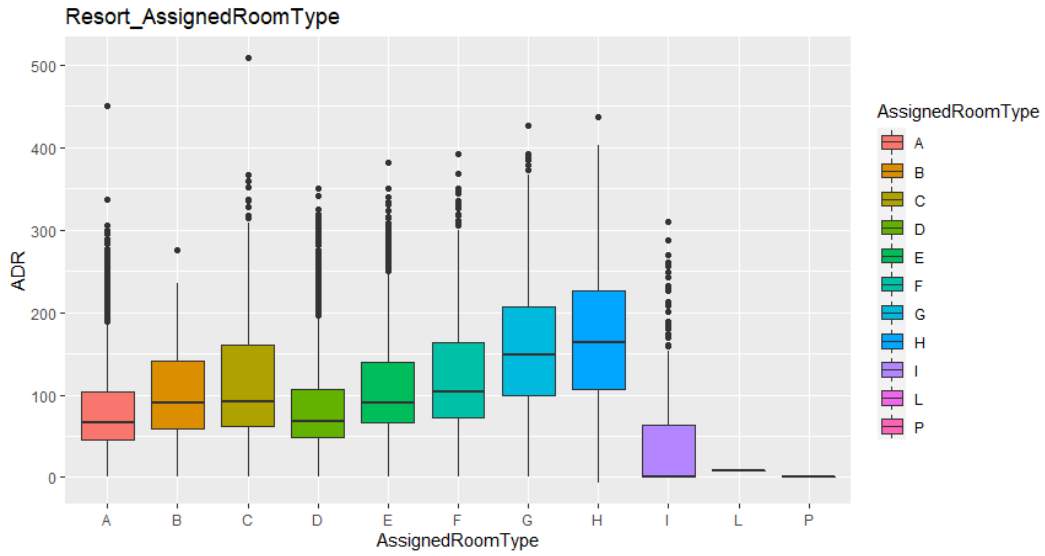
City

There are way more canceled cases for the city hotel. Therefore, the overall ADR took big hits because of these cancellations.

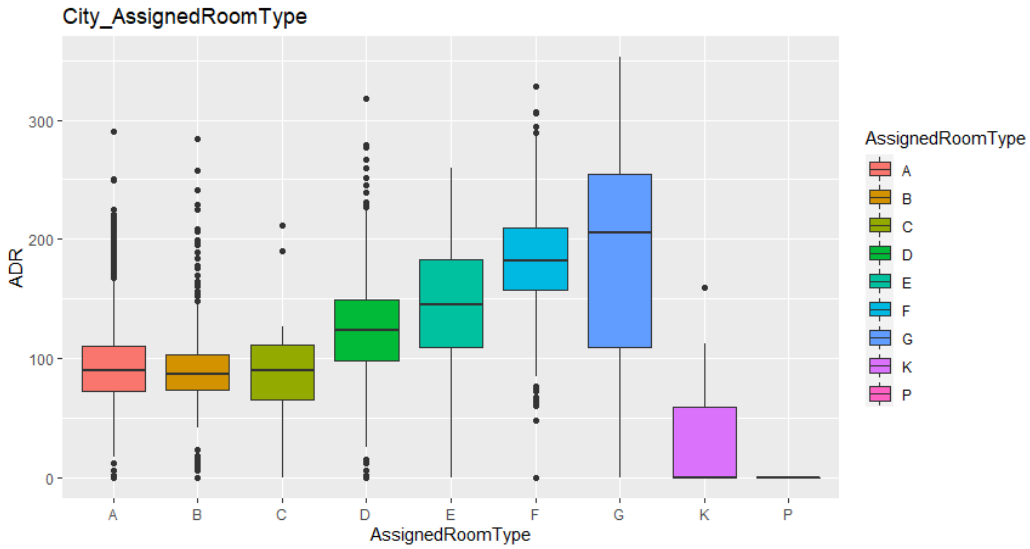
ADR - Assigned Room Type

Different room types offer different profitabilities. Two hotels have different characteristics.

Resort



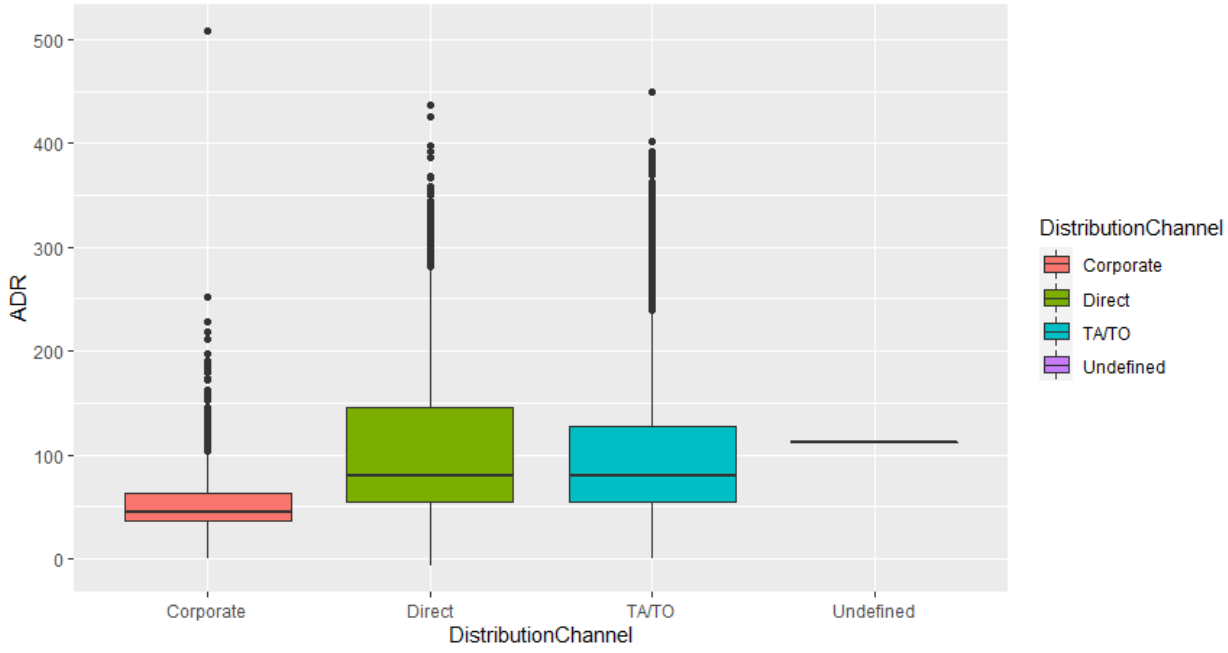
City



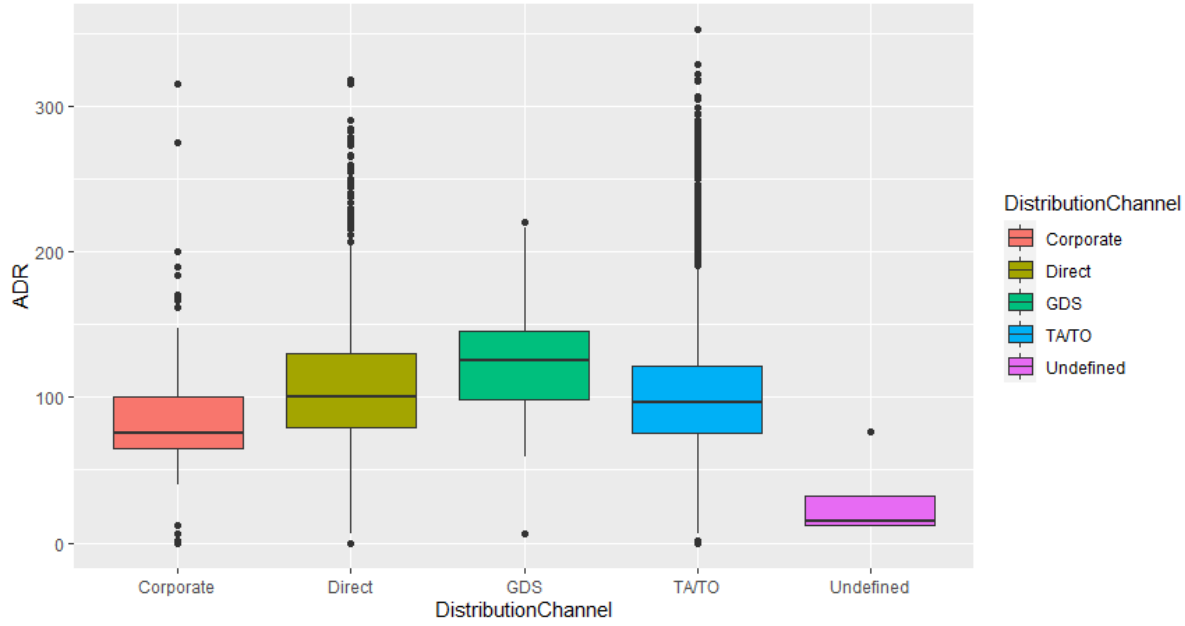
ADR - Distribution Channel

The city hotel has more distribution channels than the resort hotel. ADRs seem to have been affected by distribution channels as well.

Resort



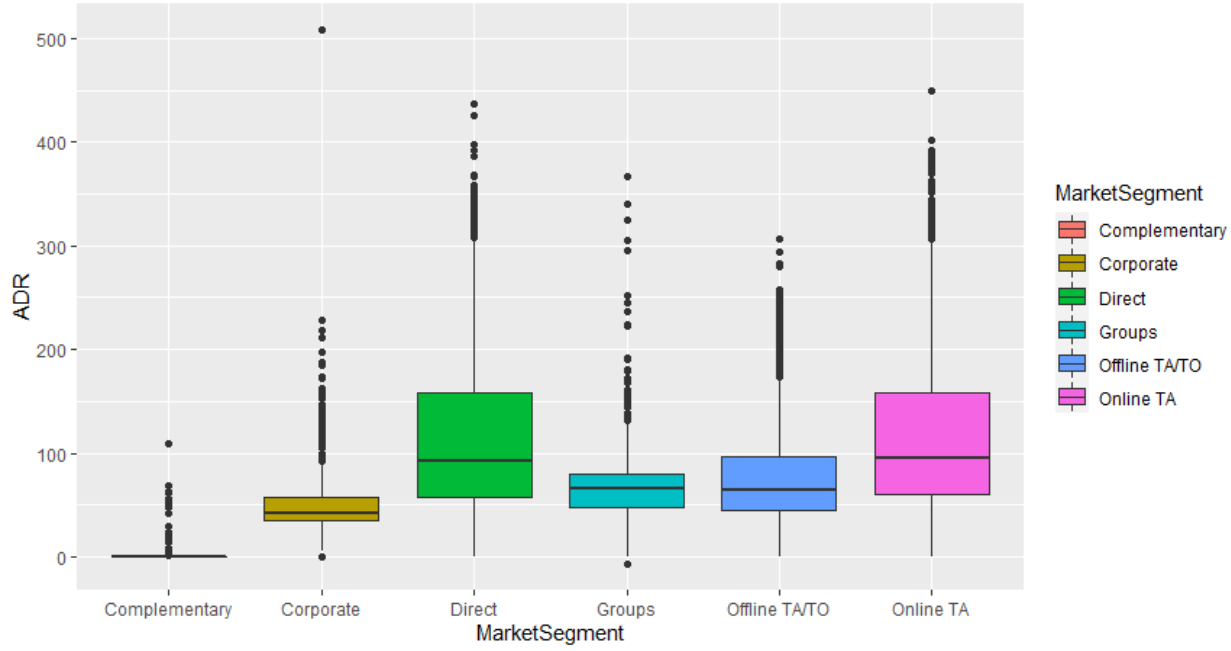
City



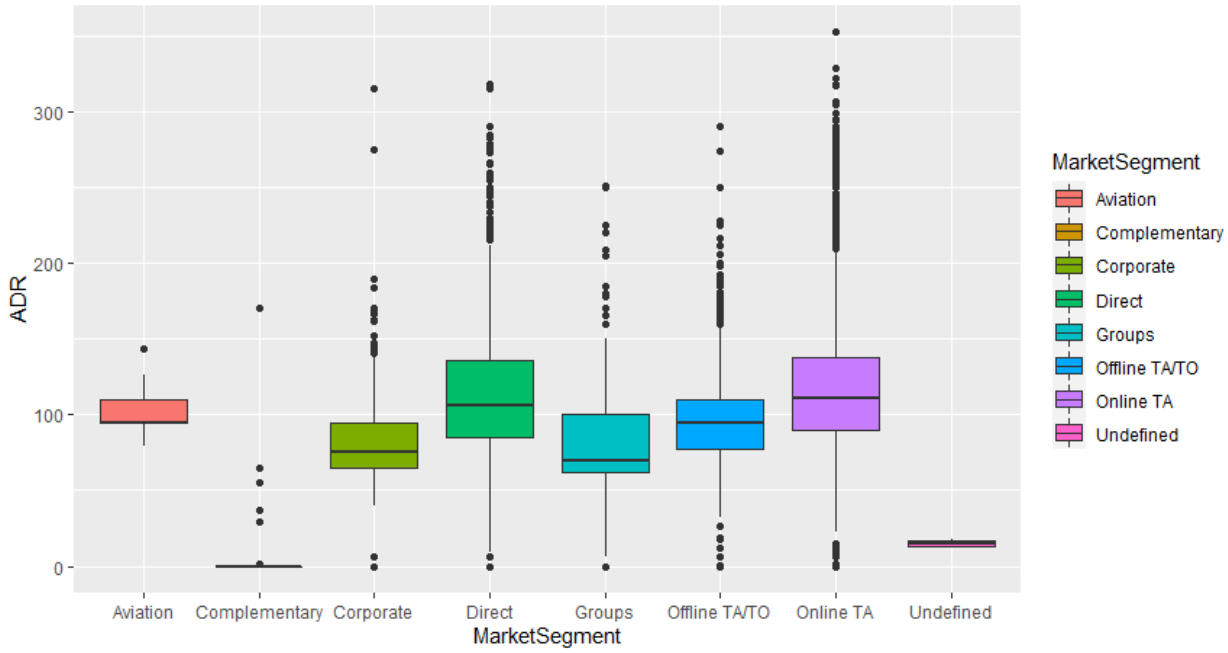
ADR - Market Segment

The city hotel get their revenue from more market segments than the resort hotel.

Resort



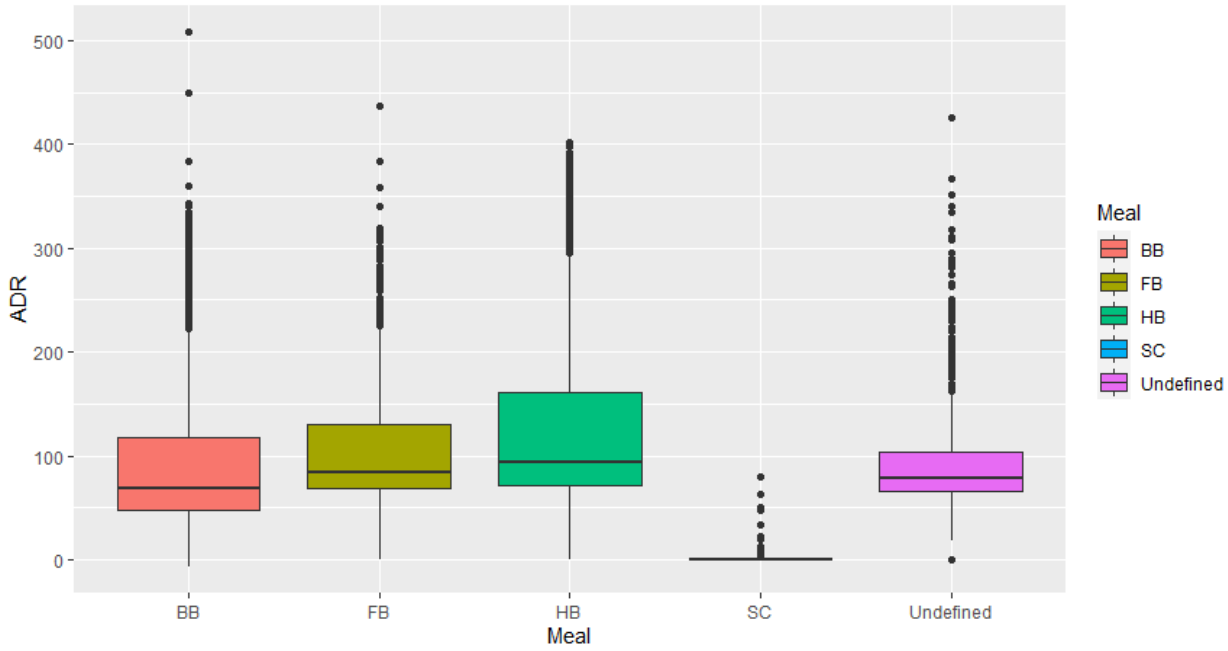
City



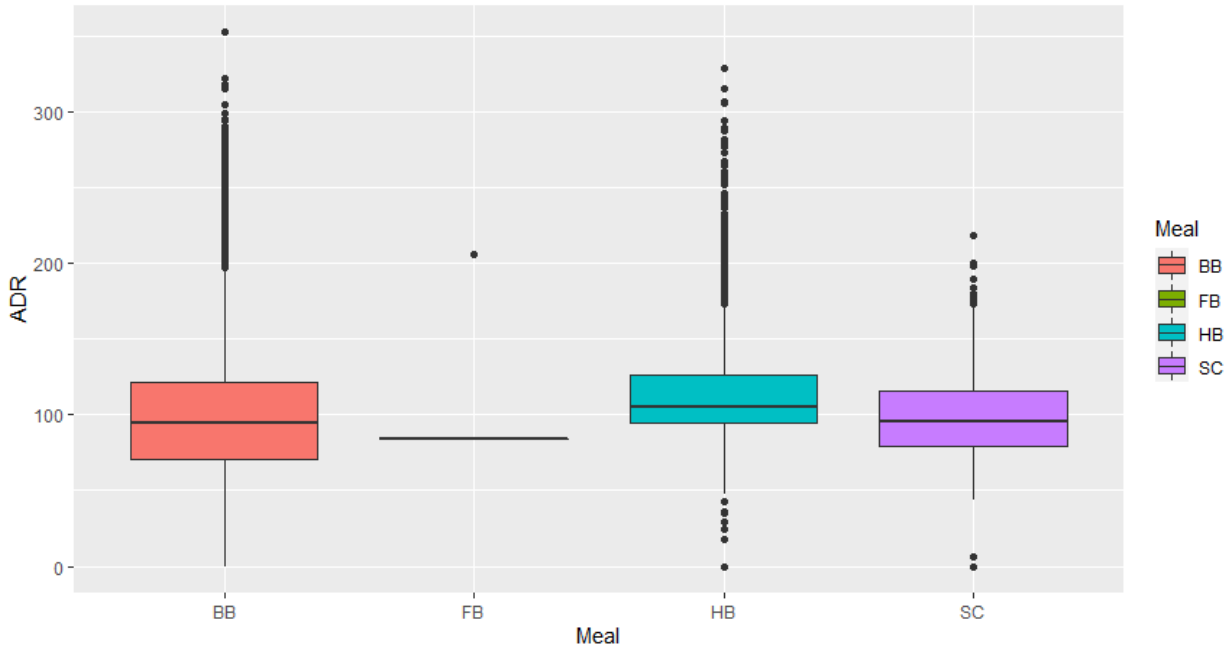
ADR - Customer Meal Choices

The customers have more focused meal choices than the resort hotels.

<u>Resort</u>

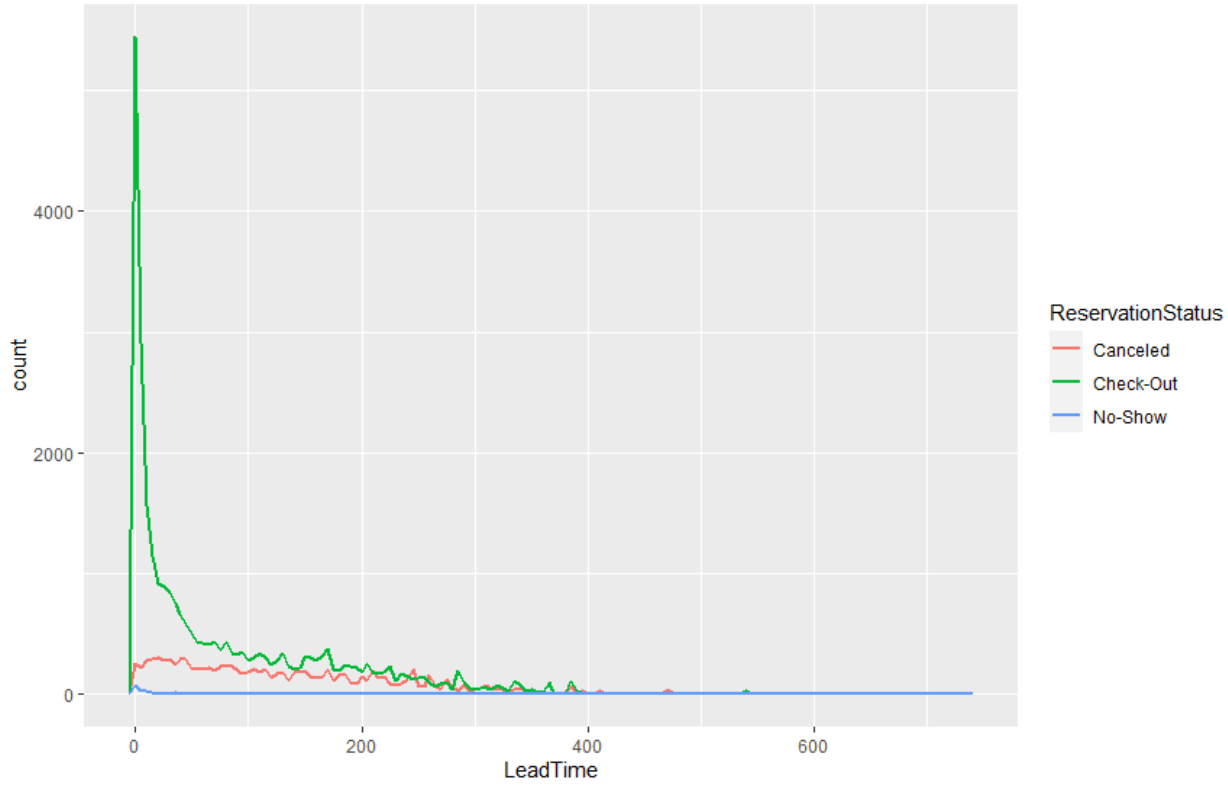


<u>City</u>

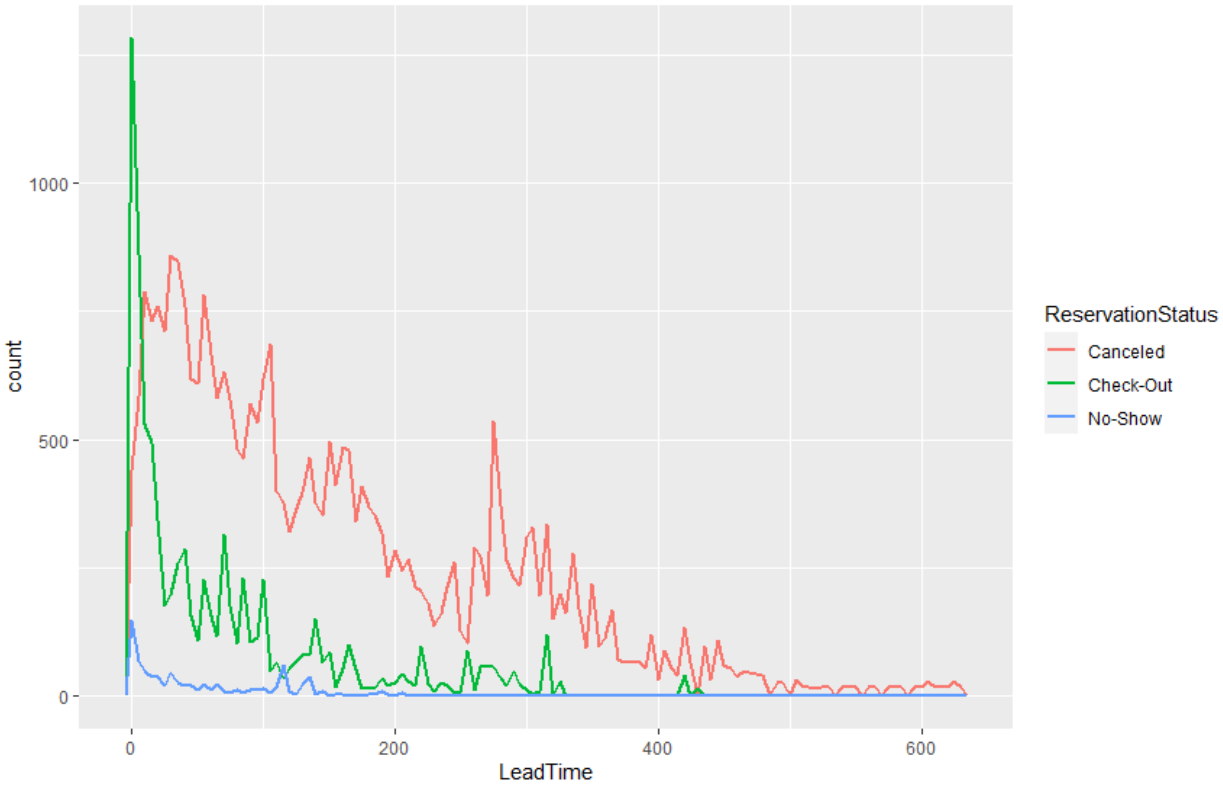


Lead Time- Cancellation

Resort



City

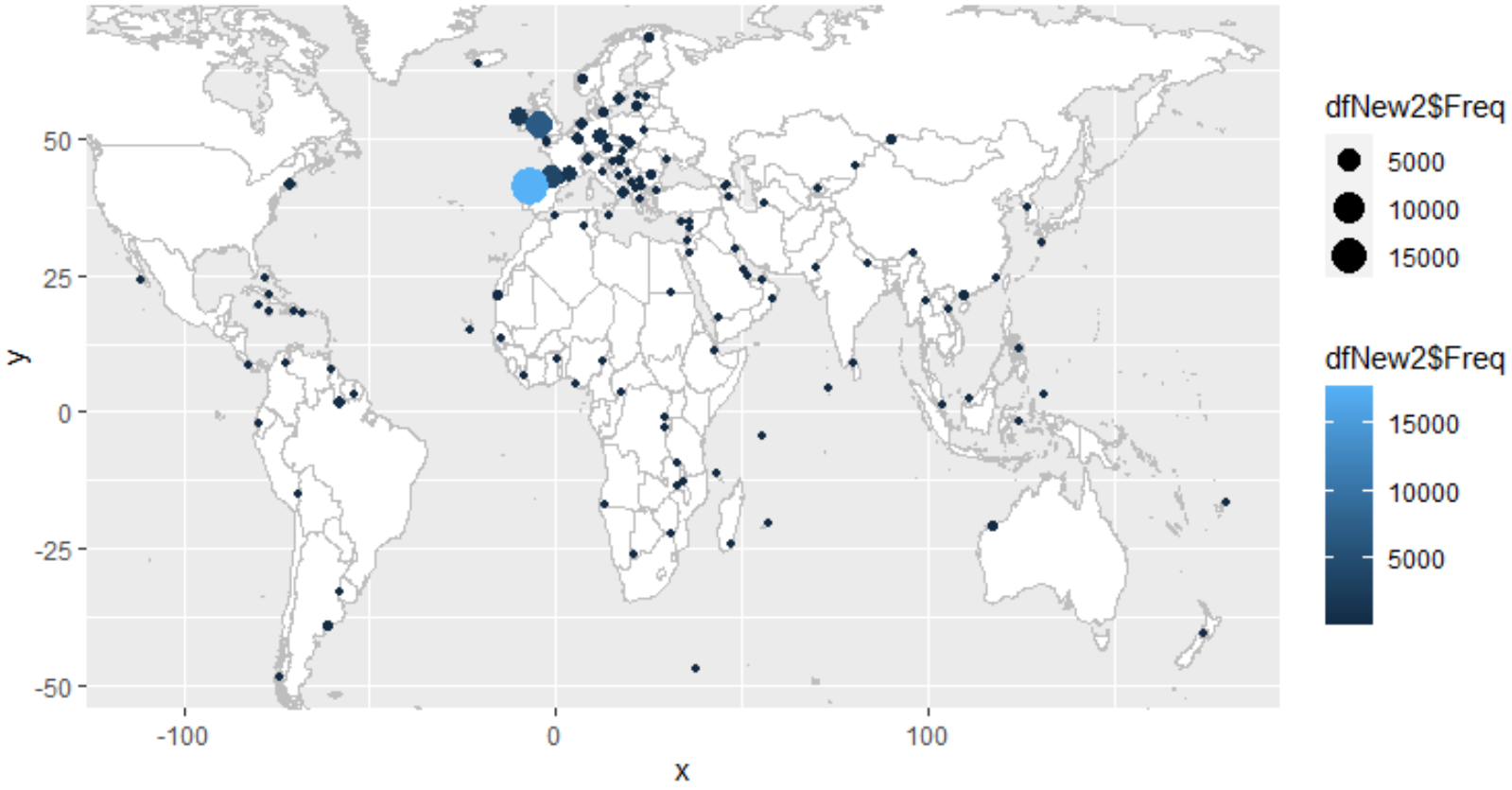


Customer Map

To analyze how customers differ for both hotels, a customer map was created to identify the customer identifications.

Resort

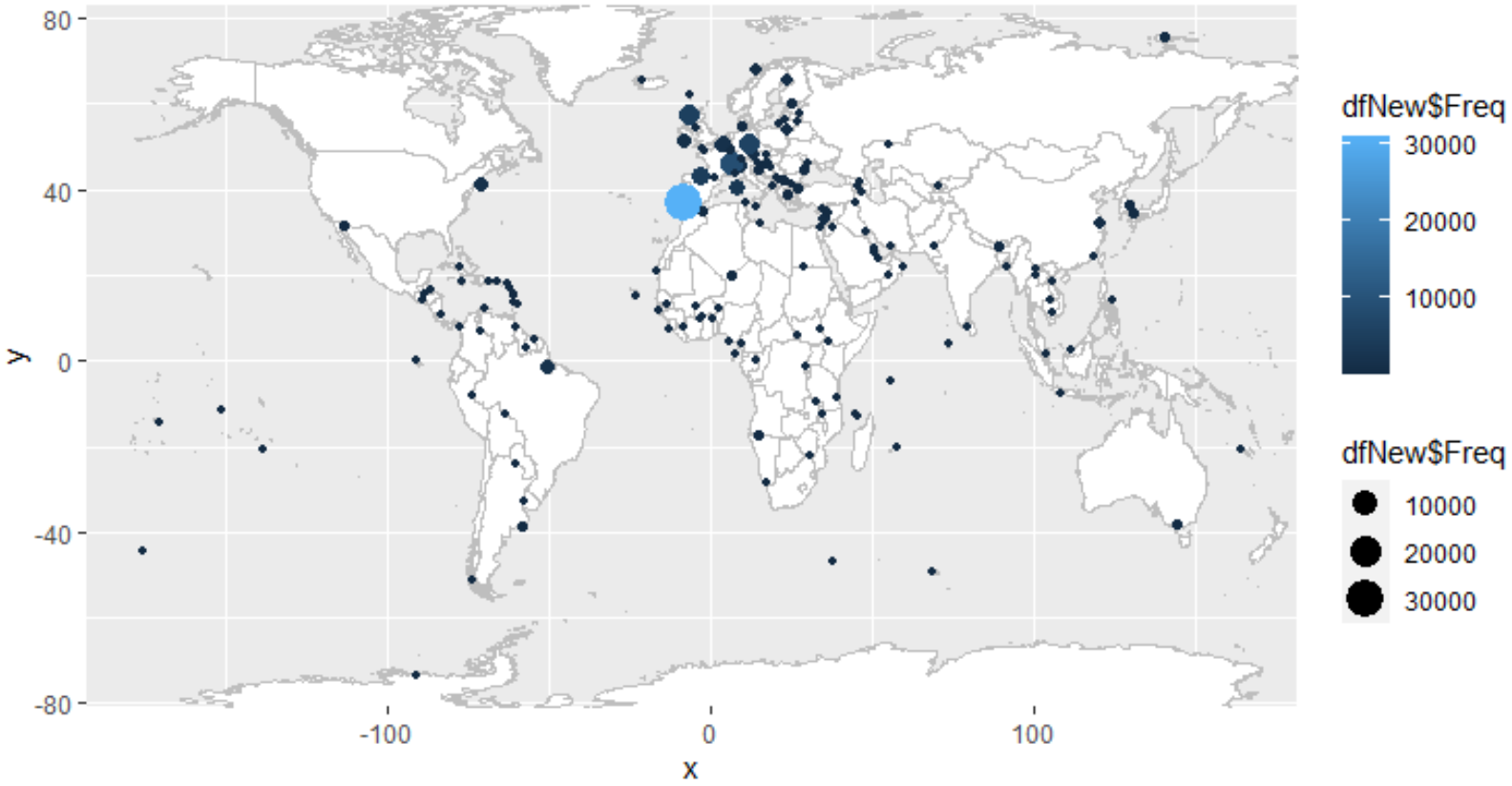
The top six countries' customers that stayed at the resort hotel come from Portugal(17,630), United Kingdom (6,814), Spain(3,957), Ireland(2,166), France(1,511) and Germany (1,203).



Portugal	United Kingdom	Spain	Ireland	France	Germany
17630	6814	3957	2166	1611	1203

City

The top six countries' customers that stayed at the city hotel come from Portugal(30,959), France (8,804), Germany(6,084), United Kingdom (5,315), Spain (4,611) and Italy (3,307).



Portugal	France	Germany	United Kingdom	Spain	Italy
30959	8804	6084	5315	4611	3307

Performance-related Factors Identification

Association Rules Mining

To further understand how performance of each hotel is affected by various factors, association rules mining was used to test potential factors that could lead to customer's cancellation. As stated in previous descriptive statistics, the cancellation rate varied significantly between the two hotels. As such, cancellation is treated as the dependent variable in the mining process.

Association rule mining aims to extract interesting correlations, associations or casual structures among items. An association rule is an implication in the form of X to Y. The rule means X implies Y.

Basic measures for association rules are support and confidence. Support is defined as the percentage of records that contain X to Y relationship to the total number of records in the database. Confidence is defined as the percentage of the number of transactions that contain $X \cup Y$ to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule $X \Rightarrow Y$ can be generated. Support can help filter out the items that have a low frequency. Confidence tells us how often the items x and y occur together, given the number times x occurs.

In this project, lift is also used to indicate potential relationships. Lift indicates the relative magnitude of the probability of observing X to Y. When lift = 1, X and Y are independent of each other. When the lift is greater than one, they are not independent. The higher the value, the greater the relevance of the relationship. It explains the strength of a rule and more the Lift more is the strength.

Two sets of association rule mining were conducted in the two datasets.

City

Cancellation was input as the dependent variable. LeadTime, previous cancellations, repeated guests status, booking changes, required car parking spaces and total of special requests were input as independent variables. Initial support was set to 0.002 with a confidence level of 0.6.

Inspection of the association rules revealed the following results:

```
> inspect(ruleset.City)
      lhs                                 rhs                support confidence    coverage     lift count
[1]   {LeadTime=178}                   => {IsCanceled=1} 0.002218609  0.7364017 0.003012770 1.764841   176
[2]   {LeadTime=277}                   => {IsCanceled=1} 0.002470723  0.8099174 0.003050587 1.941027   196
[3]   {LeadTime=113}                   => {IsCanceled=1} 0.002042128  0.6585366 0.003101010 1.578232   162
[4]   {LeadTime=105}                   => {IsCanceled=1} 0.002584175  0.6655844 0.003882565 1.595122   205
[5]   {LeadTime=104}                   => {IsCanceled=1} 0.002785866  0.6538462 0.004260737 1.566991   221
[6]   {PreviousCancellations=1}        => {IsCanceled=1} 0.062574847  0.9629486 0.064982541 2.307778  4964
[7]   {LeadTime=178,
       TotalOfSpecialRequests=0}       => {IsCanceled=1} 0.002079946  0.9217877 0.002256426 2.209133   165
[8]   {LeadTime=178,
       BookingChanges=0}               => {IsCanceled=1} 0.002193397  0.7767857 0.002823684 1.861625   174
[9]   {LeadTime=178,
       IsRepeatedGuest=0}              => {IsCanceled=1} 0.002218609  0.7364017 0.003012770 1.764841   176
[10]  {LeadTime=178,
       RequiredCarParkingSpaces=0}     => {IsCanceled=1} 0.002218609  0.7586207 0.002924529 1.818091   176
[11]  {LeadTime=277,
       TotalOfSpecialRequests=0}       => {IsCanceled=1} 0.002432906  0.9554455 0.002546358 2.289796   193
[12]  {LeadTime=277,
       BookingChanges=0}               => {IsCanceled=1} 0.002458117  0.8369099 0.002937135 2.005717   195
```

We did a further analysis by changing support to 0.005, the results are as follows:

```
> inspect(ruleset.City)
     lhs                                rhs                support confidence   coverage     lift count
[1]  {PreviousCancellations=1}       => {IsCanceled=1} 0.06257485  0.9629486 0.06498254 2.307778  4964
[2]  {PreviousCancellations=1,
      TotalOfSpecialRequests=0}      => {IsCanceled=1} 0.05803678  0.9850235 0.05891918 2.360682  4604
[3]  {PreviousCancellations=1,
      BookingChanges=0}              => {IsCanceled=1} 0.06130167  0.9696909 0.06321774 2.323936  4863
```

```
[16] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      RequiredCarParkingSpaces=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.05406598  1.0000000 0.05406598 2.396574  4289
[17] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      BookingChanges=0,
      RequiredCarParkingSpaces=0} => {IsCanceled=1} 0.05714178  1.0000000 0.05714178 2.396574  4533
[18] {IsRepeatedGuest=0,
      BookingChanges=0,
      RequiredCarParkingSpaces=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.31482812  0.6092206 0.51677192 1.460042 24975
[19] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      BookingChanges=0,
      RequiredCarParkingSpaces=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.05319618  1.0000000 0.05319618 2.396574  4220
> |
```

Resort

Cancellation was input as the dependent variable. LeadTime, previous cancellations, repeated guests status, booking changes, required car parking spaces and total of special requests were input as independent variables. Initial support was set to 0.002 with a confidence level of 0.6.

Inspection of the association rules revealed the following results:

```
> inspect(ruleset.Resort)
     lhs                                rhs                support confidence    coverage     lift count
[1]  {LeadTime=244}                  => {IsCanceled=1} 0.002121870  0.7589286 0.002795876 2.733494    85
[2]  {LeadTime=259}                  => {IsCanceled=1} 0.002546244  0.8717949 0.002920692 3.140014   102
[3]  {LeadTime=92}                   => {IsCanceled=1} 0.002196760  0.6197183 0.003544771 2.232089    88
[4]  {PreviousCancellations=1}       => {IsCanceled=1} 0.018722384  0.8370536 0.022367009 3.014883   750
[5]  {LeadTime=244,
      RequiredCarParkingSpaces=0}    => {IsCanceled=1} 0.002121870  0.7657658 0.002770913 2.758120    85
[6]  {LeadTime=244,
      IsRepeatedGuest=0}             => {IsCanceled=1} 0.002121870  0.7589286 0.002795876 2.733494    85
[7]  {LeadTime=259,
      PreviousCancellations=1}       => {IsCanceled=1} 0.002271649  1.0000000 0.002271649 3.601780    91
[8]  {LeadTime=259,
      TotalOfSpecialRequests=0}      => {IsCanceled=1} 0.002371502  0.9313725 0.002546244 3.354599    95
[9]  {LeadTime=259,
      BookingChanges=0}              => {IsCanceled=1} 0.002546244  0.9357798 0.002720987 3.370473   102
[10] {LeadTime=259,
      RequiredCarParkingSpaces=0}    => {IsCanceled=1} 0.002546244  0.8717949 0.002920692 3.140014   102
[11] {LeadTime=259,
      IsRepeatedGuest=0}             => {IsCanceled=1} 0.002546244  0.8717949 0.002920692 3.140014   102
```

We did a further analysis by changing support to 0.005, the results are as follows:

```
> inspect(ruleset.Resort)
     lhs                                rhs                support confidence   coverage     lift count
[1]  {PreviousCancellations=1}       => {IsCanceled=1} 0.01872238  0.8370536 0.02236701 3.014883   750
[2]  {PreviousCancellations=1,
      TotalOfSpecialRequests=0}      => {IsCanceled=1} 0.01495294  0.8606322 0.01737437 3.099808   599
```

```
[13] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      BookingChanges=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.01300582  0.9542125 0.01362990 3.436864   521
[14] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      RequiredCarParkingSpaces=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.01340523  0.9470899 0.01415412 3.411210   537
[15] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      BookingChanges=0,
      RequiredCarParkingSpaces=0} => {IsCanceled=1} 0.01590155  0.9564565 0.01662548 3.444946   637
[16] {PreviousCancellations=1,
      IsRepeatedGuest=0,
      BookingChanges=0,
      RequiredCarParkingSpaces=0,
      TotalOfSpecialRequests=0}   => {IsCanceled=1} 0.01300582  0.9559633 0.01360493 3.443170   521
```

With close inspection, previous cancellations, repeated guest status, booking changes, required car parking space and a total of special requests are considered to have significant correlations with consumer cancellation intentions.

Linear Modeling testing factors impacting cancellation

Two multiple linear models were carried out to investigate whether previous cancellations, repeated guest status, booking changes, required car parking space and a total of special requests significantly predict customer stay cancellation. Customer arrival months were also put into the model as control variables.

City

Linear model coefficient results were as follows:

Variable	Estimate	Std. Error	*t* value	*p* value
(Intercept)	0.7645	0.0088	86.543	< 2e-16 ***
Lead Time	0.0008	0.0000	50.126	< 2e-16 ***
Previous Cancellations	0.0571	0.0037	15.439	< 2e-16 ***
Repeated Guest	-0.3561	0.0137	-26.081	< 2e-16 ***
Booking	-0.0931	0.0035	-26.848	< 2e-16 ***

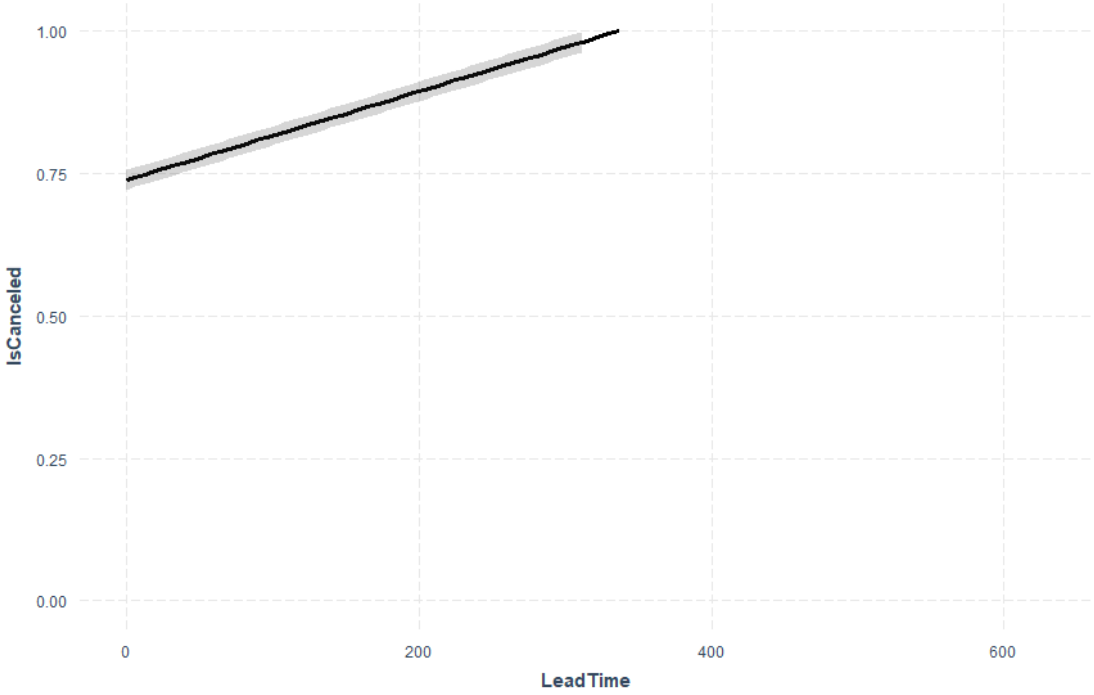
Changes				
Required Car Parking Spaces	-0.5999	0.2181	-27.502	< 2e-16 ***
Total Of Special Requests	-0.0300	0.0029	-10.490	< 2e-16 ***
Month 2	0.0949	0.0122	7.791	6.79e-15 ***
Month 3	0.0774	0.0112	6.901	5.25e-12 ***
Month 4	0.0879	0.0105	8.347	< 2e-16 ***
Month 5	0.0390	0.0103	6.747	0.000179 ***
Month 6	0.1071	0.0106	10.112	< 2e-16 ***
Month 7	-0.0261	0.0105	-2.501	0.012378 *
Month 8	-0.1084	0.0101	-10.771	< 2e-16 ***
Month 9	-0.2562	0.0099	-25.683	< 2e-16 ***
Month 10	-0.2399	0.0099	-24.137	< 2e-16 ***
Month 11	-0.0792	0.0116	-6.801	1.05e-11 ***
Month 12	-0.0039	0.0123	-0.318	0.750267

Note: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The adjusted R-squared is 0.2267 with a F-statistic of 691.6 on 17 and 40041 degrees of freedom. The overall model is significant, which means the overall model explained 22.67% of the variance of people's cancellation.

Significant linear relationships were shown below.

<u>Lead time to cancellation</u>

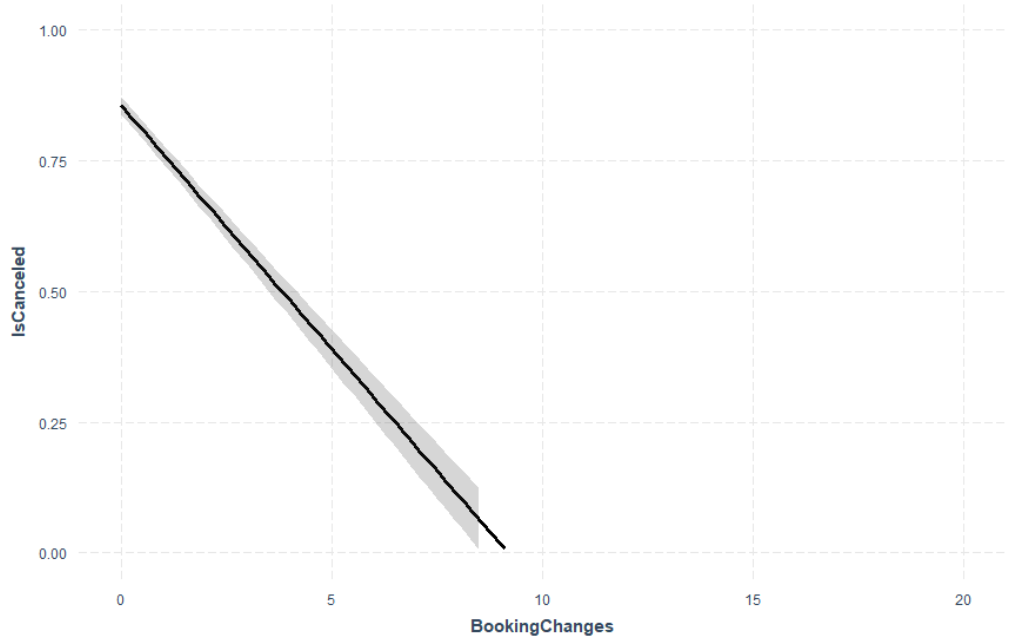


The longer lead time, the more likely the customers would cancel their stay.

Repeated Guests to cancellation

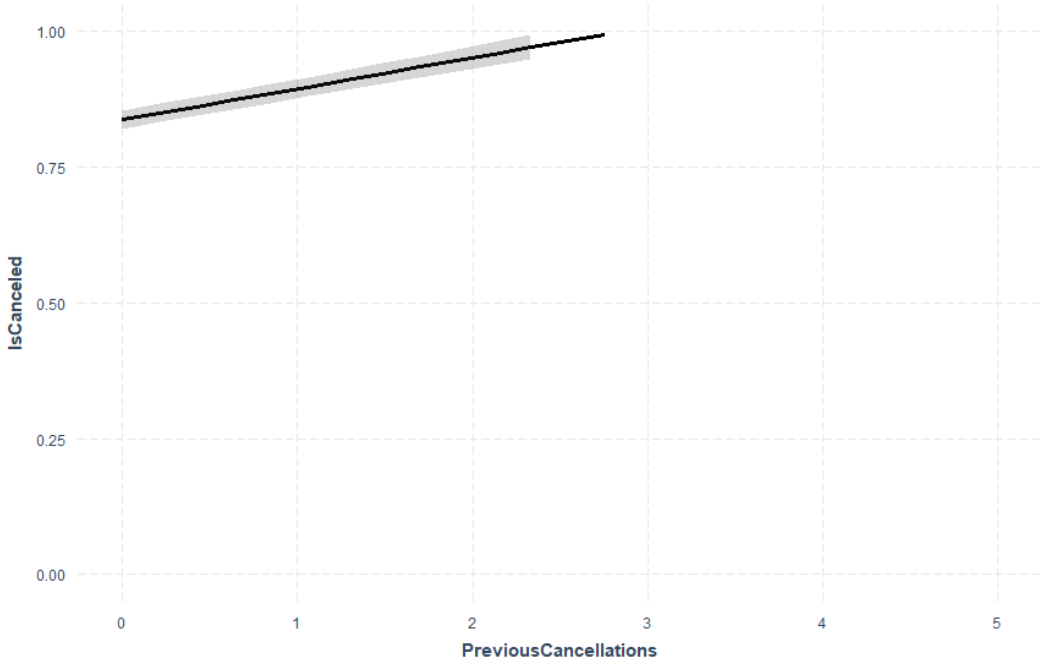
Repeated guests were less likely to cancel their stay.

Booking Changes to Cancellation



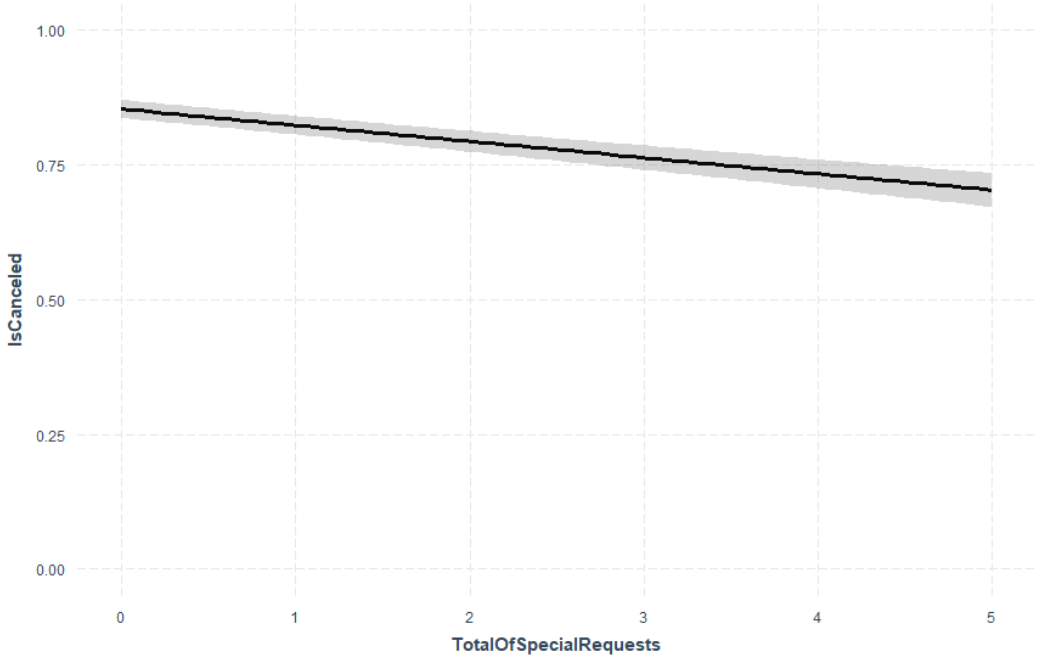
The more booking changes a customer made, the less likely they are to cancel their stay.

Previous Cancellation to Cancellation



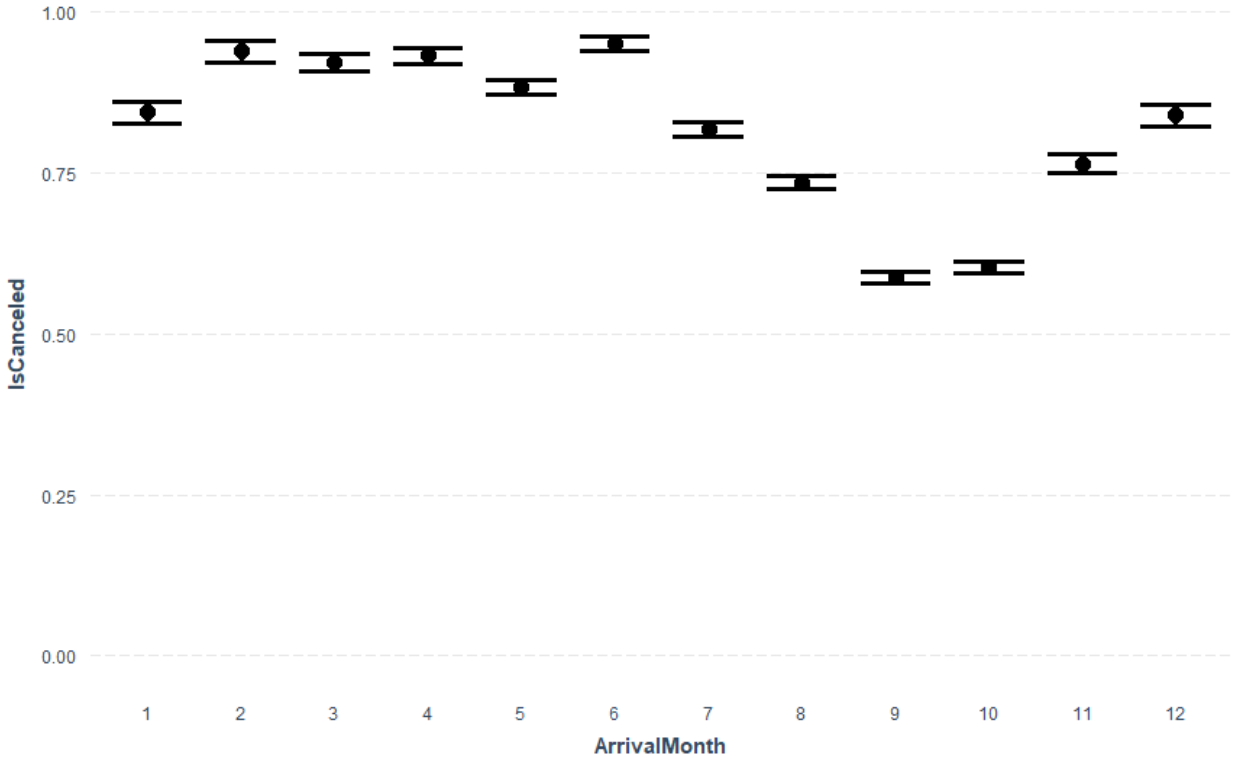
The more previous cancellation customers have, the more likely they are to cancel.

Total of Special Requests to Cancellation



The more total of special requests, the less likely they are to cancel.

Arrival Month to Cancellation.



There is a significant drop in cancellation in September and October.

Resort

Linear model coefficient results were as follows:

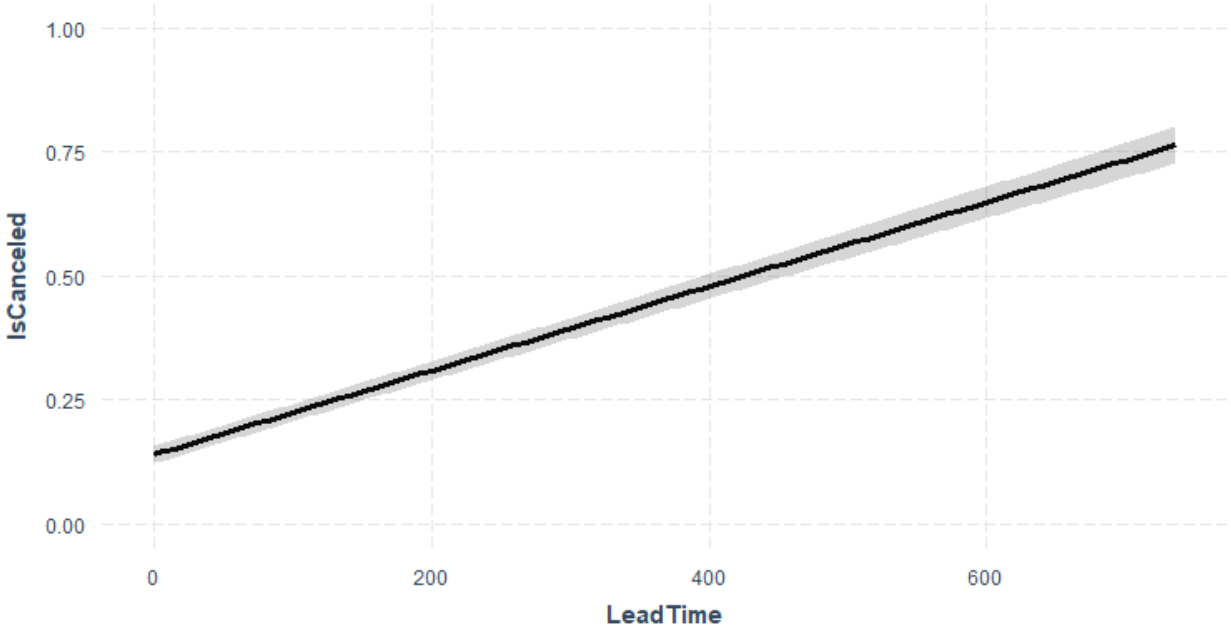
Variable	Estimate	Std. Error	*t* value	*p* value
(Intercept)	2.261e-01	9.183e-03	24.618	< 2e-16 ***
Lead Time	8.476e-04	2.382e-05	35.579	< 2e-16 ***
Previous Cancellations	2.958e-02	1.580e-03	18.722	< 2e-16 ***
Repeated Guest	-1.336e-01	1.028e-02	-13.006	< 2e-16 ***
Booking Changes	-6.861e-02	2.890e-03	-23.745	< 2e-16 ***
Required Car Parking Spaces	-2.498e-01	6.053e-03	-41.269	< 2e-16 ***
Total Of Special Requests	-4.818e-02	2.609e-03	-18.468	< 2e-16 ***
Month 2	7.774e-02	1.163e-02	6.682	2.39e-11 ***
Month 3	3.137e-02	1.150e-02	2.728	0.00637 **
Month 4	8.143e-02	1.138e-02	7.157	8.37e-13 ***
Month 5	4.546e-02	1.153e-02	3.942	8.10e-05 ***
Month 6	8.141e-02	1.197e-02	6.798	1.07e-11 ***
Month 7	8.875e-02	1.102e-02	8.057	8.06e-16 ***
Month 8	1.185e-01	1.091e-02	10.855	< 2e-16 ***
Month 9	2.545e-02	1.206e-02	2.110	0.03489 *
Month 10	2.982e-02	1.153e-02	2.586	0.00970 **
Month 11	1.657e-02	1.228e-02	1.349	0.17730
Month 12	5.483e-02	1.206e-02	4.547	5.46e-06 ***

Note: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The adjusted R-squared is 0.1341 with a F-statistic of 365.8 on 17 and 40041 degrees of freedom. The overall model is significant, which means the overall model explained 13.41% of the variance of people's cancellation.

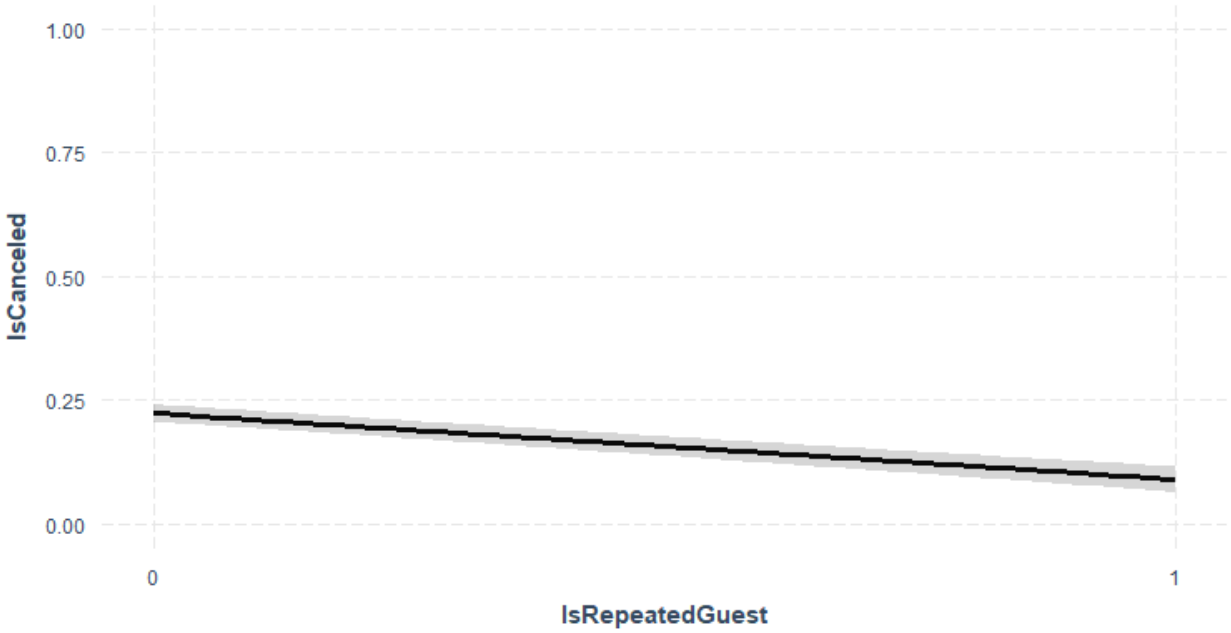
Significant linear relationships were shown below.

Lead time to cancellation



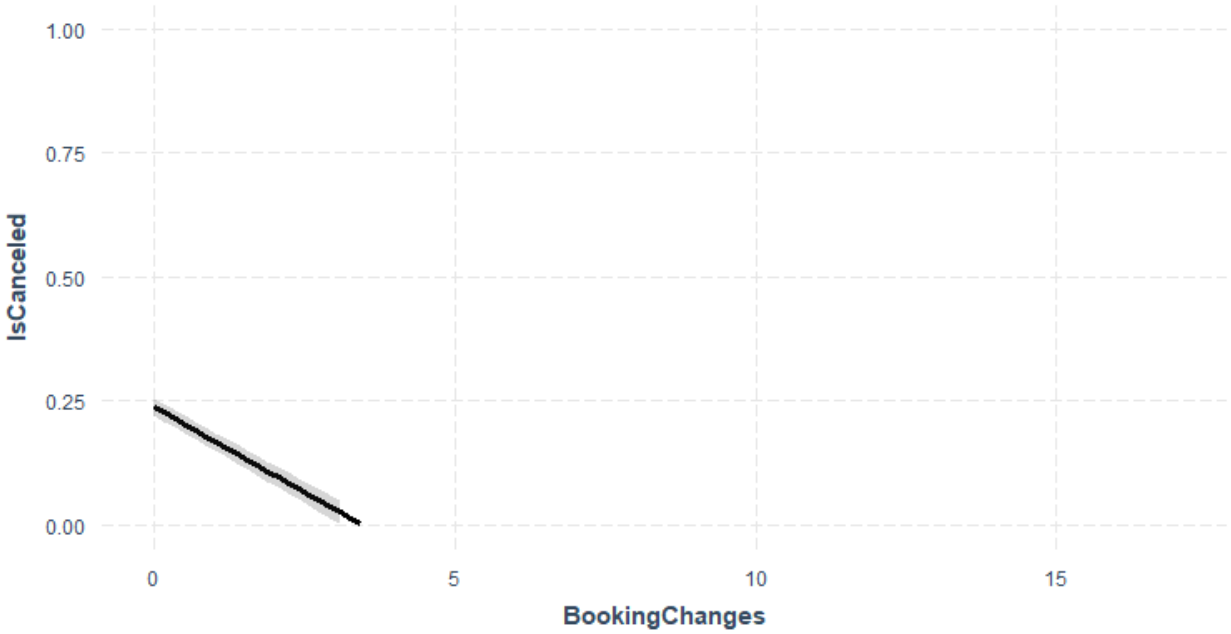
The longer lead time, the more likely the customers would cancel their stay.

Repeated Guests to cancellation



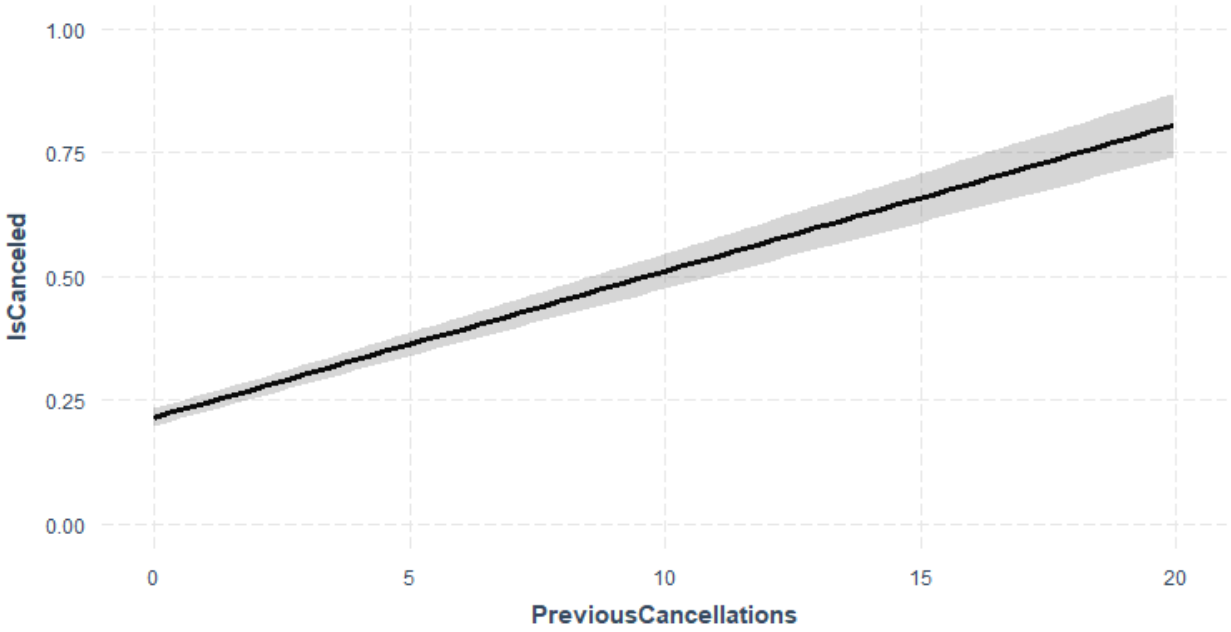
Repeated guests were less likely to cancel their stay.

Booking Changes to Cancellation



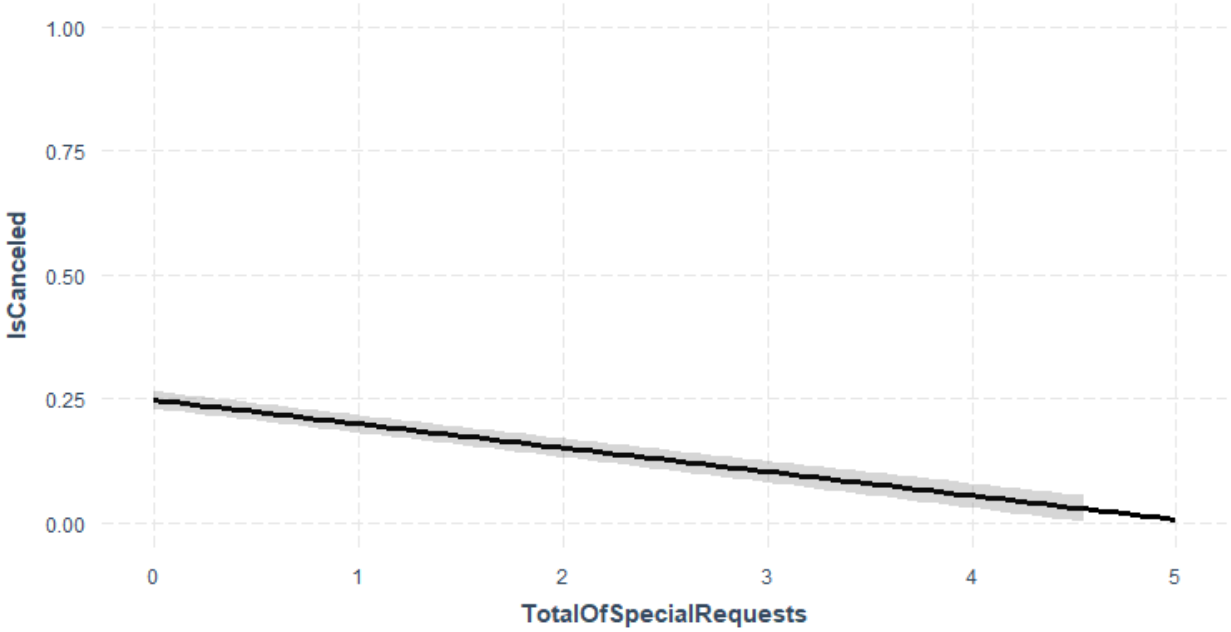
The more booking changes a customer made, the less likely they are to cancel their stay.

Previous Cancellation to Cancellation



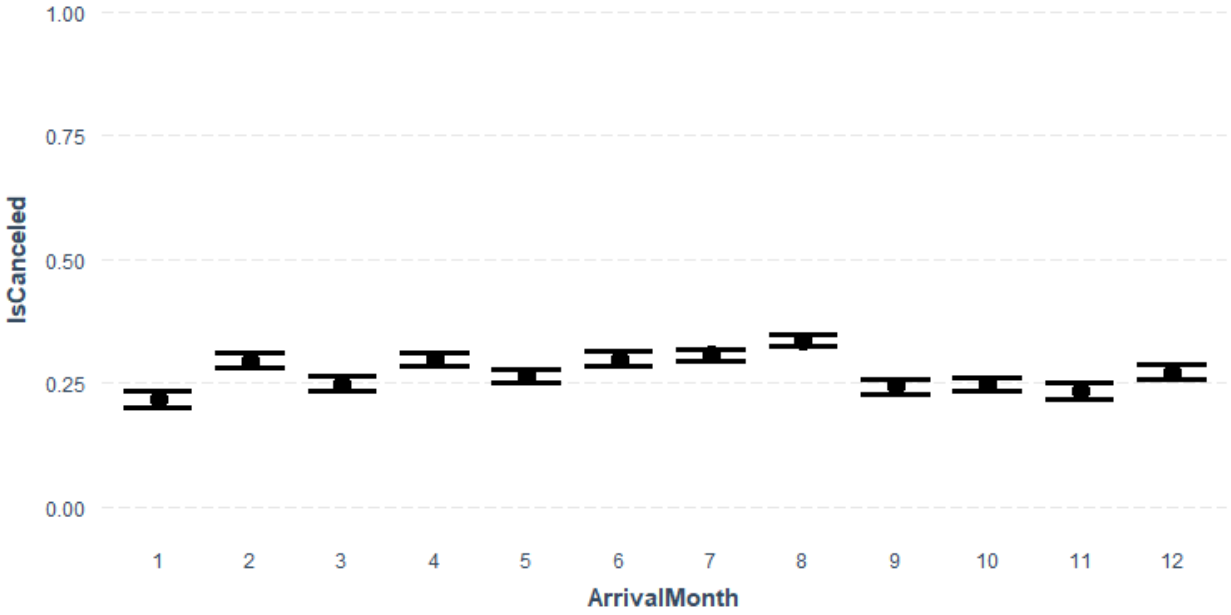
The more previous cancellation customers have, the more likely they are to cancel.

Total of Special Requests to Cancellation



The more total of special requests, the less likely they are to cancel.

Arrival Month to Cancellation.



Linear Modeling testing factors impacting ADR

Two multiple linear models were carried out to investigate whether assigned room type, reserved room type, market segment, and customer arrival month predict ADR.

City

Linear model coefficient results were as follows:

Variable	Estimate	Std. Error	*t* value	*p* value
(Intercept)	62.4191	4.3035	14.504	< 2e-16 ***
Assigned Room B	-10.0296	1.3095	-7.659	1.92e-14 ***
Assigned Room C	-1.5546	7.0269	-0.221	0.82491
Assigned Room D	-0.7854	0.7253	-1.083	0.27889
Assigned Room E	5.0928	1.7702	2.877	0.00402 **
Assigned Room F	17.1672	2.5067	6.848	7.58e-12 ***

Assigned Room G	14.1000	3.5849	3.933	8.40e-05 ***
Assigned Room K	-76.6507	4.5423	-16.875	< 2e-16 ***
Assigned Room P	-14.7332	8.7096	-1.692	0.09073
Reserved Room B	0.1848	1.6961	0.109	0.91323
Reserved Room C	7.7888	13.1393	0.593	0.55333
Reserved Room D	27.8832	0.8172	34.122	< 2e-16 ***
Reserved Room E	51.6920	2.0805	24.846	< 2e-16 ***
Reserved Room F	65.6144	2.6121	25.120	< 2e-16 ***
Reserved Room G	97.6626	4.1348	23.619	< 2e-16 ***
Reserved Room P	NA	NA	NA	NA
Complementary	-96.3488	4.6672	-20.644	< 2e-16 ***
Corporate	-2.5824	4.3335	-0.596	0.55123
Direct	18.0789	4.3042	4.200	2.67e-05 ***
Groups	0.4438	4.2621	0.104	0.91707
Offline TA/TO	9.6131	4.2641	2.254	0.02417 *
Online TA	25.1076	4.2560	5.899	3.68e-09 ***
Undefined	-56.1900	19.7318	-2.848	0.00441 **
Month 2	-1.5772	0.9222	-1.710	0.08723
Month 3	7.3873	0.8494	8.697	< 2e-16 ***
Month 4	23.7455	0.7981	29.753	< 2e-16 ***

Month 5	38.4390	0.7842	49.016	< 2e-16 ***
Month 6	38.2692	0.8016	47.740	< 2e-16 ***
Month 7	16.7883	0.7873	21.324	< 2e-16 ***
Month 8	18.6158	0.7624	24.417	< 2e-16 ***
Month 9	29.1529	0.7607	38.325	< 2e-16 ***
Month 10	16.8678	0.7580	22.253	< 2e-16 ***
Month 11	9.2697	0.8809	10.523	< 2e-16 ***
Month 12	2.8224	0.9294	3.037	0.00239 **

Note: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The adjusted R-squared is 0.4834 with a F-statistic of 1172 on 32 and 40026 degrees of freedom. The overall model is significant, which means the overall model explained 48.34% of the variance of ADR.

Resort

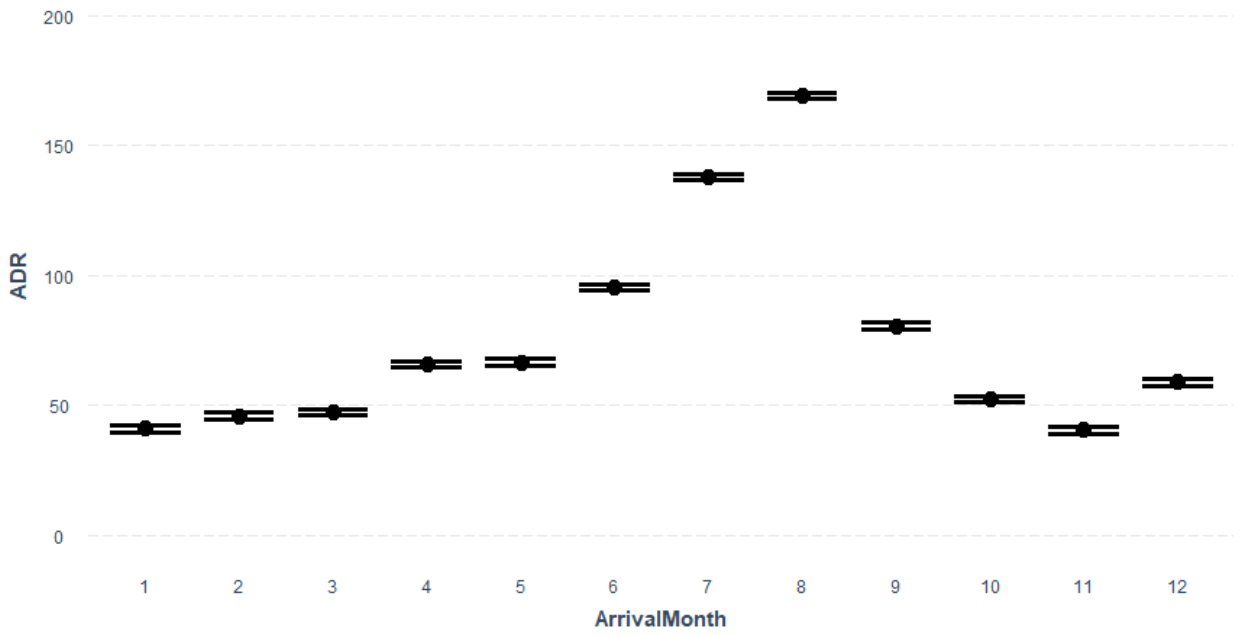
Variable	Estimate	Std. Error	*t* value	*p* value
(Intercept)	40.9300	0.7813	52.385	< 2e-16 ***
Assigned Room B	0.8500	2.7785	0.306	0.759679
Assigned Room C	6.5565	0.9658	6.789	1.15e-11 ***
Assigned Room D	-2.1683	0.6240	-3.475	0.000512 ***
Assigned Room E	-1.5675	0.9875	-1.587	0.112421
Assigned Room F	2.5751	1.3700	1.880	0.060162
Assigned Room G	4.2724	2.0576	2.076	0.037868 *

Assigned Room H	0.1339	2.9121	0.046	0.963327
Assigned Room I	-51.9852	1.8481	-28.130	< 2e-16 ***
Assigned Room L	-125.4497	37.8091	-3.318	0.000908 ***
Assigned Room P	-43.1744	24.4035	-1.769	0.076870
Reserved Room B	13.7277	20.1137	0.683	0.494922
Reserved Room C	42.0117	1.4522	28.929	< 2e-16 ***
Reserved Room D	16.2755	0.6746	24.125	< 2e-16 ***
Reserved Room E	29.6454	1.0280	28.838	< 2e-16 ***
Reserved Room F	40.6037	1.6399	24.760	< 2e-16 ***
Reserved Room G	71.1896	2.1838	32.599	< 2e-16 ***
Reserved Room H	91.2982	3.1559	28.930	< 2e-16 ***
Reserved Room L	-4.5362	15.4595	-0.293	0.769197
Reserved Room P	NA	NA	NA	NA
Month 2	4.8988	0.9652	5.075	3.88e-07 ***
Month 3	6.4490	0.9546	6.756	1.44e-11 ***
Month 4	24.8187	0.9447	26.272	< 2e-16 ***
Month 5	25.6023	0.9496	26.961	< 2e-16 ***

Month 6	54.4015	0.9830	55.342	< 2e-16 ***
Month 7	97.0559	0.9138	106.217	< 2e-16 ***
Month 8	128.3740	0.9062	141.657	< 2e-16 ***
Month 9	39.6257	0.9779	40.520	< 2e-16 ***
Month 10	11.5785	0.9467	12.230	< 2e-16 ***
Month 11	-0.4102	1.0177	-0.403	0.686914
Month 12	17.8857	0.9995	17.895	< 2e-16 ***

Note: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The adjusted R-squared is 0.6846 with a F-statistic of 3000 on 29 and 40028 degrees of freedom. The overall model is significant, which means the overall model explained 68.46% of the variance of ADR.



Future Performance Prediction

Using SVM to predict future cancellation

In this project, several SVM models are used to predict future performance of the two hotels. As stated in previous sections, cancellation can be important in predicting the performance of a hotel. Thus, the following section will be using svm models to predict cancellation. Cancellations were entered as dependent variables while lead time, repeated guests, booking changes, required parking spaces, total of special requests and arrival month were entered as independent variables.

City

	Actual 0	Actual 1
Predicted False 0	4731	1549
Predicted True 1	3745	30034

The svm model predicted 4731 rows matched to the actual not cancellation. 1549 rows were actually canceled.

The svm model predicted 30034 rows matched to the actual cancellation. 3745 that were actually not canceled.

In the svm model for city hotels, the model had a high chance of predicting cancellation right, yet it was not accurate enough with people who are not cancelling. Overall, there seems to be a pattern among the cancellation and variables.

Resort

	Actual 0	Actual 1
Predicted 0	27367	7924
Predicted 1	1570	3198

While the svm model predicted 27367 rows matched to the actual not cancellation, there were 7924 rows that were actually canceled.

While the svm model predicted 3198 rows matched to the actual cancellation, there were 1570 that were actually not canceled

The svm model for resort is not accurate enough to draw cancellation predictions.

Conclusion

Overall, the following conclusion can be drawn from the analyses:

1. Descriptive analysis showed better performance of the resort hotel than the city hotel

From the descriptive analysis, we see that the resort hotel has more types of rooms, lower cancellation rate and higher repeated guest rate and more diverse meal choices. On the other hand, the city hotel has more distribution channels, more market segments and better ADR in average. For city hotels, reducing lead time could really help with the cancellation rate. Resort hotels seem to attract guests during summer time, so they need to think of taking advantage of the summer time.

2. Most customers came from Europe

The descriptive analysis also showed us that most customers for both hotels come from Europe. Future strategies can be formed based on this specific demographic.

3. Association rules mining identified lead time, previous cancellations, repeated guests status, booking changes, required car parking spaces and total of special requests as predictors of customer cancellation

The result indicated that customers tend to cancel based on different factors. Long lead time could induce higher intention to cancel. Repeated guests are less likely to cancel before staying. More previous booking changes would also reduce the likelihood of them to cancel their stay. The more previous cancellation customers have, the more likely they are to cancel. Lastly, customers with more special requests are less likely to cancel.

4. Hotels perform differently during the year

The customer arrival time could be a significant indicator of how the hotel performs. While the resort hotel seemed to do much better in July and August, the city hotel did better in September and October.