Predicting the Cost of Hotel Booking in the United States

LAM-duh: Xuliang Deng, Leah Okamura, Megan Richards

Dec 6th, 2021

Contents

1 Introduction/Data

1.1 Introduction

Hotels are a critical component of the travel sector in the U.S., with an estimated 5.3 million guest rooms, and supporting 1 in 25 American jobs on average prior to 2020 based on an American Hotel and Lodging Association Report. A 2018 report showed that approximately two thirds of Americans book their hotels directly through hotel websites, and a 2019 analysis showed large discrepancies in consumer travel cost based on how consumers chose to prioritize booking lodging compared to activities.

We are interested in analyzing U.S. hotel bookings, with the goal of providing insight to consumers on the factors that affect their average daily hotel cost. We believe there will be several significant points of relevance for understanding these relationships: understanding predictors of room cost could be used to help travelers to plan financially for future travel, or to potentially reduce cost. In this report, we are looking to use a variety of chosen models to understand the contributing factors to the average daily rate of a hotel room, as well as identify the strongest predictors.

To better understand the contributing factors to the daily rate of hotel rooms in the U.S., we build models to:

1) Evaluate the efficacy of this model to predict daily hotel room rates
2) Evaluate the statistical significance of each predictor, identifying the strongest contributing factors.

We find a model that is a reasonable predictor of daily hotel room rates in the U.S., and find that purchase of a meal plan, booking in August or October, and having previous bookings that were not cancelled were the most correlated with larger increases in daily hotel rate, and that being a repeated guest, booking in a Resort hotel instead of a City hotel, and booking in December were most correlated with decreases in daily hotel rate. We were also able to show that our model generalized to new data, which means that it could reasonably be used to predict hotel rates in the U.S., with an expected error of about $33.

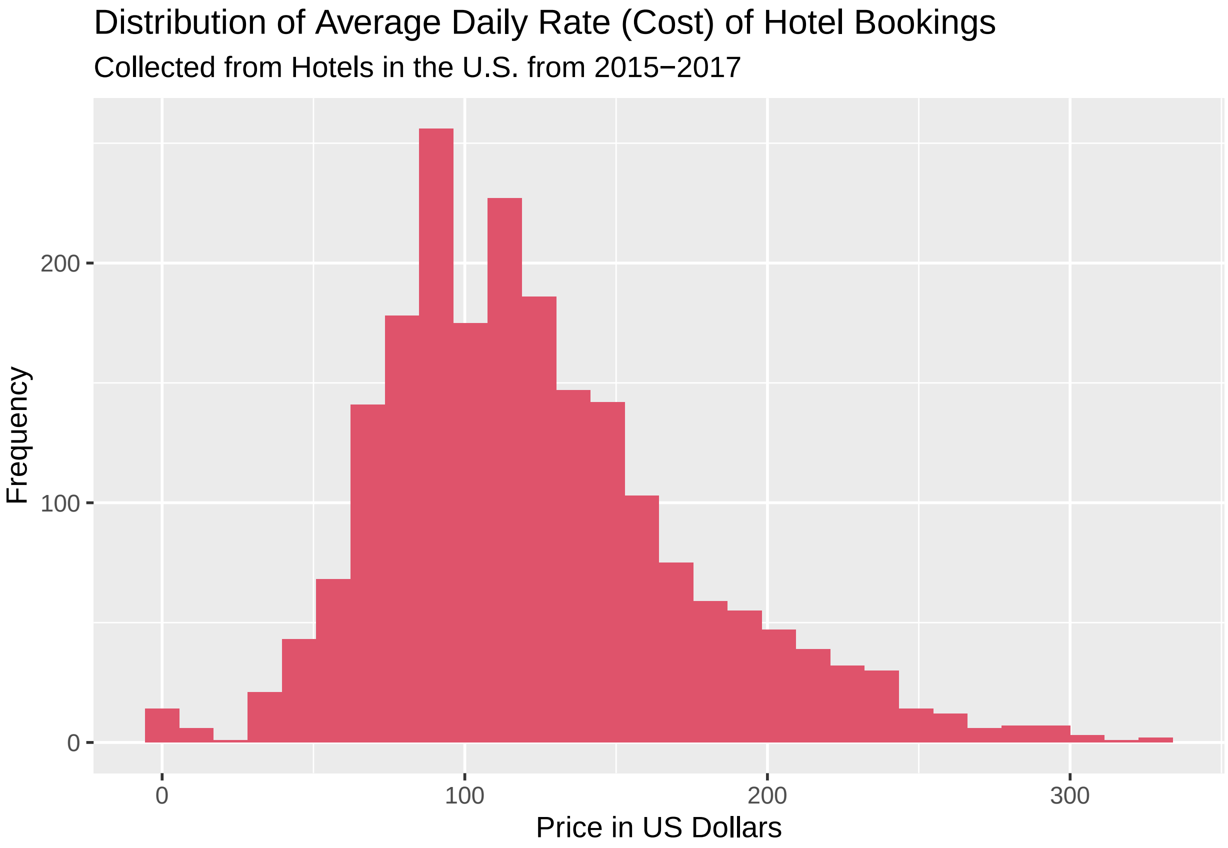
1.2 Data

Our data set is hosted on Tiny Tuesday, was originally collected through an global open hotel booking demand dataset (Antonio, de Almeida, and Nunes 2019), and was originally analyzed in a 2019 study. Due to the dataset being over 100,000 observations, we have limited the observations to be only hotels from the US.

The dataset contains characteristics of hotel bookings and hotel stays. Each observation represents one booking/stay at a specific hotel in America, containing information about the hotel itself, the booking details, and the occupants. The full data set dictionary can be found in the repository ReadMe. For this analysis, we evaluate how variables affect the average daily rate or daily cost. Our methodologies section includes information on predictor selection.

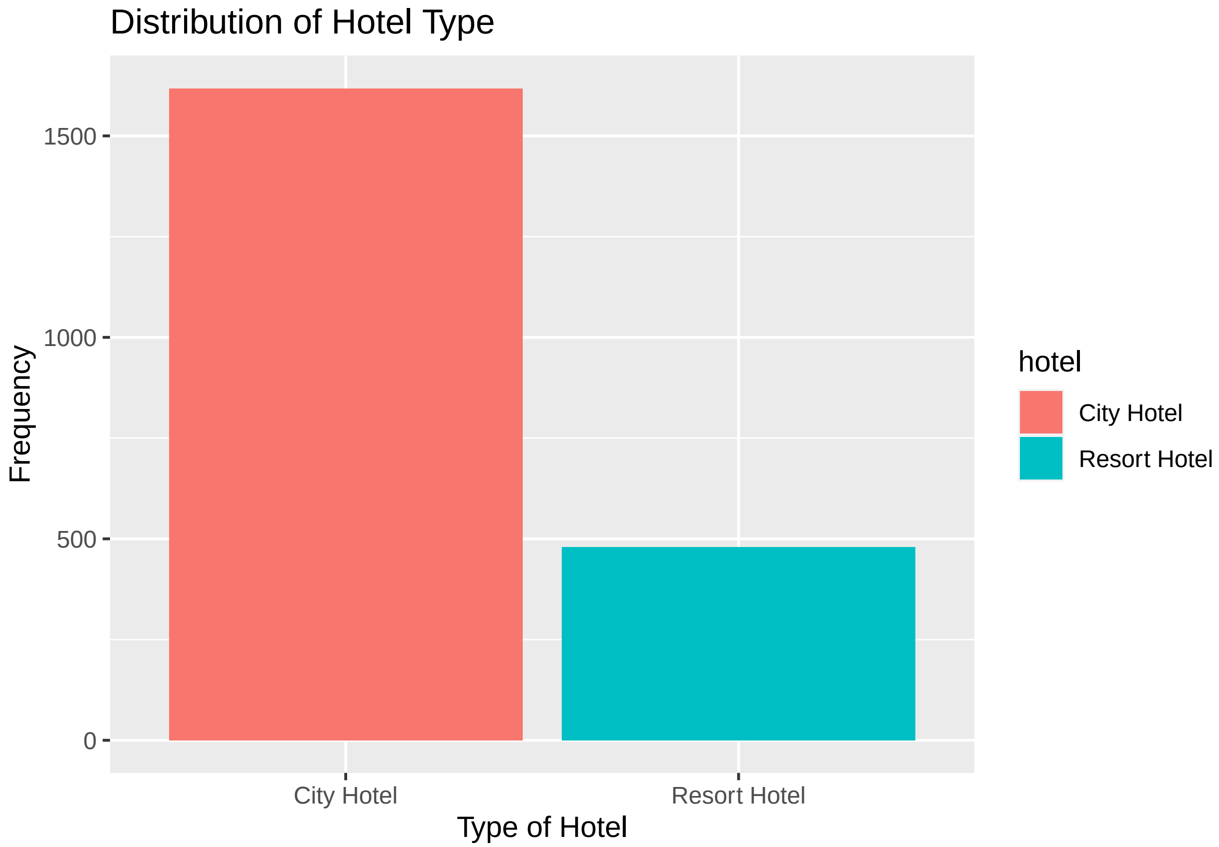
1.3 Exploratory Data Analysis

We start our analysis by exploring the data. Since we are interested in predicting the average daily rate, we first plot its distribution and calculate its summary statistics.



mean	median	sd	min	max	iqr
122.992	115	51.617	0	328.33	61.99

The response variable, average daily rate, has a somewhat skewed right, bimodal distribution. The average or mean average daily rate is $122.992 and the median is $115. Because the distribution is skewed, the median is most likely the best indicator for the center. The standard deviation is $51.617 and the data ranges from $0 to $328.33 with an interquartile range of $61.99.



```
## [1] 2.304792e-05
```

We are particularly interested in using the hotel type (Resort versus City Hotel) as a predictor of average daily rate. As an initial analysis, we perform a t-test on the average daily rates of bookings for resort versus city hotels. This t-test assesses if there is a significant difference between the two distributions, which would indicate that hotel type is likely to be a useful predictor of average daily rate. In this analysis, we find that our t-test produces a p-value of approximately 0. Because this p-value is very low, we conclude that there is a statistically significant relationship between the average daily hotel rates of resort hotels and city hotels in our dataset, which provides evidence that hotel type is a useful predictor of average daily hotel rates.

hotel	n	mean	sd
City Hotel	1618	119.773	45.189
Resort Hotel	479	133.865	67.982

We can see that resort hotel is on average more expensive, and that most customers booked at city hotels. We are interested in whether the classification of City Hotel and Resort Hotel will proportionately have similar relationships with the other variables.

```
## [1] 1.134618e-35
```

We are also particularly interested in using the presence of kids (children or babies) as a predictor of average daily rate. As an initial analysis, we perform a t-test on the the average daily rates of bookings for groups with only adults, versus groups with children or babies. This t-test assesses if there is a significant difference between the two distributions, which would indicate that the presence of kids is likely to be a useful predictor of average daily rate. In this analysis, we find that our t-test produces a p-value of approximately 0. Because this p-value is very low, we conclude that there is a statistically significant relationship between the average daily hotel rates of groups with children or babies and groups with only adults in our dataset, which provides evidence that the presence of kids (children or babies) is a useful predictor of average daily hotel rates.

2 Methodology

2.1 Data Processing

It is important to note that there are a few variables in the dataset that do not provide insight to our research question and as a result we will remove from the dataset. First is the category `hotels$country`, due to the fact that every hotel in this subset of the data is located in the U.S., this is a redundant variable. Next, the variables `hotels$reserved_room_type`, `hotels$assigned_room_type`, `hotels$agent`, `hotels$company` are categorical variables that have codes assigned to their values. However, due to the confidentiality of the customers, the classification of these codes have not been identified by those who collected the data. Therefore, we will be removing these variables from the dataset. In addition to that, we removed `hotels$reservation_status_date` and `hotels$arrival_date_week_number` as we already had variables that tracked day, month, and year. The variables `hotels$previous_cancellations` and `hotels$reservation_status` were also removed as they did not provide relevance in predicting the average daily rate.In addition to removing variables, certain variables such as `hotels$month` as well as `hotels$meal` had to be refactored and cleaned in order to function properly in the linear regression model.

```
## Warning in if (hotels$arrival_date_week_number >= 36 &
## hotels$arrival_date_week_number <= : the condition has length > 1 and only the
## first element will be used
```

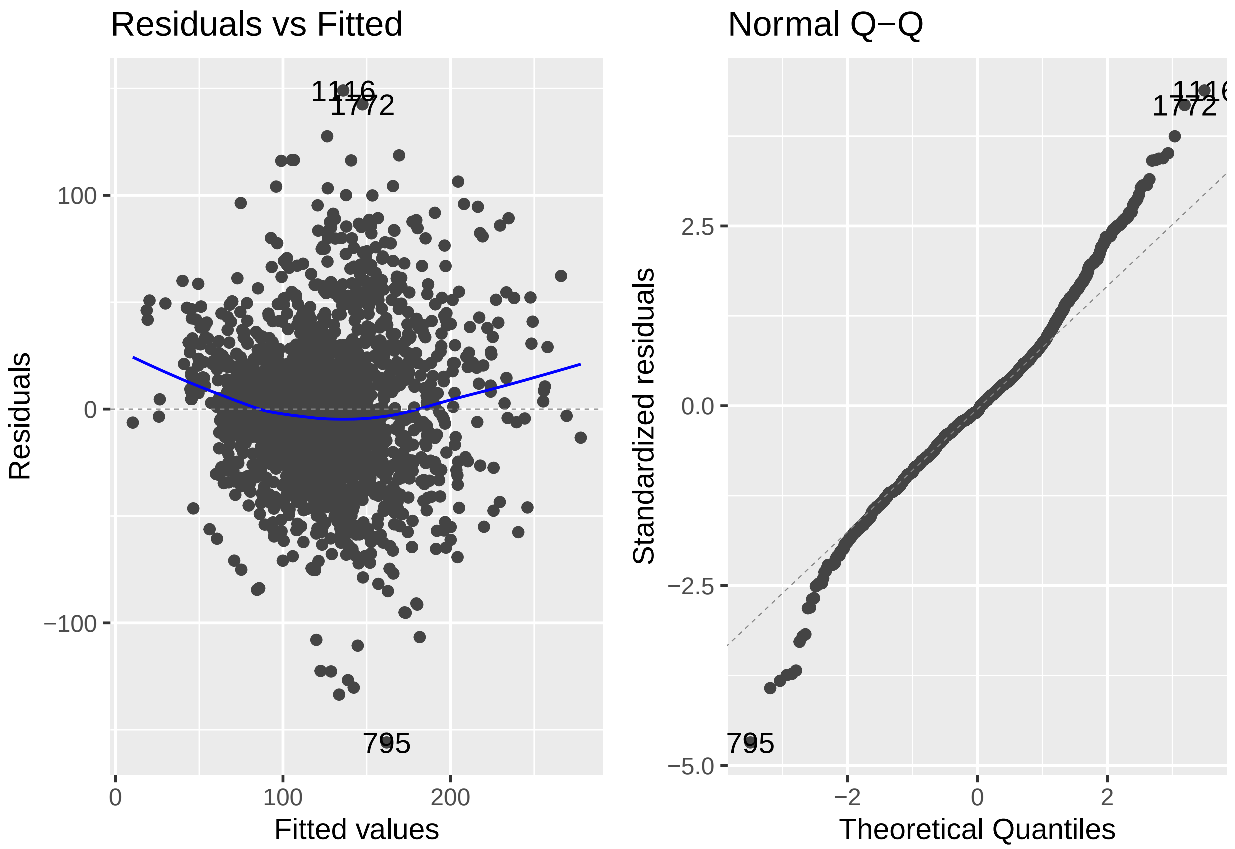
```
## Warning in if (hotels$arrival_date_week_number >= 1 &
## hotels$arrival_date_week_number <= : the condition has length > 1 and only the
## first element will be used
```

We will now fit a model using all of the variables from our cleaned dataset. In particular, we are curious in how hotel (Resort Hotel or City Hotel), stays_in_weekend_nights, stays_in_week_nights, the type of meal plan, as well as the time of the reservation as predictor variables. As we explore more of the relationships in our dataset, we will be performing selection criteria to find the best model for predicting the average daily rate of a hotel reservation. We plan to use multiple linear regression with a variety of combinations of interested predictor variables and their interaction.

2.2 Model Selection

With such a high number of predictor variables, we believed it would be best to use a backwards model with BIC as our selection criteria. With this, we are able to narrow down our predictor variables to the most significant ones that can be used to draw concrete conclusions as well as predict daily prices for different types of reservations. Additionally, we were specifically interested in how the month of the reservation and the day in the month had an affect on each other. Therefore, we added this interaction term into our model. The final model is displayed in the Results section.

2.3 Model Conditions



To check that a linear model is applicable to this dataset, we check the following conditions: Linearity, Constant Variance, Normality, and Independence.

To check for Linearity, we look at the residuals vs fitted plot. We see that the residuals are randomly scattered, which indicates that the linearity condition is met.

To check for constant variance, we look at the spread of the residuals in the residuals vs fitted plot. We see that the spread of the residuals is approximately equal as the fitted value increases, indicating that the constant variance condition is satisfied.

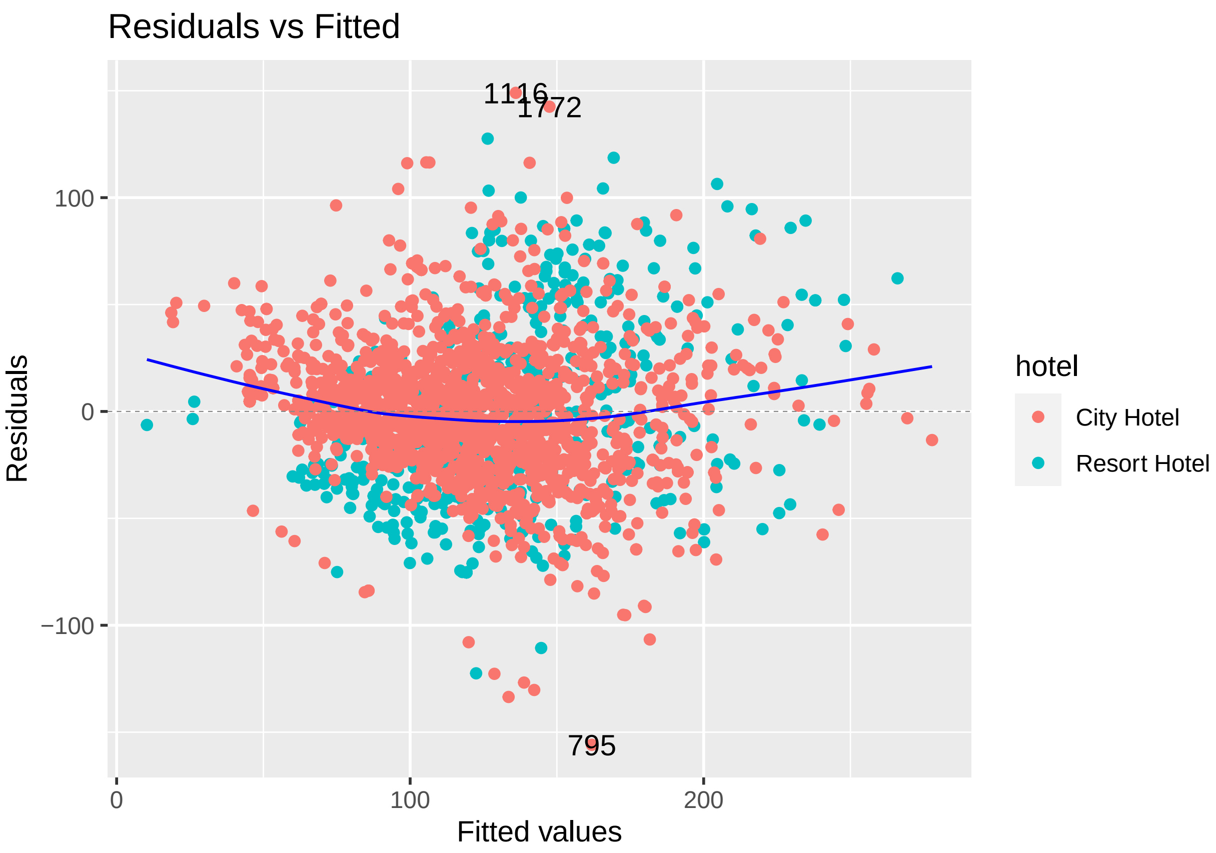
To check for normality, we look at the QQ-plot for a linear relationship. In the plot, we see a mostly linear relationship, and additionally recognize that our model is robust to deviations because our number of samples is much larger than 30. Therefore, we conclude that the normality condition is met.

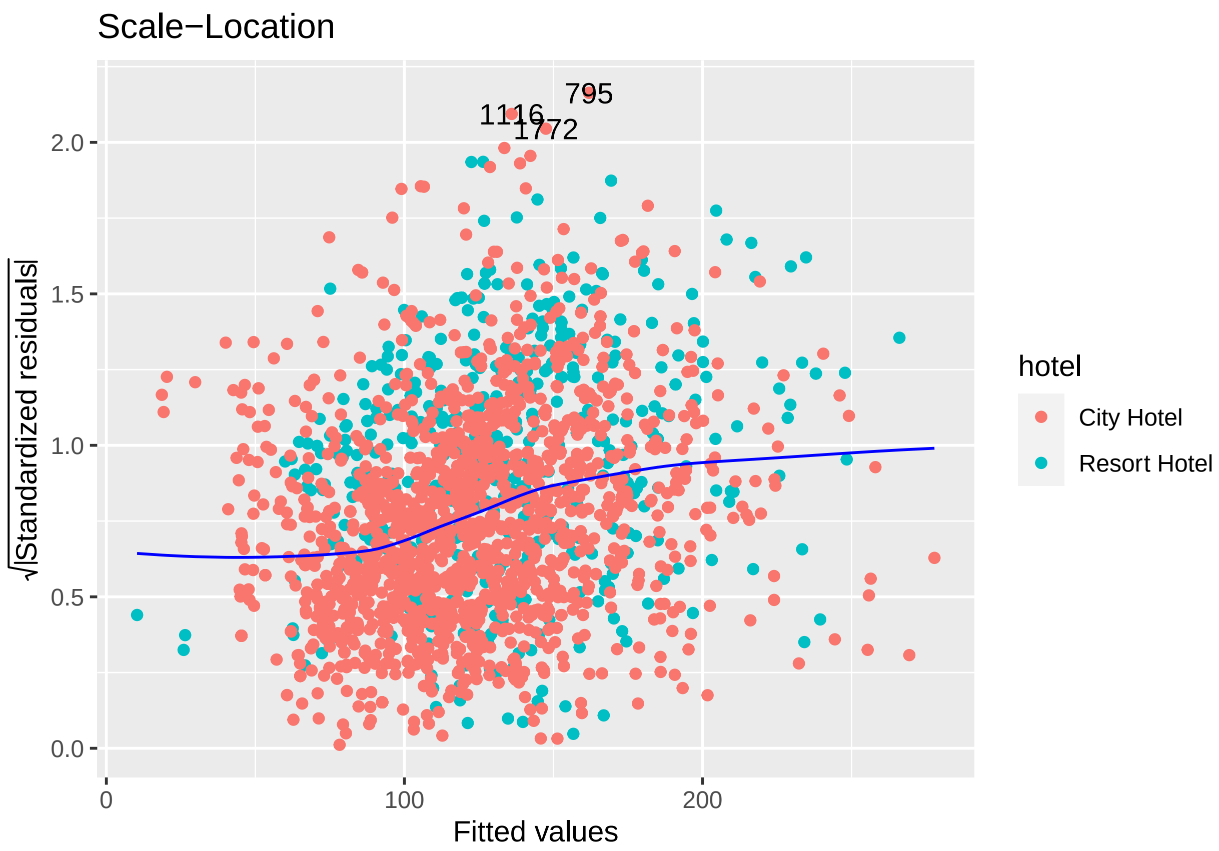
To check for independence, we look at the study design, and find that there is no reason to believe that our samples are not independent from the survey, such that individual hotel bookings can be assumed to be independent.

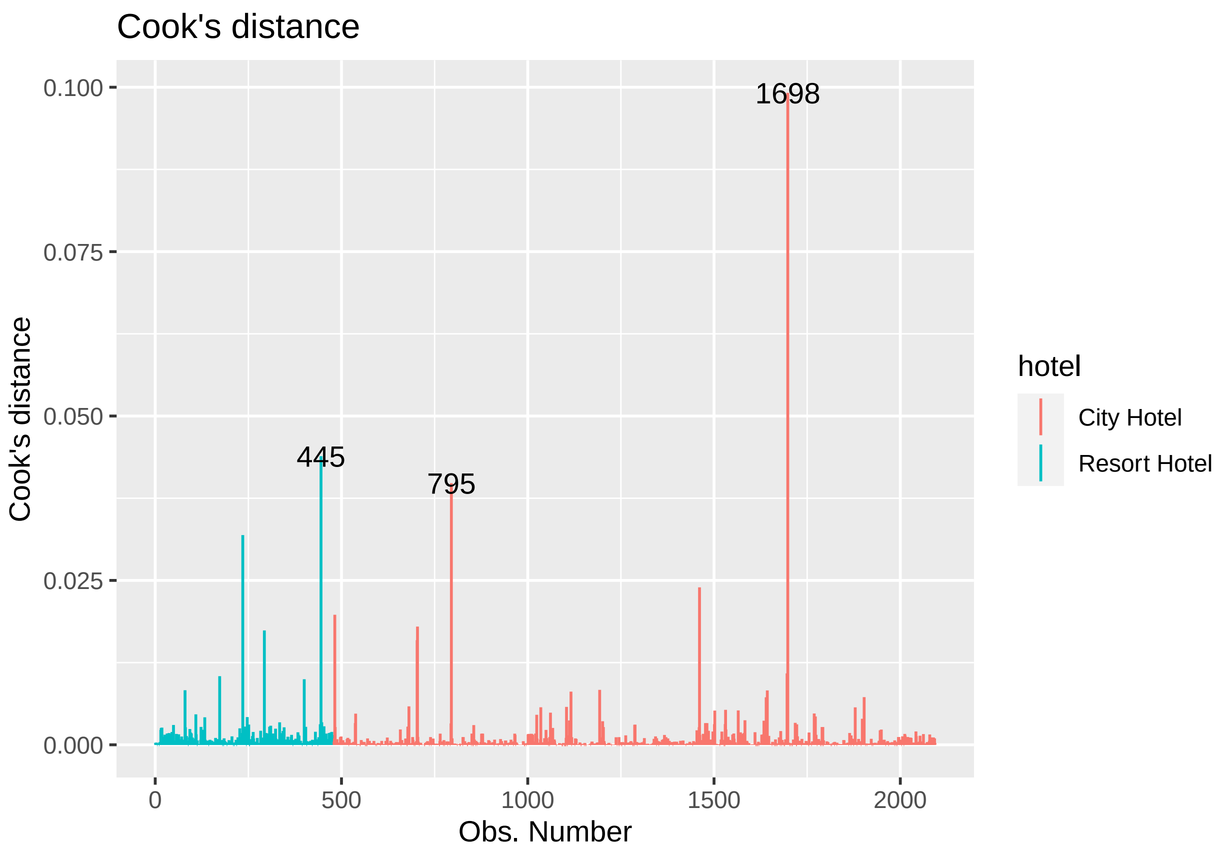
From this analysis, we conclude that the conditions are met for performing linear regression.

2.4 Model Diagnostics

Next, we evaluate model diagnostics to check for outliers and significant observations.


Residuals vs Fitted
Residuals
Fitted values
100
0
−100
0
100
200
1116
1772
795
hotel
City Hotel
Resort Hotel


Scale−Location
795
1116
1772
2.0
1.5
1.0
0.5
0.0
√|Standardized residuals|
0
100
200
Fitted values
hotel
City Hotel
Resort Hotel



We use leverage and Cook's Distance to identify influential observations in our dataset, and use standardized residuals to identify outliers.

Leverage is a measure of the distance between an observation's values of the predictor variables and the average values of the predictor variables for the entire data set. We define a high leverage point as having a leverage greater than $ 2(p+1) /n $, where p is the number of predictors and n is the number of observations. We find 448 observations to be high leverage, and consider them to be potential influential points.

Standardized residuals can be used to identify potential outliers, as observations that have standardized residuals of large magnitude don't fit the pattern determined by the regression model. We identify potential outliers as observations with standardized residuals with a magnitude greater than or equal to 3. We find 22 observations to be potential outliers.

Cook's distance is a composite measure of an observation's leverage and standardized residual, and is used to identify influential points. An observation is considered a moderately influential point if it's Cook's distance is greater than 0.5. We find 1 observation with a Cook's distance greater than 0.5. Therefore, we can conclude that there is 1 moderately influential point.

After calculating the model diagnostics, we have found that one observation can be classified as an outlier. However, due to the fact that this is a legitimate observation of our dataset, we have chosen to keep this point in our model.

3 Results

3.1 Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-11.392	9.667	-1.178	0.239	-30.350	7.567
hotelResort Hotel	-5.430	1.941	-2.797	0.005	-9.237	-1.623
is_canceled	9.622	1.915	5.024	0.000	5.866	13.377
lead_time	-0.177	0.010	-17.600	0.000	-0.197	-0.158
year2016	15.462	3.042	5.083	0.000	9.496	21.427
year2017	34.581	3.448	10.030	0.000	27.820	41.343
monthFebruary	0.906	10.663	0.085	0.932	-20.004	21.816
monthMarch	3.907	8.969	0.436	0.663	-13.683	21.497
monthApril	30.603	8.476	3.611	0.000	13.980	47.226
monthMay	52.286	7.725	6.768	0.000	37.136	67.436
monthJune	48.332	7.552	6.400	0.000	33.522	63.142
monthJuly	46.199	7.548	6.121	0.000	31.397	61.002
monthAugust	77.880	7.823	9.956	0.000	62.539	93.221
monthSeptember	51.447	8.161	6.304	0.000	35.442	67.453
monthOctober	62.318	8.492	7.339	0.000	45.664	78.971
monthNovember	48.329	10.033	4.817	0.000	28.654	68.005
monthDecember	-1.419	10.900	-0.130	0.896	-22.795	19.957
day_of_month	-0.562	0.446	-1.261	0.208	-1.435	0.312
stays_in_week_nights	1.513	0.521	2.903	0.004	0.491	2.536
adults	21.456	1.537	13.959	0.000	18.441	24.470
children	32.022	1.451	22.065	0.000	29.176	34.868
mealBB	21.739	1.994	10.904	0.000	17.830	25.649
mealHB	50.500	5.092	9.917	0.000	40.513	60.486
distribution_channelDirect	17.531	6.195	2.830	0.005	5.383	29.679
distribution_channelGDS	30.240	9.591	3.153	0.002	11.431	49.050
distribution_channelTA/TO	5.200	6.042	0.861	0.390	-6.650	17.050
is_repeated_guest	-53.217	9.627	-5.528	0.000	-72.097	-34.338
previous_bookings_not_canceled	45.803	11.550	3.966	0.000	23.152	68.454
days_in_waiting_list	-0.343	0.077	-4.436	0.000	-0.495	-0.192
total_of_special_requests	4.951	0.952	5.202	0.000	3.085	6.818
monthFebruary:day_of_month	0.088	0.671	0.131	0.896	-1.229	1.405
monthMarch:day_of_month	1.410	0.575	2.451	0.014	0.282	2.539
monthApril:day_of_month	0.965	0.533	1.809	0.071	-0.081	2.011
monthMay:day_of_month	0.464	0.505	0.919	0.358	-0.526	1.455
monthJune:day_of_month	0.577	0.500	1.154	0.249	-0.403	1.558
monthJuly:day_of_month	1.712	0.500	3.423	0.001	0.731	2.692
monthAugust:day_of_month	0.621	0.504	1.233	0.218	-0.367	1.610
monthSeptember:day_of_month	1.309	0.526	2.489	0.013	0.278	2.340
monthOctober:day_of_month	-0.583	0.569	-1.024	0.306	-1.699	0.534
monthNovember:day_of_month	-0.691	0.740	-0.934	0.350	-2.142	0.760
monthDecember:day_of_month	1.825	0.605	3.015	0.003	0.638	3.012

3.2 Interpretation

The model provides a number of useful predictors for determining how the average daily rate or price of a hotel is affected by its customization. The first aspect when choosing a hotel may be the type of hotel one would like to visit. In this model we looked at the difference between staying in a hotel located in a city/urban setting versus staying in a resort style hotel. According the model, the coefficient for a resort hotel is -5.430. This can be interpreted as if a guest decides to stay in a resort hotel versus a city hotel, they are expected to pay on average $5.43 less per day, holding all else constant. Because the p-value is very close to 0, we can also conclude that this is a significant statistic. Therefore, when deciding on a travel destination, it would be beneficial to consider the difference in price of a city or resort hotel.

Another important factor when planning a trip or booking a hotel is when you would like to travel or stay. The model covers a variety of predictors that involve the timing of a booking. For example, when considering which month to travel in, there is a possibility that depending on the month you book your stay, the average daily rate may be more expensive or less expensive. For example, the model identifies that the arrival date of a stay in August has a coefficient of 77.880. This can be interpreted as if a guest decides to book a stay during the month of August, they would be expected to pay on average $77.88 more per day, holding all else constant. Additionally, according to the model, if a guest stayed in a hotel with an arrival date in April, May, June, July, August, September, October, or November, they would be expected to pay on average a higher daily rate than staying in the month of January. This is proved by the fact that the p-value for each coefficient of these months is approximately 0, and the 95% confidence interval does not include 0 and has both thresholds being positive values. On the other hand, the months of February, March, and December all have fairly large p-values and have 0 fall in the 95% confidence interval. Therefore, we can say that these months are not statistically different from the month of January, meaning that there will not be a significant difference in the average daily rate whether a guest stays in January or February, March, or December.

Not only is the month of stay important, but the day within a month may also be an important factor in the price of a hotel stay. The model gives a coefficient of -0.562 for the arrival date day of the month, but has a p-value of 0.208. Because this p-value is larger than the $\alpha = 0.05$ level, we can say that the date of the month is not statistically significant. This is further proved by the 95% confidence interval (-1.435, 0.312), where 0 falls in the interval. Therefore, the arrival date in relation to the day of the month is not a useful predictor for the average daily rate of a hotel stay.

Additionally, when looking at the interaction between the day of the month and the month, we can see that maybe staying at a certain time within a certain month may have an effect in the cost. The months of March, July, September, and December all had p-values less than $\alpha = 0.05$ level, so we can conclude that they are statistically significant. This means that the effect of Month of arrival date on the average daily rate is statistically different for July, September, and December.

Another attribute that may effect the daily price of the hotel is the number of guests and more specifically, the number of adults in the reservation. The model gives a coefficient of 21.456 and a p-value of 0 showing this coefficient is statistically significant. Therefore, we can conclude that for every additional adult added to the reservation, the average daily price of a hotel will increase on average $21.46, holding all else constant. From this, a pretty good estimate can be made about how much the daily rate will change with the number of adults under the reservation. This can similarly be applied to the number of children staying under a reservation. For every additional child added to the reservation, the average daily price of a hotel will increase on average $32.02, holding all else constant. We can be confident in this interpretation as the p-value is approximately 0 and 0 does not fall in the 95% confidence interval. Therefore, more children under a reservation will result in an increase of the average daily rate.

Lastly, we will discuss how a meal plan can possibly change the average daily price of a hotel stay. In the variable meal, the baseline is if a guest were to get no meal plan. The coefficient for Bed and Breakfast has a coefficient of 21.739 and a p-value of approximately 0. Therefore, we can say that if one were to choose the meal plan Bed and Breakfast, they would pay on average $21.74 more than if they did not buy a meal plan. Similarly for a Half Board meal plan (breakfast and one other meal), the guest is expected to pay on average $50.50 more per day for their hotel stay if they were to get the Half Board plan versus no meal plan. This interpretation is supported by the approximately 0 p-value for this coefficient.

3.3 Prediction

In addition to discussing how the different predictors affect the average daily rate, one can also use the model to predict how much their hotel stay may cost in relation to these factors. For example, we can predict the average daily rate for a guest who had a reservation at a Resort Hotel that was not canceled, had 90 days elapsed between the entering date of the booking into the PMS and their arrival date, had an arrival date of February 26, 2016, stayed for 3 weekend nights, consisted of 2 adults and 3 children, got a Half Board meal plan, booked directly from the hotel, is not a repeated guest, has not had previous bookings canceled, had 0 days in the wait list and had a total of 4 special requests.

fit	lwr	upr
210.401	196.489	224.313

Based on the above table, we are 95% confident that the average daily rate for this guest will fall somewhere between $196.49 and $224.31. Therefore, we can predict that this guest's reservation will have an average daily rate of about $210.40. This model can be applied to a variety of new bookings and reservations to predict how much one's hotel stay would cost.

.metric	.estimator	.estimate
rmse	standard	33.767

.metric	.estimator	.estimate
rmse	standard	34.714

To determine the accuracy of the predictive ability of our model, we use the Root Mean Squared Error on both our training and test datasets below. We find a very similar error between our training and test sets. For the training set, the model had an RMSE of $33.767 and the testing set had an RMSE of $34.714. Given that the RMSEs for the training and test set are very close, we can conclude that the model is generalizable based on these values. However, due to the fact that the RMSE is approximately $33-35, we would argue that there is a limitation as to what can be predicted from different data using our model.

4 Discussion and Conclusion

Our model assumes that hotel bookings are independent from each other, which could potentially be broken in limited circumstances, such as competing hotel holiday promotions.

Our analysis shows that purchase of a meal plan, booking in August or October, and having previous bookings that were not cancelled were the most correlated with larger increases in daily hotel rate, and that being a repeated guest, booking in a Resort hotel instead of a City hotel, and booking in December were most correlated with decreases in daily hotel rate. We were also able to show that our model generalized to new data,showing that the error rate on a training and test set are similar (around $33). This means that our model could reasonably be used to predict hotel rates in the U.S., with the above limitations in mind.

It's important to recognize that our model has several limitations. First, we recognize that the travel industry (including hotel booking) is sensitive to tragic circumstances or disasters, which are not accounted for in our model. For example, we recognize that our model is not likely to characterize hotel bookings during COVID, given that our data is limited to the years 2015 to 2017 (before COVID was in the U.S.). Additionally, we anticipate that the hotel room type (such as suite versus individual room) would be an important predictor of hotel room price, but is not included in our model, as described in our methodologies section. Additionally, in our exploratory analysis we found that there were far more hotel bookings in cities than in resorts. Given this skew, our model may better characterize city bookings than resort bookings.

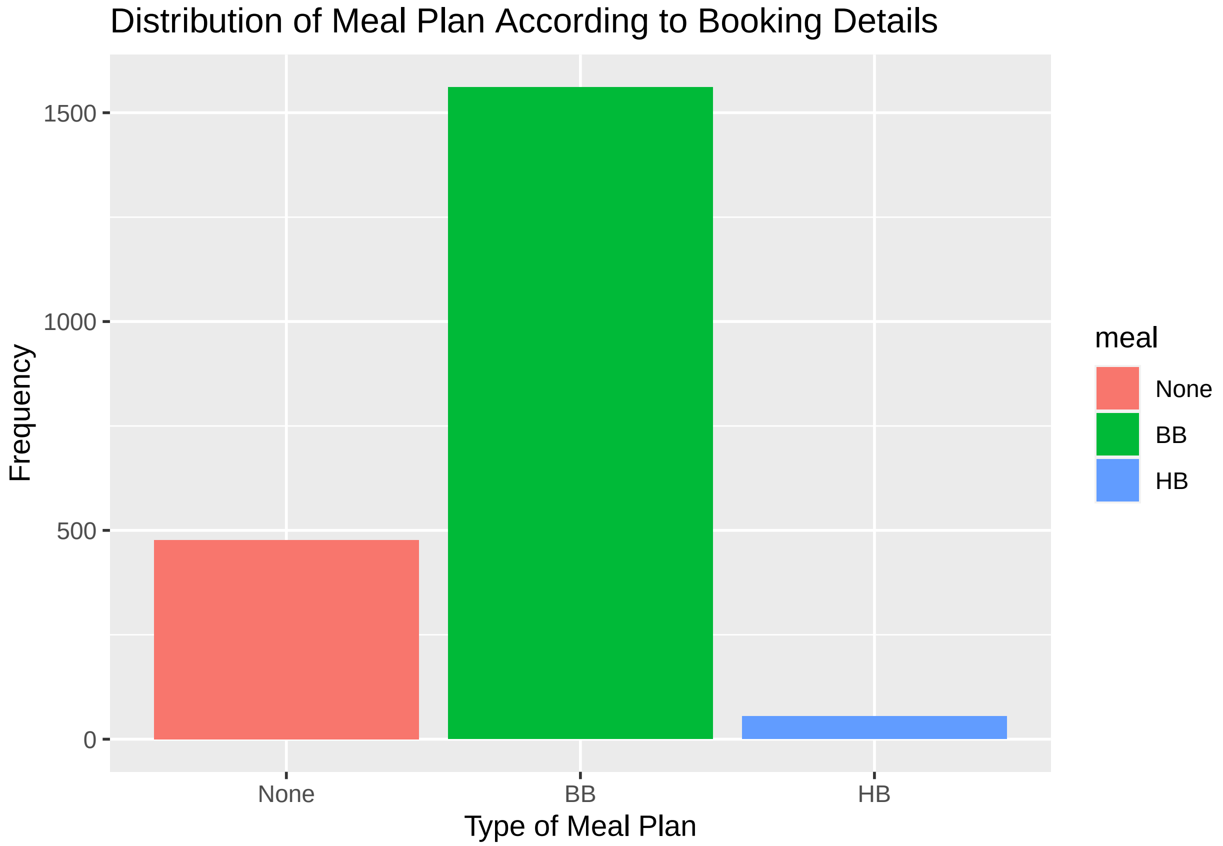
For future work, we are interested in expanding our analysis to include data from other countries, where we could then draw comparisons and differences in the average daily rate across a wide range of areas. Additionally, because we are all personally interested in traveling around the world, we hope to expand our model to not only be the average daily rate of a hotel, but possibly have the response variable be the average daily rate of a trip. We believe this we could then incorporate more predictors such as transportation cost, food cost, etc. that will add interesting insights to our interpretations and conclusions.

5 References

Antonio, Nuno, Ana de Almeida, and Luis Nunes. 2019. "Hotel Booking Demand Datasets." *Data in Brief* 22: 41–49. https://doi.org/https://doi.org/10.1016/j.dib.2018.11.126.

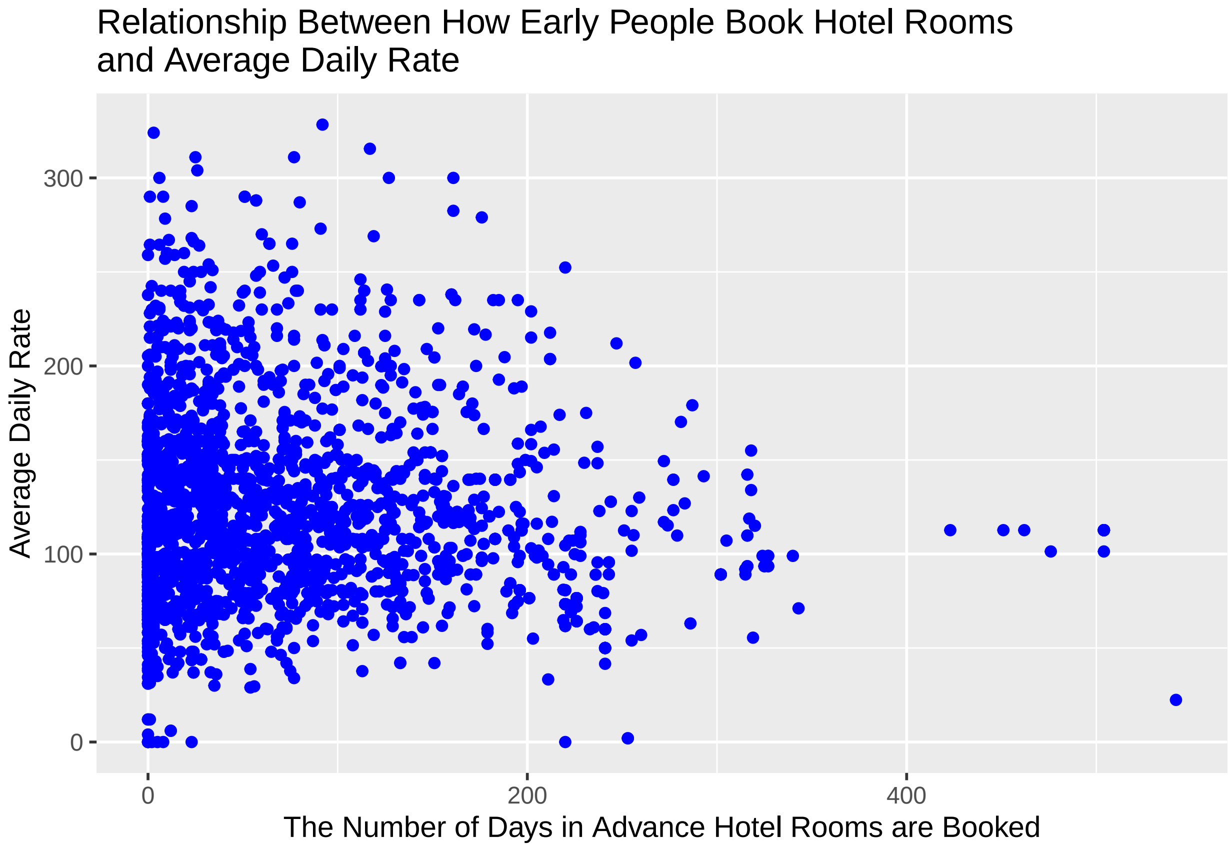
6 Appendix

6.1 Exploratory Data Analysis



meal	n	mean	sd
None	477	102.530	30.013
BB	1561	127.998	54.009
HB	55	167.216	60.824

We can see that majority of customer bookings chose the Bed and Breakfast plan (BB), rather than Half Board plan or Full Board (HB and SC). While this does reflect the choice of those booking the hotel, this visualization and dataset can also be influenced by what type of meal plan each hotel is offering.



We can see that the more last minute visitors book a hotel, the more likely it is that the price in the US dollars varies. Besides, we can also see that the highest price in a hotel decreases as the difference between entering data and arrival date increases. This could suggest that the earlier you book a hotel, the price is more likely to be cheap.

6.2 Full Model

term	estimate	std.error	statistic	p.value
(Intercept)	-18.084	35.481	-0.510	0.610
hotelResort Hotel	-7.763	2.076	-3.739	0.000
is_canceled	8.217	2.011	4.086	0.000
lead_time	-0.170	0.011	-15.692	0.000
year2016	15.902	3.186	4.991	0.000
year2017	34.131	3.580	9.534	0.000
monthFebruary	-0.427	5.498	-0.078	0.938
monthMarch	22.355	4.645	4.813	0.000
monthApril	41.858	4.593	9.113	0.000
monthMay	55.480	4.385	12.652	0.000
monthJune	52.970	4.345	12.190	0.000
monthJuly	68.106	4.443	15.330	0.000
monthAugust	85.096	4.384	19.410	0.000
monthSeptember	68.097	4.780	14.246	0.000
monthOctober	51.333	4.935	10.401	0.000
monthNovember	40.892	5.860	6.978	0.000
monthDecember	27.386	5.636	4.860	0.000
day_of_month	0.263	0.088	2.996	0.003

term	estimate	std.error	statistic	p.value
stays_in_weekend_nights	-0.121	0.910	-0.133	0.895
stays_in_week_nights	1.775	0.529	3.352	0.001
adults	21.149	1.566	13.506	0.000
children	43.841	4.406	9.951	0.000
babies	15.422	15.949	0.967	0.334
mealBB	23.057	2.039	11.306	0.000
mealHB	52.188	5.145	10.144	0.000
market_segmentComplementary	44.392	52.825	0.840	0.401
market_segmentCorporate	0.609	34.626	0.018	0.986
market_segmentDirect	50.348	38.969	1.292	0.197
market_segmentGroups	10.706	37.495	0.286	0.775
market_segmentOffline TA/TO	0.324	37.400	0.009	0.993
market_segmentOnline TA	21.389	37.387	0.572	0.567
distribution_channelDirect	-26.418	19.030	-1.388	0.165
distribution_channelGDS	30.275	17.092	1.771	0.077
distribution_channelTA/TO	-8.721	15.489	-0.563	0.573
is_repeated_guest	-54.187	9.633	-5.625	0.000
previous_bookings_not_canceled	43.828	11.635	3.767	0.000
booking_changes	1.569	0.947	1.656	0.098
deposit_typeRefundable	-30.109	34.010	-0.885	0.376
days_in_waiting_list	-0.283	0.082	-3.434	0.001
customer_typeGroup	-11.439	12.592	-0.908	0.364
customer_typeTransient	-9.109	8.369	-1.088	0.277
customer_typeTransient-Party	-0.839	8.980	-0.093	0.926
required_car_parking_spaces	6.333	3.255	1.946	0.052
total_of_special_requests	3.950	0.975	4.051	0.000
children_present1	-23.500	7.647	-3.073	0.002

The original full model with all variables as predictors is printed above.