Hotel Bookings Analysis

Camille van der Watt, Matthew Wong, Timmy Li, Shreya Kumar

16th May 2023

1 Abstract

Vacations are some of the most fundamental activities for most people. They represent a break from the monotonous routine of their work days, helping with both their physical and mental health. However, not everyone makes use of their vacation days: CNBC reports that only 54% of people's vacation days are used up. This leaves a massive amount of time for relaxation wasted. This is likely due to the fact that the luxury of vacation is accompanied by luxury prices that compound from travel to and from the destination, lodging, dining, activities and excursions, souvenir shopping, and more. Therefore, this report analyzes what determines the nightly rate of hotel rooms, as well as what gives incentives to customers to return to hotels they have stayed at previously.

The data used in this report came from the Kaggle Dataset "Hotel Booking Demand," which was sourced from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The dataset contains over 119,000 records for bookings made at city and resort hotels from people in over 170 countries across the world in the years 2015-2017. The records include fields relevant to many different areas of interest, such as when and how the booking was made, the start and end dates of the reservation, the composition of the group of people staying, the type and nightly price rate of the room booked, and any changes, special requests, or cancellations made by the guests.

2 Assumptions and Limitations

Despite the global breadth of this dataset, these results cannot be generalized to all hotels. This dataset only classifies each booking record's hotel as a "city hotel" or a "resort hotel." Not only are these two categories insufficient to encompass all hotels, but they are also too broad to capture the high levels of nuance and variation between the hotels that do fall within these categories. Broadening the analysis to more types of hotels or specifying which sector of the hospitality industry these hotels fall within could provide better-informed insight. Additionally, these records are for hotel stays occurring between 2015-2017–notably years before the global COVID-19 Pandemic disrupted all travel and tourism, leaving such industries largely altered through the time that this report was written. A recreation of this sample drawn from bookings occurring in the year 2020 or later would likely yield very different observations.

The creator of the dataset implemented the measure of average nightly rate in terms of the profit made by the hotel per night of the reservation, rather than in units of currency. This was to eliminate the complications related to the many different currencies used in the over 170 countries appearing in records in this database. Additionally, the analysis of the relationship between holidays and other variables in this report is limited to United States federal holidays for the sake of simplicity. This implies the assumption that all the countries with records in this dataset are affected by approximately the same holiday seasons. Additionally, the dataset was filtered to remove one extreme outlier with average nightly rate > 5000, and other anomalous records with average nightly rate < 1, in order to train and test our models on typical data.

3 Methods

Three of the four models included in this report utilized training and testing data sets generated by the test_train_split function where 70% of the data was used for training and 30% was used for testing, in order to avoid overfitting. The random_state parameter sets the seed for the random number generator used in the splitting process so that the results could be reproduced. The model was fitted to the training data and later validated using the testing data. The model for Question 3 did not utilize a train-test split because the data for the model was grouped, which greatly reduced the number of observations to train the model with. Three of the four models do not include constants as explanatory variables. This is because dates and lengths of hotel stays are focal explanatory variables in those models, so the constant would represent the predicted outcome value for hotel stays occurring on the “0th day” of the year or stays for 0 days, which does not make sense in context.

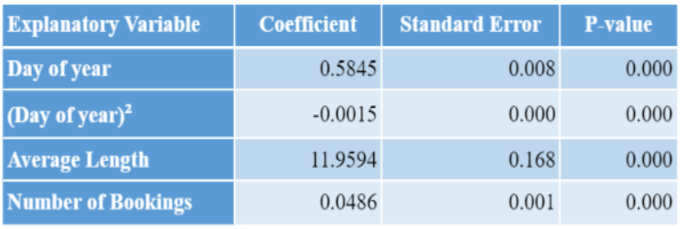
4 Question 1: Have you ever wondered when the best time of year to book a hotel room is?

We were interested in seeing how an interested customer could strategically plan the timing of their vacation in order to minimize the nightly rate they pay for their room, as rates fluctuate seasonally. This fluctuation is due to a basic principle of economics: suppliers will set prices high when demand is high, and they will set prices low when demand is low. Since demand fluctuates seasonally due to human behaviors, so do prices.

In order to isolate the effect on the daily rate of just changing the arrival date of a booking, we controlled for demand on the start date of each reservation. The measures of demand which we utilized were the average length of bookings and the total number of bookings on the day of the year on which the reservation started. Also due to this cyclic fluctuation, we modeled a quadratic polynomial relationship between nightly rate and the number day of the year on which the reservation started. We proposed the following model:

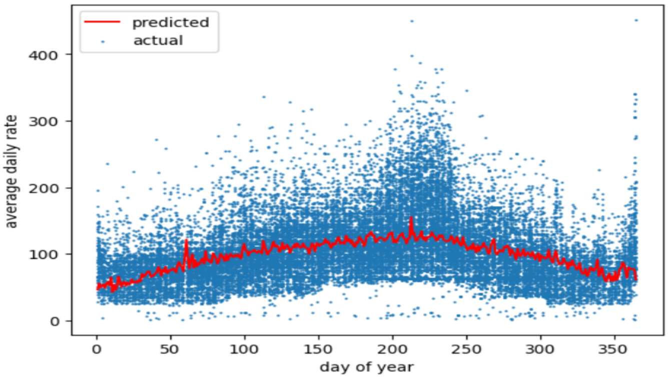
$$rate_i = \beta_1 \cdot (day) + \beta_2 \cdot (day) + \beta_3 \cdot (avglength) + \beta_4 \cdot (bookings) + \epsilon$$

An OLS regression using this model produced the following results:



Explanatory Variable	Coefficient	Standard Error	P-value
Day of year	0.5845	0.008	0.000
(Day of year)^2	-0.0015	0.000	0.000
Average Length	11.9594	0.168	0.000
Number of Bookings	0.0486	0.001	0.000

All of the explanatory variables have p-values less than 0.05, which means there is evidence to support the hypotheses that each of them individually are related to the dependent variable of average daily rate. The positive coefficient on day of year in conjunction with the negative coefficient on day of year squared indicates that on average, over the course of a year, prices first increase and then later decrease over time. The variable representing the average length of all bookings which start on the same day as the given observation has a relatively high positive coefficient of 11.96. This means that a 1-day increase in average length of stay increases the average daily rate by 11.96 units on average, which is a strong effect. The number of bookings does not have as high of a coefficient, but its positive effect on average daily rate is still statistically significant. Our two measures of demand, average length of stay and number of bookings, have statistically significant positive relationships with price, which is consistent with the economic principle that inspired the design of this model.

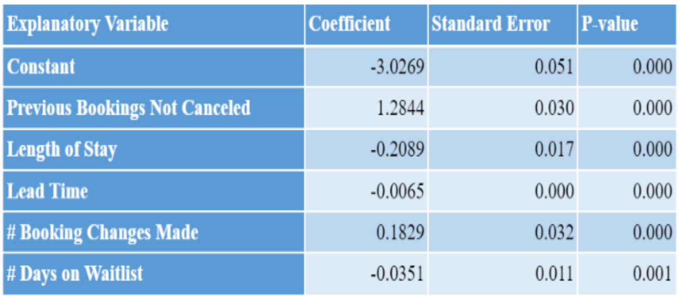


5 Question 2: What incentivizes customers to be repeat customers or what do repeat customers have in common?

We created a predictive model using logistic regression in order to show the relationship between several predictor variables and the likelihood that a customer is a repeated guest in a hotel. The predictor variables are: the previous_bookings_not_canceled, length_of_stay, lead_time, booking_changes, and days_in_waiting_list whereas the dependent variable is is_repeated_guest. We proposed the following model:

$$
\begin{aligned}
D_i^{repeat} &= \beta_0 + \beta_1 \cdot (\#prev.\ bookings) + \beta_3 \cdot (length) + \beta_4 \cdot (lead\ time) \\
&+ \beta_5 \cdot (\#changes) + \beta_6 \cdot (\#days\ on\ waitlist) + \epsilon
\end{aligned}
$$

A logistic regression using this model produced the following results:



Explanatory Variable	Coefficient	Standard Error	P-value
Constant	-3.0269	0.051	0.000
Previous Bookings Not Canceled	1.2844	0.030	0.000
Length of Stay	-0.2089	0.017	0.000
Lead Time	-0.0065	0.000	0.000
# Booking Changes Made	0.1829	0.032	0.000
# Days on Waitlist	-0.0351	0.011	0.001

The first predictive variable, previous_bookings_not_canceled, has a coefficient value of 1.2844 and a p-value of 0. Holding all other variables constant, for each additional previous booking that wasn't canceled, the log odds of being a repeated guest increase by 1.2844. This positive relationship between previous_bookings_not_canceled and is_repeated_guest is statistically significant and is unlikely due to chance since the p-value is less than 0.05.

The second predictive variable, length_of_stay, has a coefficient value of -0.2089 and a p-value of 3.348e-34. Holding all other variables constant, for each additional day of stay, the log odds of being a repeated guest decrease by 0.2089. This negative relationship between length_of_stay and is_repeated_guest is statistically significant and is unlikely due to chance since the p-value is less than 0.05.

The third predictive variable, lead_time, has a coefficient value of -0.0065 and a p-value of 8.325e-52. Holding all other variables constant, for each additional day of lead time (the number of days between booking and arrival date), the log odds of being a repeated guest decrease by 0.0065. This negative relationship between lead_time and is_repeated_guest is statistically significant and is unlikely due to chance since the p-value is less than 0.05.

The fourth predictive variable, booking_changes, has a coefficient value of 0.1829 and a p-value of 9.434e-09. Holding all other variables constant, for each additional change in the booking, the log odds of being a repeated increase by 0.1829. This positive relationship between booking_changes and is_repeated_guest is statistically significant and is unlikely due to chance since the p-value is less than 0.05.

The fifth predictive variable, days_in_waiting_list, has a coefficient value of -0.0351 and a p-value of 1.046e-03. Holding all other variables constant, for each additional day spent on the waiting list, the log odds of being a repeated guest decrease by 0.0351. This negative relationship between days_in_waiting_list and is_repeated_guest is statistically significant and is unlikely due to chance since the p-value is less than 0.05.

Although the relationships between each predictive model and is_repeated_guest is statistically significant, based on the magnitudes of the coefficients, it seems that the variable previous_bookings_not_canceled has the highest impact on is_repeated_guest, followed by both length_of_stay and booking_changes, then with days_in_waiting_list and finally with lead_time. The results based on this specific dataset seems to show that guests who tend to not cancel previous bookings, stay for short periods, book close to their arrival date, make more booking changes, and spend little time on a hotel's waiting list tend to be frequent, repeated guests. Repeated guests for certain hotels tend to be content with these hotels' services and are unlikely to cancel existing bookings they make. The short stays of many repeated guests can be attributed to the fact that repeated guests can be customers on work trips or on short weekend trips at nearby hotels, which are usually short and frequent. Bookings for short trips are more easily changeable and less expensive compared to bookings for trips on rare, special occasions, which is why repeated customers may tend to change bookings more often than non-frequent guests. Guests who frequent certain hotels often are familiar with the booking process and thus don't feel the need to book a trip months in advance (especially if it's a short, normal work trip and not on a special occasion). Finally, guests who spend a long time on a hotel's waiting list may become dissatisfied with the hotel's service and feel disinclined to visit the hotel a second time.

We used the predict method to generate predictions from the logistic regression model using the testing data. Next we use the accuracy_score function to compare predicted outcomes with the actual outcomes from the testing data. When we validated the model using the testing data, the logistic regression model achieved an accuracy score of 0.98, which indicates that the model is very good at predicting whether a guest is a repeated guest based on the predictor variables used in the model. Although it is common practice to determine the model's ability to generalize to new, unseen data by using testing data from the same dataset as the training data, a way we could improve validating the model in the future is to use independent test data. If the test data is drawn from the same dataset as the training data, there may exist certain patterns in both the training and testing data that the model becomes accustomed to when fitted to the training data. This means if the model was exposed to a new dataset without the same patterns, it may not be able to predict as accurately or perform as well as it did when the testing and training data were from the same dataset.

6 Question 3: What is the optimal length of stay in order to get the best daily rate?

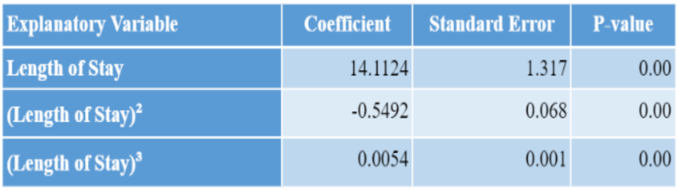
Another question our group brainstormed while initially analyzing the data set is the optimal length of stay to get the best daily rate price. We aimed to find a correlation between the length of stay at a hotel and the price the customer pays. We used linear regression to create the prediction model to show the relationship between the length of stay and the average daily rate for each night at the hotel during this time period. The predictor variable is the average daily price rate stored in the series "adr3", and the explanatory variables are length of stay and the length squared. These two variables are stored in the "expvars3" DataFrame, which we will explain in the following paragraph.

We will first explain how we obtained the following data. To start, we grouped the hotelbookings DataFrame by the column "length_of_stay" and calculated three aggregate values for each group: the sum of stays_in_week_nights, the sum of stays_in_weekend_nights, and the mean of adr (average daily rate), which we added to a new DataFrame "bylength."

From there, we created a pandas DataFrame called "expvars3" and added a column called "length" that contains the values from the first column of the "bylength" DataFrame. This column was used as one of the explanatory variables in our linear regression model. We also added a column called "length_sqrd" to the expvars3 DataFrame that contains the squared values of the length column. "length_sqrd" was used as another explanatory variable in the linear regression model. We then created a pandas Series called "adr3" that contains the values from the adr column of "bylength," which we used as the explanatory variable. We proposed the following model:

$$rate_i = \beta_0 + \beta_1 \cdot (length) + \beta_2 \cdot (length^2) + \beta_3 \cdot (length^3) + \epsilon$$

An OLS Regression using this model produced the following results:



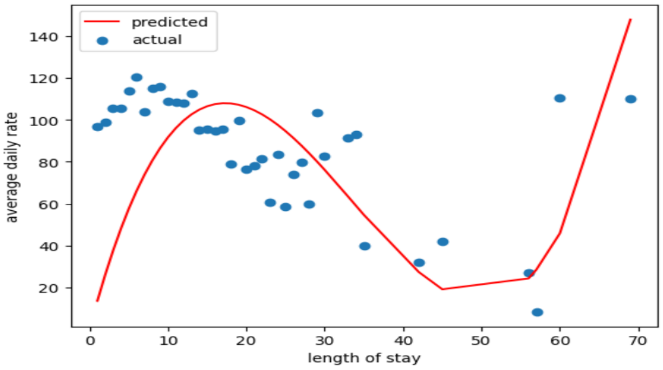
Explanatory Variable	Coefficient	Standard Error	P-value
Length of Stay	14.1124	1.317	0.00
(Length of Stay)^2	-0.5492	0.068	0.00
(Length of Stay)^3	0.0054	0.001	0.00

The first explanatory variable, length of stay, has a coefficient value of 14.1124 and a p-value of 0.000. The coefficient of 14.1124 for the length of stay variable tells us that, on average, the predicted average daily rate increases by 14.1124 units (average rate of profit per room) for every one unit increase in the length of stay, while holding all other variables constant.

The second explanatory variable, length squared, has a coefficient value of -0.5492 and a p-value of 0.00. The coefficient of -0.5492 for the length squared variable indicates that, on average, the predicted average daily rate decreases by 0.5492 units (average rate of profit per room) for every one-unit increase in the squared length of stay, holding all other variables constant.

Finally, the third variable, length cubed, has a coefficient value of 0.0054 and a p-value of 0.00. This means that the average daily rate increases by 0.0054 units (average rate of profit per room) for every one-unit increase in the cubed length of the stay, holding all other variables constant.

Jointly, these coefficients show that on average, over the course of a year, prices first increase with time, then decrease, then increase again. Since all of the variables have p-values of 0.00, there is strong evidence against the null hypothesis that there is no relationship between each of the explanatory variables and the response variable. In other words, the data shows that the length of stay, squared length of stay, and cubed length of stay are each strong predictors of the average daily rate, controlling for the effect of the other explanatory variables.



The above graph is used to evaluate the performance of the linear regression model on the data by visualizing the predicted values against the actual values. The scatter plot created by the code shows the predicted average daily rates and actual average daily rates for different values of the length of stay. The red line graph represents the predicted average daily rates for the data points, based on the linear regression model. The blue dots represent the actual average daily rates for the data points. The graph shows that the predicted values generally follow the same trend as the actual values, which suggests that the model is making reasonable predictions, but there is still some unexplained variability in the data that the model is not capturing.

7 Question 4: How strongly do holidays affect hotel prices?

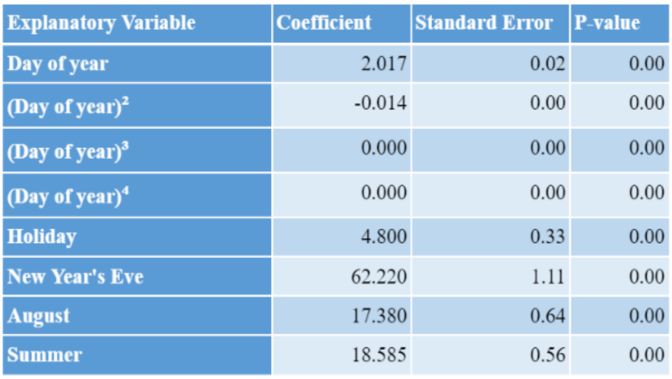
The last question we wanted to answer was how holidays affect hotel prices so that people can avoid certain times of the year to travel. In order for our model to take into account certain times of the year, we decided to create four indicator variables for each of the following times: federal holidays, New Year's Eve, August, and the summer in general. These were decided upon by looking at the previous visualization for the question. Each of the four times seemed to have a trend that differed from the normal curve.

In addition to the indicator variables, we also decided to use a cubic regression on the graph for the general curse; outside of the times outlined above. The reason why we chose to use a quartic regression, as opposed to a quadratic regression was that the graph seemed to curve upwards very slightly towards the beginning and end of the year. Since quartic polynomials form a kind of 'w' shape, we thought that this curve would better account for the trends we saw in the graph

Putting these two factors together, we proposed the following model:

$$rate_i = \beta_1 \cdot (day) + \beta_2 \cdot (day)^2 + \beta_3 \cdot (day)^3 + \beta_4 \cdot D^{Holiday}$$
$$+\beta_5 \cdot D^{NYE} + \beta_6 \cdot D^{August} + \beta_7 \cdot D^{Summer} + \epsilon$$

Then using an OLS Regression, this model produced the following results:

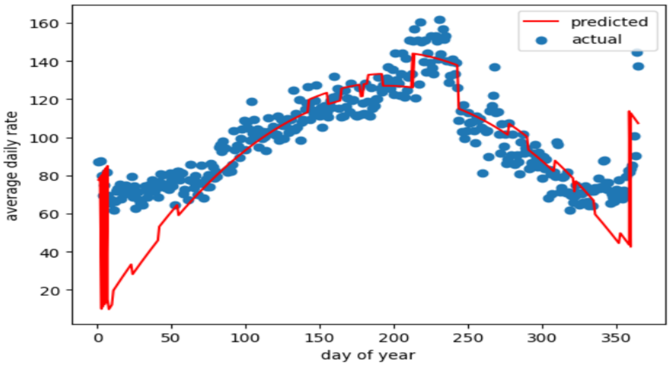


Explanatory Variable	Coefficient	Standard Error	P-value
Day of year	2.017	0.02	0.00
(Day of year)^2	-0.014	0.00	0.00
(Day of year)^3	0.000	0.00	0.00
(Day of year)^4	0.000	0.00	0.00
Holiday	4.800	0.33	0.00
New Year's Eve	62.220	1.11	0.00
August	17.380	0.64	0.00
Summer	18.585	0.56	0.00

As can be seen, the cubic and quartic variables within the model barely did anything, with coefficients very close to 0, however, that is likely because these variables can take on values as large as 365^3 and 365^4, respectively. We eventually decided to keep them in the calculation since their corresponding F-Statistic of 1510.85 provides strong evidence to support the hypothesis that these two variables are jointly significant.

$$F = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n-k)} = \frac{(.864 - .859)/2}{(1 - .864)/(82198 - 8)} = 1510.85$$

Using the test data from a test_train_split and then plugging it into the model we obtained, the following graph is formed.



The segments of the line graph at the far left and right that look thicker than the rest of the line, are due to the fact that our model includes a dummy variable for a date being in the first or last week of the year, but the same "week of the year" is composed of different dates depending on whether the year is a leap year or not since our dataset contains data from one leap year and two non-leap years. Therefore, our model predicts very different outcomes for observations with the same day that are considered parts of different weeks. The "thick segments" are actually just the line graph fluctuating back and forth very frequently when this happens.

The model seemed to generally fit the data, and since we did not want to add even more variables, maybe causing the model to over-fit the data, the graph is exactly what we wanted. As can be seen, the largest change by far in the prices of hotels in the data set is during New Year's. This makes sense, as we previously discussed, New Year's is a holiday celebrated all around the world, not just in the United States, so there would probably be much more demand for hotels in comparison to America-centric holidays. The other large spike in prices can be seen in the summer, generally around July/August. This would also make sense as not only are schools no longer in session but it is generally seen as a period of relaxation and travel.

8 Conclusion

Our takeaway for our first question is that on average, prices first increase and then later decrease over the span of a year. In general, it is better to book a hotel later in the year and for short periods of time in order to obtain a better price. Our takeaway for our second question is that by using logistic regression, our predictive model showed that repeated guests tend to not cancel previous bookings, to stay for short periods, to book close to their arrival date, to make more booking changes, and to spend little time on a hotel's waiting list. Our takeaway for the third question is that the average daily rate increases while the length of stay increases till around the fifth day. From then onwards, the average daily rate decreases steadily until it reaches its lowest rate around the 60th day. After the 60th day, there is a sharp increase in the average daily rate. Our fourth takeaway is that while general holidays don't appear to influence prices greatly, international holidays like New Years Eve seem to have a huge impact on prices.