MSIS510 Final Report: Descriptive and Predictive Analysis on Hospitality Industry

Gold Team 3: Hai-Au Vo, Ritu Garg, Elaine Zhang, Charles Gay, & Naveed Safavi

Project Goal

The ultimate goal of this project is to provide business leaders in the hospitality industry about the factors that influence customer's decision making when it comes to their choice of lodging. Within this goal we created several sub goals to answer the questions that focus on particular factors. In accomplishing these sub goals, we seek to answer the following questions: how likely are customers going to cancel? What type of hotels do customers prefer? Finally, how likely are customers going to reserve a parking space?

Description of Dataset

The single dataset we are using for this project is 16.07MB in size, consists of 32 columns and has 119,391 entries where each row represents a hotel reservation for city and resort hotels and contains information pertaining to this reservation which is key to answering our business questions. These include columns such as hotel type, whether or not the hotel booking was canceled, and how many parking spaces customers would need, among others. The range of dates for these reservations in the

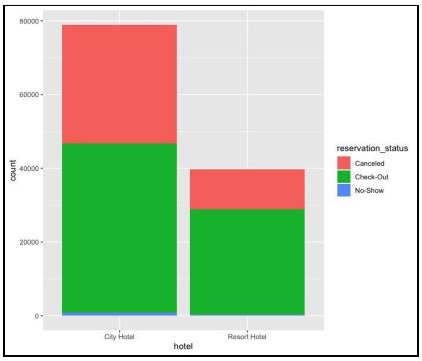
Variable	Type	Description
Hotel	Categorical	(H1 = Resort Hotel or H2 = City Hotel)
ADR	Numeric	Average Daily Rate -Calculated by dividing the sum of all lodging transactions by the total number of staying nights
Adults	Integer	Number of adults
Agent	Categorical	The ID of the travel agency that made the booking
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date
ArrivalDateMonth	Categorical	The month of arrival date with 12 categories: "January" to "December"
ArrivalDateWeek Number	Integer	Week number of the arrival date
ArrivalDateYear	Integer	Year of arrival date
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
Babies	Integer	Number of babies
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
Children	Integer	Number of children
Company	Categorical	The 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 [6]
Customer type	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated with it; Group – when the booking is associated with a group; Transient – when the booking is not part of a group or contract and is not associated with another transient booking; Transient-party – when the booking is transient but is associated with at least another transient booking
DaysInWaitingList	Integer	Number of days the booking was on the waiting list before it was confirmed to the customer

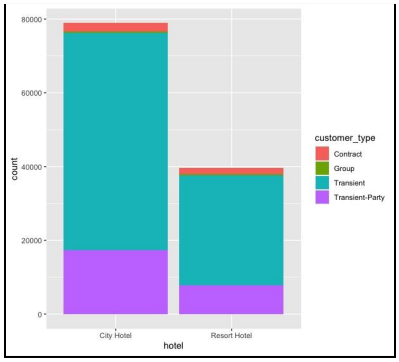
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; In case no payments were found the value is "No Deposit". Non Refund - if the payment was equal or exceeded the total cost of stay Refundable – a deposit was made with a value under the total cost of the stay.
DistributionChannel	Categorical	Distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
IsCanceled	Categorical	A value indicating if the booking was canceled (1) or not (0)
IsRepeatedGuest	Categorical	A value indicating if the booking name was a repeated guest (1) or not (0)
LeadTime	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. 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 canceled by the customer prior to the current booking
PreviousCancellations	Integer	Number of previous bookings that were canceled by the customer prior to the current booking
RequiredCardParking Spaces	Integer	Number of car parking spaces required by the customer
ReservationStatus	Categorical	Reservation the 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 not inform the hotel
ReservationStatusDate	Date	The 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
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	Integer	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Integer	Number of special requests made by the customer

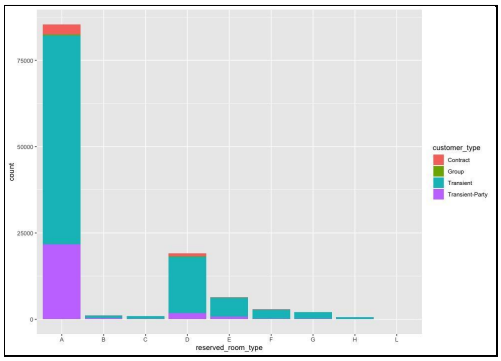
dataset occur between 2015 and 2017. As the question changes, the desired columns will also change accordingly through the course of this project as we explore several variations of this dataset.

Data Visualization and Exploratory Analysis

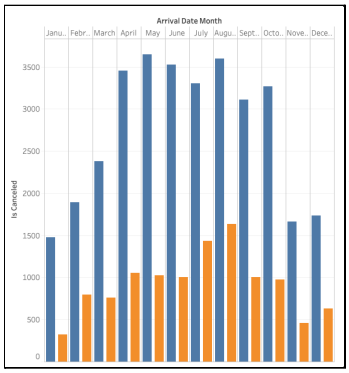
By utilizing ggplot and tableau, we are able to extract and visualize information from the data set to help guide our models and insights:







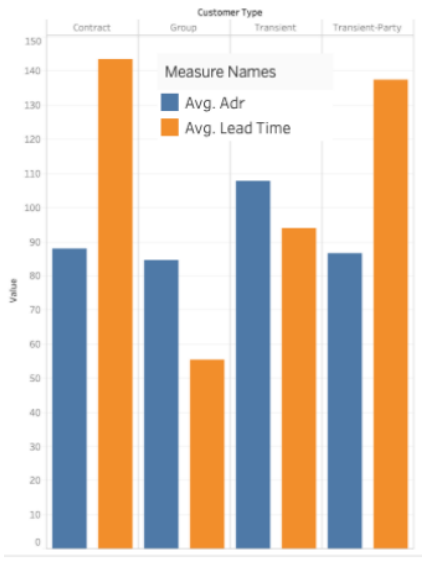
On average, City Hotel captures the majority of the customers however, with a higher cancelation rate proportionally. For both, the most popular room types are A, D, and E with H, D, and F associated at a higher average day rate. Transient customer types make up the majority of the market as well as the higher average day rate cluster.

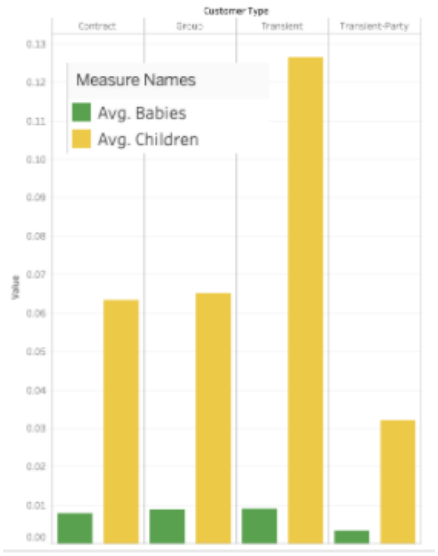


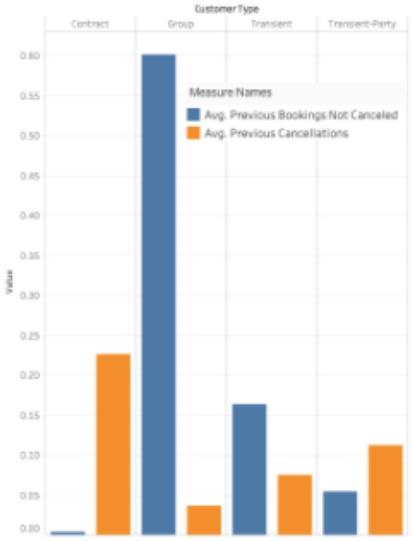
On average, cancellation rates are the lowest during winter months and highest during warmer months among both hotel types. proportionally, resort hotels (orange bars) run a smaller rate of cancellations than city hotel

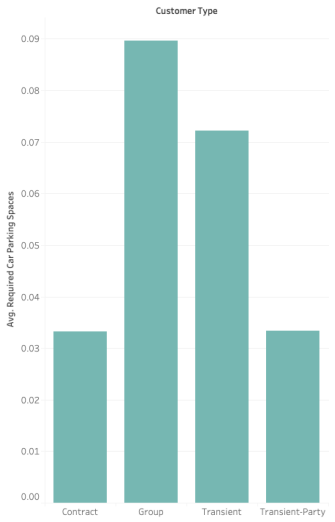
On average, customers with a history of high frequent cancellations, low numbers of special requests, and low numbers of booking changes are more likely to cancel. This make intuitive sense because it shows

that the customer is making less of a investment in their stay (planning & adjusting)









Building a Customer Profile (left to right of x-axis)	
Contract	Average paying customer, tend to bring children + babies, highest rate of cancellations & lead time, less sparking space
Group	Lowest paying, tend to bring children + babies, cleanest record for cancellations, lowest lead time, require parking
Transient	Highest paying, more children + babies, tend to require parking space
Transient-group	Average paying, less likely to bring children + babies, tend to book further out, require less parking space

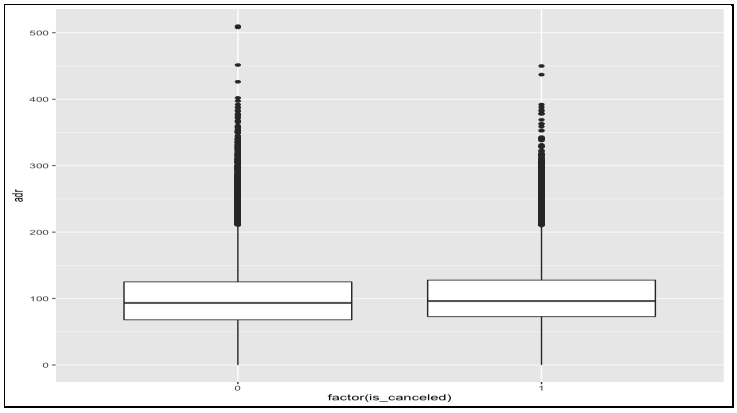
Data Preparation Steps

We took following data preparation steps to make data more manageable:

1) **Missing data handling:**

- Replaced company (90% of rows) and agent column(16k rows) null values with 0.
- Replaced 4 null values in the children column with the average number of children.
- Replaced 488 nulls in the country column with the most frequently occurring country.

2) **Outlier Detection:**



Removed one row in column ‘children’ with value 10, two rows for column ‘babies’ with value 9 and 10, one row for column ‘adr’ with value <0, and one with value 5400(as shown left)

3) Feature engineering:

- Created a new column 'stays_in_total_nights' that is the sum of 'stays_in_week_nights' and 'stays_in_weekend_nights' column.
- Created another column 'total_guests’ that is the sum of ‘adults', 'children' and 'babies'.
- Removed 645 rows in stays_in_total_nights column with value 0.
- Removed 181 rows in Total_guests column with value 0.
- Dropped Reservation_status and reservation_status_date columns as we won't have them at the time of prediction.
- Dropped the company column with more than 90% nulls.

This leaves us with 118,560 rows and 31 columns in total.

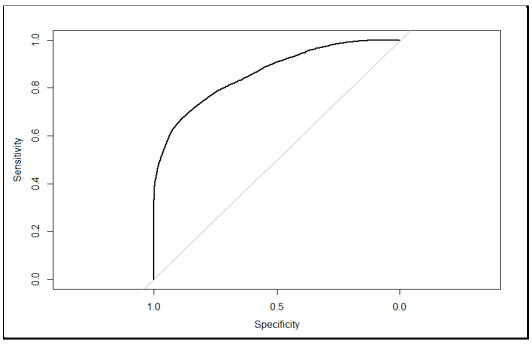
Data Modeling and Methodology

Because we use different data models and analysis methods for each of 3 business questions, we will categorize the methods by questions below, including cluster analysis, logistic regression,decision tree, random forest, correlation matrix and multiple linear regression.

1. How likely are customers to cancel their booking?

Model 1: Logistic Regression

We use Logistic Regression to predict the outcome of a target variable, y, given a set of predictors. In this case, we chose the target variable to be is_canceled, which predicts if a customer will cancel. After



removing variables that would be unavailable prior to a customer's stay, we changed categorical data to factors and releveled them to a known base level.

Afterwards, we ran the logistic regression model with variables in *Appendix 2.1*, using seed 52 and a training split of 70:30. Nearly every variable chosen has a significant impact on the target variable. In terms of cutoff value, we begin with 0.5, giving us the confusion matrix shown in *Appendix 2.2* with sensitivity of 0.62, and specificity of 0.93. From here, we want to choose the best cutoff value, for which there are a few methods. First, we will try to maximize the sensitivity and specificity values using the pROC library to chart the receiver operating characteristic (ROC) curve.

	threshold	specificity	sensitivity
1	0.1	0.2879038	0.9768747
2	0.2	0.5034631	0.9059064
3	0.3	0.7003888	0.8074153
4	0.4	0.8399392	0.7128668
5	0.5	0.9230975	0.6259004

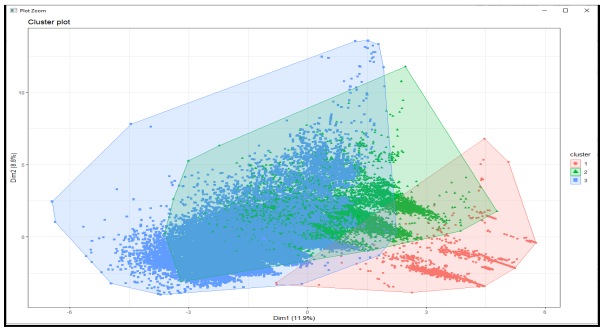
Using the coords function, we find that a cutoff of .445 will produce the best combined specificity and sensitivity, being 0.88 and 0.68 respectively. If we are more interested in predicting customers that will cancel rather than customers that will not, it may be best for us to pick a different cutoff. We can again use the coords function to examine a number of cutoff points at once, giving us the table above. From these models, we decided that we will go with the maximized values, and will choose .445 as our cutoff point, and generated the confusion matrix in *Appendix 2.3*. This leaves us with a final accuracy, sensitivity, and specificity of 80.5%, 0.68, and 0.88.

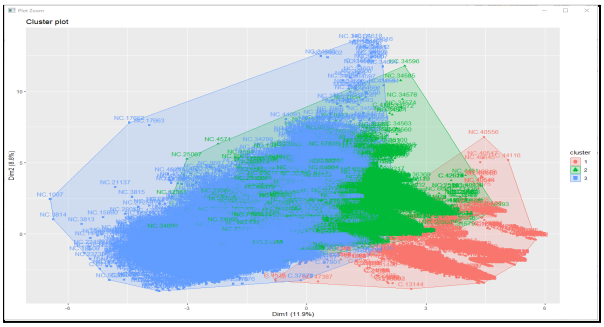
Model 2: Cluster Analysis

Cluster analysis is best for when you don't know what you're looking for in your data. Points within the cluster will be similar to each other. We applied clustering to help us answer the questions of how likely customers are to cancel and what type of hotels do customers prefer.

To start, we decided to pare down some of the columns that are irrelevant to customer profiles. After cutting the columns down there is still some preprocessing required. We first applied one-hot encoding to get a binary numerical representation of categorical columns. Next we normalized the data to make sure it's scaled properly. The categorical variables are all binary representations but the original numeric columns; adults, children, babies, and total_guests, can take on values between 1-50 so we have to make sure the clustering is dependent on just these variables. With these steps done, the data looks like this (refer to *Appendix 1.1*).

We perform the clustering setting centers (clusters) to 3 and nstart to 25 (selects best of 25 configurations). Looking at the visualizations we notice that there is good clustering with some overlap between all the clusters and relatively few. While some overlap is normal in k-means clustering this would suggest that the green and blue clusters have very similar points but there seems to be enough distinct points between the red and blue clusters that we can investigate. Unfortunately the labels (cancelled reservation versus did not cancel) don't give us the best idea about how to label the clusters because of the number of data points so we can print out the values from the kmeans to get an idea for how to label and the number of data points in each cluster.





When looking at the three clusters and the labeled data points that indicate whether the data point represents a canceled reservation or not, we can observe that there is quite a bit of overlap between the three clusters but

	Cluster 1	Cluster 2	Cluster 3
Canceled	14493	5821	23842
Did not cancel (NC)	90	20840	53455

clusters 1 and 3 have some dissimilar points that we can get some insights about.

While cluster 2 and 3 overlap (more NC than canceled points) we can see they are to some extent the opposite of cluster 1 which consists of almost entirely canceled data points. Let's look at the centers (refer to *Appendix 1.2*). From normalized averages of the cluster centers data, we can connect some points in the data in the above table. Cluster 1 almost entirely consists of cancelled reservations. This makes sense when looking at the centers. For example, lead time, the number of days ahead of arrival that a customer makes a reservation, is significantly higher than other clusters. This would suggest that reservations with a higher lead time are more likely to be canceled which is further reinforced by the correlation matrix (*Appendix 3.1*) that classifies lead time as an important predictor. Remaining factors like previous cancellations are higher while booking changes and required number of parking spaces are lower than other clusters. Briefly looking at cluster 2 and 3 (mostly consisting of not canceled reservations) we can see the lead time is significantly lower further indicating the importance of lead time as a factor. It should be noted that because we chose a smaller number of cluster we have a greater total variance of km$tot.withins = 2896009. We chose less clusters because of the distinct clusters. But this can be lowered by increasing the number of clusters based on the elbow method which is determined to be 10 clusters.

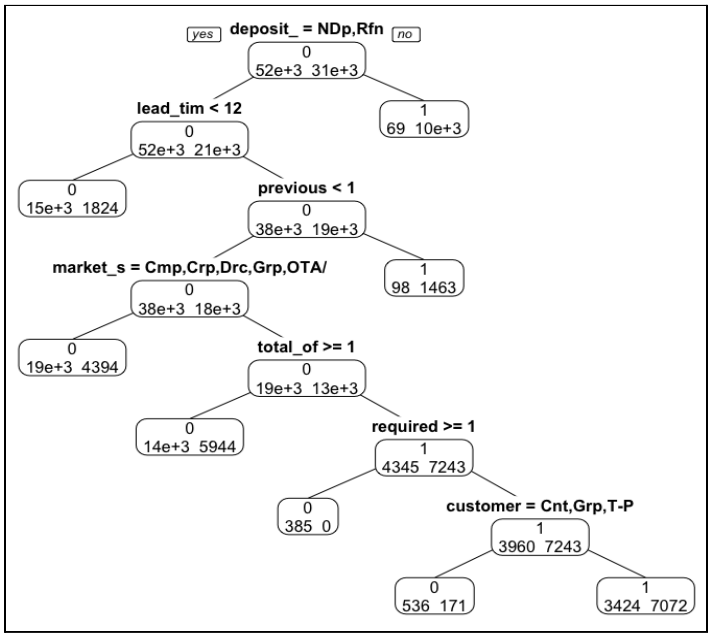
Model 3: Decision tree

We chose to use classification trees in this analysis to identify top predictors and to create classification rules to predict if a customer will cancel. First, we created a correlation matrix shown in *Appendix 3.1* to get an early assessment of predictor variables against the target variable is_cancelled. Based on this matrix we learned that Deposit_type, lead_time, distribution_channel, previous_cancellations,market_segment have high positive correlation and Total_of_special_requests, required_car_parking_spaces, assigned_room_type, booking_changes have high negative correlation

with is_cancelled. We then removed variables Country , Agent, Company , Arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month , meal that we thought were less relevant to our prediction. On running this model, we are able to identify seven important predictors : deposit_type, lead_time, previous_cancellations, market_segment, total_of_special_requests, required_parking and customer_type. This model gave us fair results with 81.5% accuracy and 61% sensitivity shown in *Appendix 3.2.* Then we wrote classification rules and found out that customers who either don't pay any deposit or pay only refundable deposit are more likely to cancel their bookings in cases where it was made more than 11 days prior to the arrival date with some cancellation history. On the other side, customers who either don't pay any deposit or pay only a refundable deposit and made bookings within 11 days from the arrival date are more likely to check-in.

Here are the classification rules :

- IF deposit_type= No Deposit or Refundable AND Lead_time<12 THEN is_canceled=0
- IFdeposit_type= Non Refundable THEN is_canceled=1
- IF deposit_type= No Deposit OR Refundable AND Lead_time>=12 and previous_cancellations>=1 THEN is_canceled=1
- IF deposit_type= No Deposit OR Refundable AND Lead_time>=12 and previous_cancellations<1 AND market_segment= "Complementary" or



"Corporate" or "Direct" or "Groups" or "Online TA" THEN is_canceled=0

- IF deposit_type= No Deposit or Refundable AND Lead_time>=12 AND previous_cancellations<1 AND market_segment= "Aviation" or "offline TA/TO" or "undefined" AND total_of_special_requests>=1 THEN is_canceled=0
- IF deposit_type= No Deposit or Refundable AND Lead_time>=12 AND previous_cancellations<1 AND market_segment= "Aviation" or "offline TA/TO" or "undefined" AND total_of_special_requests<1 AND required_parking>=1 then is_canceled=0
- IF deposit_type= No Deposit or Refundable AND Lead_time>=12 AND previous_cancellations<1 AND market_segment= "Aviation" or "offline TA/TO" or "undefined" AND total_of_special_requests<1 AND required_parking<1 AND customer_type=Contract or group or transient-party THEN is_canceled=0
- IF deposit_type= No Deposit or Refundable AND Lead_time>=12 AND previous_cancellations<1 AND market_segment= "Aviation" or "offline TA/TO" or "undefined" AND total_of_special_requests<1 AND required_parking<1 AND customer_type=transient THEN is_canceled=1

Model 5: Random forest

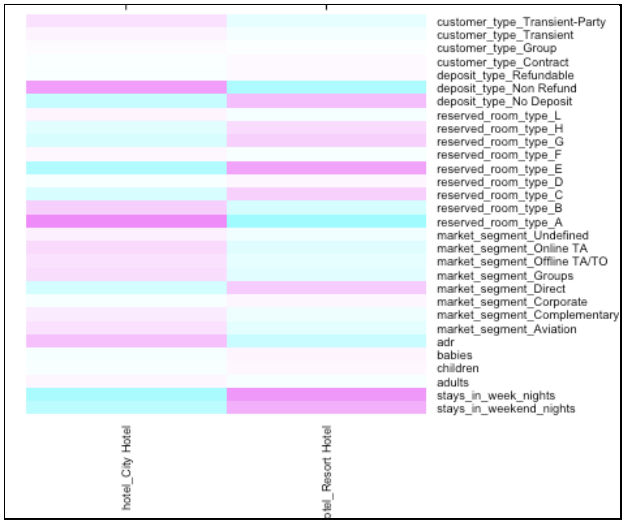
Now that we have the classification tree we want to further enhance its accuracy, particularly sensitivity because our focus is to accurately predict an important class which is 'is_cancelled=1 '.Therefore, we decided to use random forest .It is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation, jointly referred to as bagging, and tends to outperform most other classification methods in terms of accuracy without issues of overfitting. Using this on the same set of predictors we see a substantial improvement in our model accuracy which went up by 4% to give

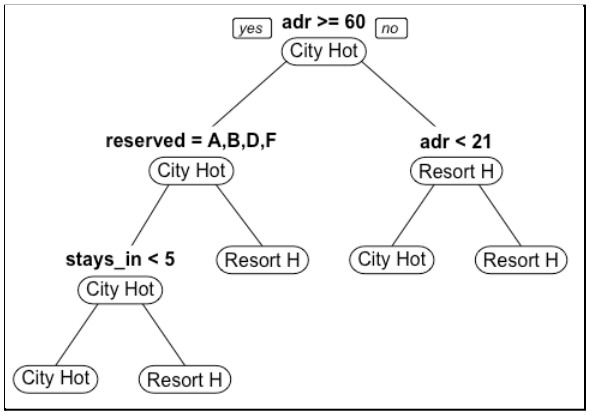
85.5%. However our main focus is model sensitivity which has increased dramatically by 11% to 72%. Results are shown in *Appendix 3.3*.

2. What factors influence the hotel type choice and how can each hotel type target to the right customers?

In order to answer this question, we first run a correlation analysis, followed by a decision tree.

Correlation matrix - Because a correlation matrix shows correlation coefficients between variables, this could be a great starting point for further analysis. In this question, we want to understand the target markets for two hotel types: resort and city hotel. From the matrix, we learn that the correlation between hotel type and "adult", "children" and "babies" are nearly zero. We eliminated these irrelevant variables to simplify the model and save compute time running the decision tree.





Decision Tree - This method helps hotel owners utilize history booking data to gain insights for targeted marketing campaigns and other operational needs. Factors included in the analysis are: customer type, deposit type, reserved room type, average hotel rate, stay in total nights. Accuracy of this model is 82.55%.

From this decision tree, we learned that city hotels should target customers with low budget (avr <21) or customers with a higher budget, but reserves room type of A, B, D, F and book a shorter trip (staying less than 5 days).

Resort hotels should target customers in the budget range of 21-60, customers with a higher budget and reserve room type of C, E, G, H, L, or customers that have a higher budget, reserve cheaper room type(A, B, D, F) but stay more than 5 days.

3. How likely are customers to reserve a parking space and how many are needed if they do?

This question is answered by first a correlation matrix and then a multiple linear regression.

Correlation Matrix - *Appendix 4.1* is the correlation coefficient result when comparing "required car parking spaces" and other factors in the database. Factors on both ends of the bar chart show a strong correlation: hotel type and reserve room type. Factors in the middle part show a weak correlation, which can be eliminated in further linear regression analysis - arrival date month.

Multiple Linear regression - This method is used to predict the number of parking spaces one reservation needs, which further answers the question of the likelihood of reserving a parking space. It is a great model to predict a continuous dependent variable. One challenge is the singularity issue between two groups of columns: "Stays_in_total_nights" = "stays_in_weekend_nights" + "stays_in_week_nights" , "total_guests" = "adults" + "children" + "babies". Because of that, we ran the regression model twice (*Appendix 4.2 and 4.3)*.

The first regression includes total nights and total guests, which shows a general coefficient with these two variables. This is also helpful if future data is incomplete and only has these two data types instead. Second regression includes the other 5 detailed attributes (weekend vs weekdays, adults vs children vs babies) to find out more granular relationship with those attributes. Both *Appendix 4.4 and 4.5* contains only significant attributes and *4.5* contains columns excluded in the second regression. Both model RMSE are 0.24. From the combination of two regressions, we see a clearer picture here: Parking spaces are significantly related to hotel type, year, reserved room type, deposit type, customer type, hotel rate

and total of special requests. Customers with children have less need for parking, while having babies increases the needs.

Results and Insights :

1. Hotel managers can accurately predict bookings that are likely to get cancelled using this model which can be utilized to estimate net demand , improve cancellation policies, and define better overbooking tactics to maximize revenue.
2. City and resort hotels marketing teams should target two different types of customers. For a city hotel, it should reach out to customers who have a very low budget but want to stay in the city: cheaper room type, less staying nights and lower hotel rate. For a resort hotel, it should reach out to those who have a higher budget: either staying in higher end room type or staying in cheaper rooms for more than 5 days.
3. Resort hotel owners should expect more parking spaces requests. All owners should expect a higher need from customers travelling with babies and reserved room type H.

Additional data that may be needed:

- Customer review data to help understand the reason behind cancellations.
- Hotel room and services pricing data to dig deeper into factors leading to cancellations.
- More customer profile data like demographics.

Reflection

Overall, this was a very beneficial project for us. We generated useful insights and developed valuable skills in data analysis. We hadn't yet practiced with such a large dataset, either in attributes or entities. We had a couple issues with using the same attributes as each other, so a more cohesive approach would be beneficial. In addition, we had to convert a large number of categorical variables into numeric for the k-means clustering, and learned about one hot encoding to do so. Finally, we found that we were missing

some extra information that would be extremely beneficial in answering a few of our questions, such as review text or the policies and procedures of the hotels in question.

References

https://www.kaggle.com/jessemostipak/hotel-booking-demand

https://www.sciencedirect.com/science/article/pii/S2352340918315191

Appendix

Appendix 1.1

```
'data.frame':	118560 obs. of  25 variables:
 $ adults                        : num  -1.495 -1.495 0.243 0.243 0.243 ...
 $ children                      : num  -0.262 -0.262 -0.262 -0.262 -0.262 ...
 $ babies                        : num  -0.0873 -0.0873 -0.0873 -0.0873 -0.0873 ...
 $ is_repeated_guest             : num  -0.174 -0.174 -0.174 -0.174 -0.174 ...
 $ booking_changes               : num  -0.342 -0.342 -0.342 -0.342 -0.342 ...
 $ required_car_parking_spaces   : num  -0.255 -0.255 -0.255 -0.255 -0.255 ...
 $ total_of_special_requests     : num  -0.721 -0.721 0.54 0.54 -0.721 ...
 $ total_guests                  : num  -1.3545 -1.3545 0.0391 0.0391 0.0391 ...
 $ hotel_City Hotel              : num  -1.41 -1.41 -1.41 -1.41 -1.41 ...
 $ hotel_Resort Hotel            : num  1.41 1.41 1.41 1.41 1.41 ...
 $ market_segment_Aviation       : num  -0.0442 -0.0442 -0.0442 -0.0442 -0.0442 ...
 $ market_segment_Complementary  : num  -0.0779 -0.0779 -0.0779 -0.0779 -0.0779 ...
 $ market_segment_Corporate      : num  -0.215 4.655 -0.215 -0.215 -0.215 ...
 $ market_segment_Direct         : num  2.92 -0.343 -0.343 -0.343 2.92 ...
 $ market_segment_Groups         : num  -0.447 -0.447 -0.447 -0.447 -0.447 ...
 $ market_segment_Offline TA/TO  : num  -0.504 -0.504 -0.504 -0.504 -0.504 ...
 $ market_segment_Online TA      : num  -0.948 -0.948 1.055 1.055 -0.948 ...
 $ market_segment_Undefined      : num  -0.00411 -0.00411 -0.00411 -0.00411 -0.00411 ...
 $ deposit_type_No Deposit       : num  0.377 0.377 0.377 0.377 0.377 ...
 $ deposit_type_Non Refund       : num  -0.375 -0.375 -0.375 -0.375 -0.375 ...
 $ deposit_type_Refundable       : num  -0.037 -0.037 -0.037 -0.037 -0.037 ...
 $ customer_type_Contract        : num  -0.188 -0.188 -0.188 -0.188 -0.188 ...
 $ customer_type_Group           : num  -0.0694 -0.0694 -0.0694 -0.0694 -0.0694 ...
 $ customer_type_Transient       : num  0.577 0.577 0.577 0.577 0.577 ...
 $ customer_type_Transient-Party: num  -0.517 -0.517 -0.517 -0.517 -0.517 ...
```

Appendix 1.2

```
> km12$centers
  lead_time stays_in_weekend_nights stays_in_week_nights      adults    children      babies is_repeated_guest previous_cancellations
1 1.0144024             -0.31173677           -0.22269632 -0.08443436 -0.2599882 -0.08726022      -0.14882790           0.38237012
2 0.2657927             -0.00555500            0.01353531 -0.21240863 -0.1859743 -0.04824633      -0.02074887          -0.00316269
3 -0.2833937             0.06062131            0.03719252  0.08122355  0.1132500  0.03312196       0.03527221          -0.07102323
  previous_bookings_not_canceled booking_changes required_car_parking_spaces total_of_special_requests total_guests hotel_City Hotel
1                    -0.08432097    -0.323042800                 -0.25511109                -0.7188080   -0.2228270       0.45964591
2                    -0.05296374     0.196520453                 -0.12153527                -0.2700929   -0.2795342      -0.05370757
3                     0.03419544    -0.006782057                  0.09010309                 0.2288749    0.1320964      -0.06788112
  hotel_Resort Hotel market_segment_Aviation market_segment_Complementary market_segment_Corporate market_segment_Direct
1        -0.45964591             -0.04418334                 -0.07789249                -0.10395341            -0.3382694
2         0.05370757             -0.02802258                 -0.05899884                 0.08857258            -0.1927601
3         0.06788112              0.01801041                  0.03506124                -0.01089055             0.1299968
  market_segment_Groups market_segment_Offline TA/TO market_segment_Online TA market_segment_Undefined deposit_type_No Deposit
1            1.2405307                    0.3491181              -0.9406122             -0.004107218             -2.6531108
2            0.5733871                    0.4003036              -0.6569183              0.014157354              0.3604336
3           -0.4317020                   -0.2040134               0.4041858             -0.004107218              0.3761289
  deposit_type_Non Refund deposit_type_Refundable customer_type_Contract customer_type_Group customer_type_Transient
1              2.6698843             -0.03698993             0.01706542         -0.06944322              0.3118841
2             -0.3742023              0.10719457             0.28647381          0.21874344             -1.7315105
3             -0.3745449             -0.02998566            -0.10198507         -0.06476348              0.5386332
  customer_type_Transient-Party
1                    -0.3268433
2                     1.6729749
3                    -0.5152465
```

Appendix 2.2 - Confusion Matrix 0.5 Cutoff (Upper)

Appendix 2.3 - Confusion Matrix 0.445 Cutoff (Lower)

Appendix 2.1 - Logistic Regression Results

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.7444  -0.3574   0.2445   5.7443

Coefficients:
                                 Estimate   Std. Error z value             Pr(>|z|)
(Intercept)                   -1.5977704    0.0671077 -23.809 < 0.0000000000000002 ***
hotelResort Hotel              0.1746224    0.0231246   7.551 0.000000000000043072 ***
lead_time                      0.0035934    0.0001106  32.495 < 0.0000000000000002 ***
stays_in_weekend_nights        0.0374363    0.0107005   3.499             0.000468 ***
stays_in_week_nights           0.0385152    0.0056395   6.830 0.000000000008520582 ***
adults                         0.1513290    0.0202750   7.464 0.000000000000084053 ***
children                       0.2430730    0.0298475   8.144 0.000000000000000383 ***
babies                         0.3724246    0.1176959   3.164             0.001555 **
market_segmentAviation        -0.0859702    0.2102945  -0.409             0.682680
market_segmentComplementary    0.6012341    0.1899397   3.165             0.001549 **
market_segmentDirect           0.1571927    0.1229735   1.278             0.201156
market_segmentGroups           0.1659646    0.0946376   1.754             0.079484 .
market_segmentOffline TA/TO   -0.4662817    0.0979422  -4.761 0.000001928406692236 ***
market_segmentOnline TA        0.8684587    0.0974283   8.914 < 0.0000000000000002 ***
market_segmentUndefined       -5.9515071 6627.1774623  -0.001             0.999283
distribution_channelDirect    -0.5537616    0.1112864  -4.976 0.000000649099229042 ***
distribution_channelGDS       -1.2887263    0.2644166  -4.874 0.000001094453411195 ***
distribution_channelTA/TO     -0.1238383    0.0843091  -1.469             0.141871
distribution_channelUndefined 21.2743275 6623.2532785   0.003             0.997437
is_repeated_guestY            -0.5849426    0.1050607  -5.568 0.000000025818004448 ***
previous_cancellations         2.7063716    0.0712264  37.997 < 0.0000000000000002 ***
previous_bookings_not_canceled -0.4692777    0.0287424 -16.327 < 0.0000000000000002 ***
reserved_room_typeB            0.6003561    0.1214912   4.942 0.000000774992692637 ***
reserved_room_typeC            1.4302506    0.1577996   9.064 < 0.0000000000000002 ***
reserved_room_typeD            1.0934548    0.0545367  20.050 < 0.0000000000000002 ***
reserved_room_typeE            2.0205233    0.1132000  17.849 < 0.0000000000000002 ***
reserved_room_typeF            2.0780340    0.1676349  12.396 < 0.0000000000000002 ***
reserved_room_typeG            3.0202711    0.2543107  11.876 < 0.0000000000000002 ***
reserved_room_typeH            2.1675150    0.5235193   4.140 0.000034688643603647 ***
reserved_room_typeL          -10.1593823  177.6146086  -0.057             0.954387
assigned_room_typeB           -0.6534572    0.0966119  -6.764 0.000000000013447982 ***
assigned_room_typeC           -1.4930680    0.1205361 -12.387 < 0.0000000000000002 ***
assigned_room_typeD           -1.2765317    0.0520504 -24.525 < 0.0000000000000002 ***
assigned_room_typeE           -2.1025330    0.1095806 -19.187 < 0.0000000000000002 ***
assigned_room_typeF           -2.6476320    0.1587853 -16.674 < 0.0000000000000002 ***
assigned_room_typeG           -3.4743685    0.2463202 -14.105 < 0.0000000000000002 ***
assigned_room_typeH           -2.4350652    0.5106756  -4.768 0.000001857676500890 ***
assigned_room_typeI           -3.4081216    0.5404760  -6.306 0.000000000286750622 ***
assigned_room_typeK           -1.9990078    0.3951800  -5.058 0.000000422624322339 ***
assigned_room_typeL           22.4074286  370.1424343   0.061             0.951728
booking_changes               -0.3532588    0.0183510 -19.250 < 0.0000000000000002 ***
deposit_typeNon Refund         5.4568074    0.1356275  40.234 < 0.0000000000000002 ***
deposit_typeRefundable        -0.0515132    0.2642851  -0.195             0.845459
days_in_waiting_list          -0.0003464    0.0005937  -0.584             0.559532
customer_typeContract         -0.8709972    0.0635262 -13.711 < 0.0000000000000002 ***
customer_typeGroup            -1.1939440    0.2097060  -5.693 0.000000012452018701 ***
customer_typeTransient-Party  -0.4935477    0.0334230 -14.767 < 0.0000000000000002 ***
adr                            0.0035930    0.0002436  14.750 < 0.0000000000000002 ***
required_car_parking_spaces  -38.9657553   81.3231969  -0.479             0.631834
total_of_special_requests     -0.7254480    0.0138512 -52.374 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20727  4999
         1  1570  8272

               Accuracy : 0.8153
                 95% CI : (0.8112, 0.8193)
    No Information Rate : 0.6269
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5834

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6233
            Specificity : 0.9296
         Pos Pred Value : 0.8405
         Neg Pred Value : 0.8057
             Prevalence : 0.3731
         Detection Rate : 0.2326
   Detection Prevalence : 0.2767
      Balanced Accuracy : 0.7765

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19730  4265
         1  2567  9006

               Accuracy : 0.8079
                 95% CI : (0.8038, 0.812)
    No Information Rate : 0.6269
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5785

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6786
            Specificity : 0.8849
         Pos Pred Value : 0.7782
         Neg Pred Value : 0.8223
             Prevalence : 0.3731
         Detection Rate : 0.2532
   Detection Prevalence : 0.3254
      Balanced Accuracy : 0.7817

       'Positive' Class : 1
```

Appendix 3.1- Correlation Matrix

Attribute	Correlation value
deposit_type	0.468675519
lead_time	0.292875656
distribution_channel	0.16770699
previous_cancellations	0.110139263
market_segment	0.059419331
days_in_waiting_list	0.054301413
adr	0.046491987
total_guests	0.044826266
stays_in_week_nights	0.02554232
stays_in_total_nights	0.018553877
stays_in_weekend_nights	-0.001323252
reserved_room_type	-0.062216205
customer_type	-0.068206356
is_repeated_guest	-0.08374545
hotel	-0.137082143
booking_changes	-0.144831563
assigned_room_type	-0.175842975
required_car_parking_spaces	-0.195701443
total_of_special_requests	-0.234877003

Appendix 3.2 -Confusion Matrix for Decision Tree

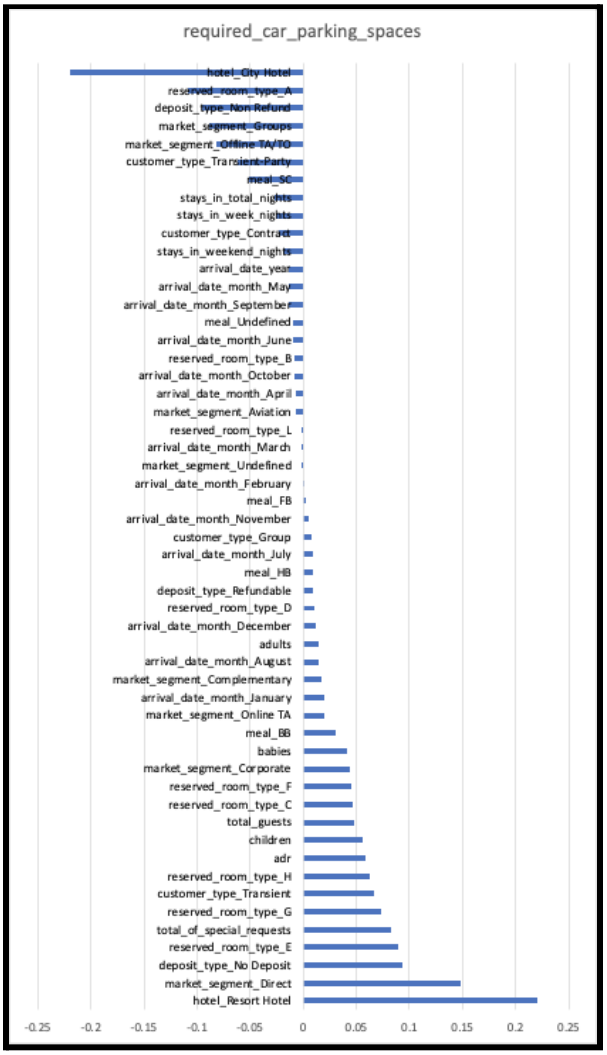
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21102  5138
         1  1482  8041

               Accuracy : 0.8149
                 95% CI : (0.8108, 0.8189)
    No Information Rate : 0.6315
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa: 0.5779

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6101
            Specificity : 0.9344
         Pos Pred Value : 0.8444
         Neg Pred Value : 0.8042
             Prevalence : 0.3685
         Detection Rate : 0.2248
   Detection Prevalence : 0.2663
      Balanced Accuracy: 0.7723

       'Positive' Class: 1
```

Appendix 3.3 - Random Forest Model results

```
Call:
 randomForest(formula = factor(is_canceled) ~ .,
 data = train.df, type = "classification", ntree = 500,
 importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

 OOB estimate of  error rate: 15.14%
Confusion matrix:
      0     1 class.error
0 48584  3843  0.07330192
1  8791 22229  0.28339781
```

```
          Reference
Prediction     0     1
         0 21028  3646
         1  1556  9533

               Accuracy : 0.8545
                 95% CI : (0.8508, 0.8582)
    No Information Rate : 0.6315
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa: 0.6768

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7233
            Specificity : 0.9311
         Pos Pred Value : 0.8597
         Neg Pred Value : 0.8522
             Prevalence : 0.3685
         Detection Rate : 0.2666
   Detection Prevalence : 0.3101
      Balanced Accuracy: 0.8272

       'Positive' Class: 1
```

Appendix 4.1



Appendix 4.2

```
Coefficients:
                                 Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                  27.92303548  2.63283901  10.606 < 0.0000000000000002 ***
hotelResort Hotel             0.10705274  0.00211950  50.509 < 0.0000000000000002 ***
arrival_date_year            -0.01384485  0.00130624 -10.599 < 0.0000000000000002 ***
adults                        0.00553138  0.00312726   1.769              0.076938 .
market_segmentComplementary   0.05018066  0.02318986   2.164              0.030475 *
market_segmentCorporate       0.04268372  0.02055105   2.077              0.037809 *
market_segmentDirect          0.07441880  0.02032448   3.662              0.000251 ***
market_segmentGroups         -0.02471036  0.02041386  -1.210              0.226103
market_segmentOffline TA/TO  -0.02279160  0.02026220  -1.125              0.260663
market_segmentOnline TA      -0.00798938  0.02017712  -0.396              0.692134
market_segmentUndefined      -0.04817662  0.23385330  -0.206              0.836782
reserved_room_typeB           0.00283970  0.00947773   0.300              0.764469
reserved_room_typeC           0.02782147  0.01091498   2.549              0.010808 *
reserved_room_typeD          -0.00315966  0.00260280  -1.214              0.224772
reserved_room_typeE           0.03781168  0.00413362   9.147 < 0.0000000000000002 ***
reserved_room_typeF           0.02807106  0.00659717   4.255   0.0000209303049127 ***
reserved_room_typeG           0.06008456  0.00785978   7.645 < 0.0000000000000212 ***
reserved_room_typeH           0.12353724  0.01337679   9.235 < 0.0000000000000002 ***
reserved_room_typeL          -0.24100552  0.11643205  -2.070              0.038463 *
deposit_typeNon Refund       -0.01730363  0.00366768  -4.718   0.0000023878002533 ***
deposit_typeRefundable        0.03583834  0.02425506   1.478              0.139530
customer_typeGroup            0.01008907  0.01380583   0.731              0.464914
customer_typeTransient        0.01983942  0.00506856   3.914   0.0000907840583940 ***
customer_typeTransient-Party -0.00200518  0.00533481  -0.376              0.707016
adr                           0.00023103  0.00002329   9.918 < 0.0000000000000002 ***
total_of_special_requests     0.01671030  0.00123637  13.516 < 0.0000000000000002 ***
stays_in_total_nights        -0.00863455  0.00037138 -23.250 < 0.0000000000000002 ***
total_guests                 -0.00507776  0.00283069  -1.794              0.072846 .
```

Appendix 4.3

```
Coefficients:
                                 Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                  27.66506408  2.65280601  10.429 < 0.0000000000000002 ***
hotelResort Hotel             0.11140238  0.00226552  49.173 < 0.0000000000000002 ***
arrival_date_year            -0.01371958  0.00131619 -10.424 < 0.0000000000000002 ***
stays_in_weekend_nights      -0.00799682  0.00102100  -7.832  0.0000000000000485 ***
stays_in_week_nights         -0.00867287  0.00054489 -15.917 < 0.0000000000000002 ***
adults                        0.00058312  0.00153400   0.380              0.703850
children                     -0.00935487  0.00294093  -3.181              0.001469 **
babies                        0.04314201  0.00996343   4.330  0.0000149287293308 ***
mealFB                       -0.05197709  0.01087001  -4.782  0.0000017416756433 8 ***
mealHB                       -0.01566215  0.00294056  -5.326  0.0000001005680085 8 ***
mealSC                       -0.00931817  0.00344813  -2.702              0.006886 **
mealUndefined                -0.06521453  0.00923853  -7.059  0.0000000000169269 ***
market_segmentComplementary   0.05531342  0.02318796   2.385              0.017061 *
market_segmentCorporate       0.04211971  0.02053755   2.051              0.040284 *
market_segmentDirect          0.07388855  0.02031333   3.637              0.000276 ***
market_segmentGroups         -0.01944731  0.02040426  -0.953              0.340542
market_segmentOffline TA/TO  -0.01954358  0.02025353  -0.965              0.334575
market_segmentOnline TA      -0.00645641  0.02017720  -0.320              0.748980
market_segmentUndefined      -0.04568797  0.23366901  -0.196              0.844983
reserved_room_typeB           0.00348165  0.00949338   0.367              0.713810
reserved_room_typeC           0.02786391  0.01091752   2.552              0.010706 *
reserved_room_typeD          -0.00594008  0.00266114  -2.232              0.025608 *
reserved_room_typeE           0.03503560  0.00414801   8.446 < 0.0000000000000002 ***
reserved_room_typeF           0.02774567  0.00664477   4.176  0.0000297606042703 ***
reserved_room_typeG           0.05794316  0.00791177   7.324  0.0000000000024383 ***
reserved_room_typeH           0.11936408  0.01339973   8.908 < 0.0000000000000002 ***
reserved_room_typeL          -0.24547548  0.11634088  -2.110              0.034865 *
deposit_typeNon Refund       -0.01908945  0.00369162  -5.171  0.0000002334432561 1 ***
deposit_typeRefundable        0.02770585  0.02425822   1.142              0.253407
customer_typeGroup            0.01089109  0.01379543   0.789              0.429839
customer_typeTransient        0.02080102  0.00506787   4.104  0.0000405660610542 9 ***
customer_typeTransient-Party  0.00093331  0.00535470   0.174              0.861632
adr                           0.00027806  0.00002434  11.422 < 0.0000000000000002 ***
total_of_special_requests     0.01588052  0.00124196  12.787 < 0.0000000000000002 ***
```

Appendix 4.4

	Estimate	Std. Error	t value		Pr(>|t|)	
(Intercept)	27.92303548	2.63283901	10.606	<	2E-16	***
hotelResort Hotel	0.10705274	0.0021195	50.509	<	2E-16	***
arrival_date_year	-0.01384485	0.00130624	-10.599	<	2E-16	***
market_segmentDirect	0.0744188	0.02032448	3.662		0.000251	***
reserved_room_typeE	0.03781168	0.00413362	9.147	<	2E-16	***
reserved_room_typeF	0.02807106	0.00659717	4.255		2.09303E-05	***
reserved_room_typeG	0.06008456	0.00785978	7.645		2.12E-14	***
reserved_room_typeH	0.12353724	0.01337679	9.235	<	2E-16	***
deposit_typeNon Refund	-0.01730363	0.00366768	-4.718		2.3878E-06	***
customer_typeTransient	0.01983942	0.00506856	3.914		9.07841E-05	***
adr	0.00023103	0.00002329	9.918	<	2E-16	***
total_of_special_requests	0.0167103	0.00123637	13.516	<	2E-16	***
stays_in_total_nights	-0.00863455	0.00037138	-23.25	<	2E-16	***

Appendix 4.5

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	27.66506408	2.65280601	10.429	< 0.0000000000000002	***
stays_in_weekend_nights	-0.00799682	0.001021	-7.832	4.85E-15	***
stays_in_week_nights	-0.00867287	0.00054489	-15.917	< 0.0000000000000002	***
children	-0.00935487	0.00294093	-3.181	0.001469	**
babies	0.04314201	0.00996343	4.33	1.49287E-05	***