Analyzing City Hotel Behavior

Josh Janda

07 May 2020

Section 1 - Introduction

For this project, I will be analyzing the Hotel Booking Demand dataset, which can be found on Kaggle. Specifically, I will be focusing on hotels that are **City Hotel's**.

The goal for this project is to utilize this dataset to help predict the average daily rate of a hotel given a number of factors. The average daily rate is computed by dividing the sum of all lodging transactions by the total number of staying nights. This is important to know as a consumer as one would want to know their average price per night if staying multiple nights as it is known that hotel prices fluctuate heavily by date.

The data represents hotel bookings from July 1st, 2015 until August 31st, 2017. While this is somewhat outdated, I may still be able to draw insight on future hotel average daily rates as historical data is important.

Overall, using multiple techniques in data analysis and predictive modeling I hope to create an accurate model for predicting a city hotel's average daily rate.

Section 2 - Exploratory Data Analysis

The first steps for exploratory data analysis is to give a brief description on all variables in the model.

- **is_canceled** (categorical) - was booking canceled or not
- **lead_time** (numeric) - number of days between booking and arrival
- **arrival_date_year** (categorical) - year of arrival date
- **arrival_date_month** (categorical) - month of arrival date
- **arrival_date_week_number** (categorical) - week number of year for arrival date
- **arrival_date_day_of_month** (categorical) - day of arrival date
- **stay_in_weekend_nights** (numeric) - number of weekend nights stayed
- **stay_in_week_nights** (numeric) - number of week nights stayed
- **adults** (categorical) - number of adults
- **children** (categorical) - number of children
- **babies** (categorical) - number of babies
- **meal** (categorical) - type of meal booked
- **market_segment** (categorical) - market segment designation
- **reserved_room_type** (categorical) - type of room reserved
- **customer_type** (categorical) - type of customer
- **adr** (numeric, target variable) - average daily rate
- **total_of_special_requests** (categorical) - number of special requests

With variables defined by type and meaning, we can begin analysis. The first variables I want to look into are the time variables. My initial thoughts are that the best time variables to keep will be the arrival date months and day of month. These values to me will hold the most information for predicting average daily rates.

Section 2.1 - Categorical/Time Variables

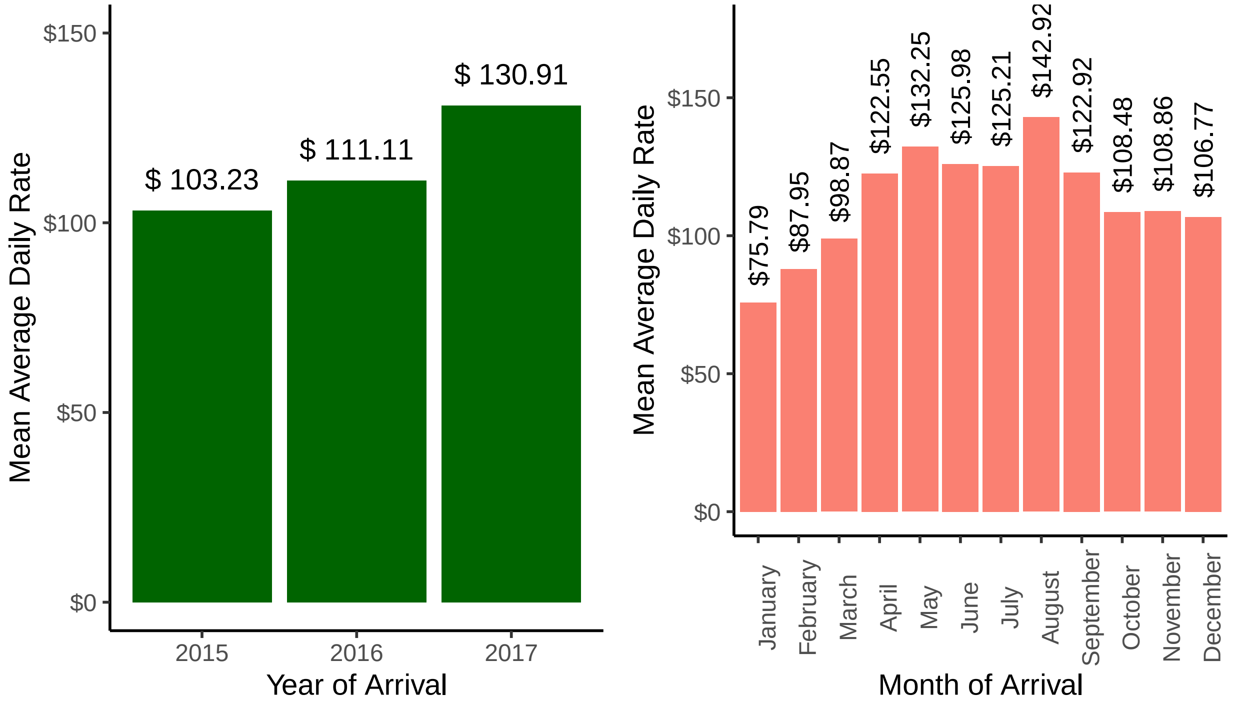

Figure 1: Mean Average Daily Rate by Year & Month of Arrival

Figure 1 above shows bar plots containing information on the mean average daily rate against year and month of arrival.

For starters, it is useful to look at the average daily rates over the years. We see that mean daily rates increase over each year. This indicates to me that year can be a useful predictor as future years may mean higher daily rates than previous years.

Next up, I look at the month plot. The first three months of the year have the lowest daily rate which is most likely due to low travel during these times. The average daily rate is at it's highest during the spring/summer months, and then begins to decline during the beginning of winter.

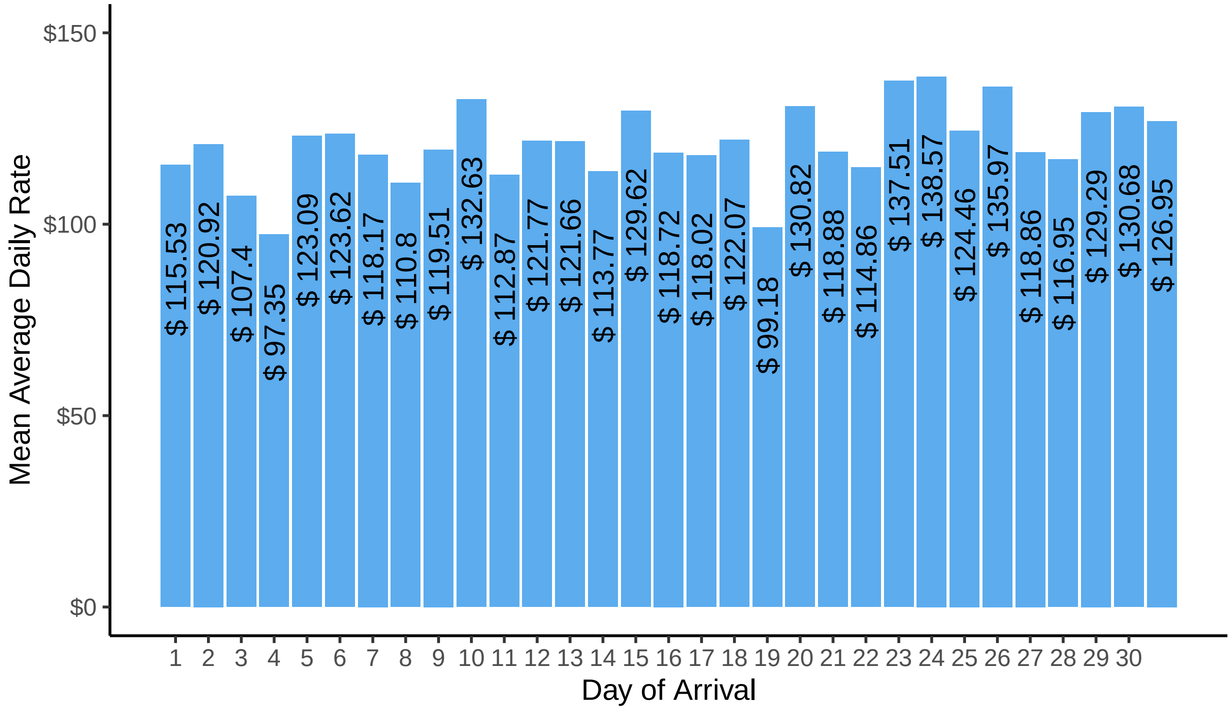

Figure 2: Mean Average Daily Rate by Day of Arrival

Figure 2 above shows bar plots containing information on the mean average daily rate against day of arrival.

Using the plot above, I can see that there is not much association between day of arrival and the average daily rate. There is some fluctuation, but there is no real trend to see that tells me there is any specific day of the month that has the lowest daily rate. I believe this variable is not worth keeping as there is not much information.

Lastly, I want to look at the arrival week number. I believe I will see similar results to arrival month.

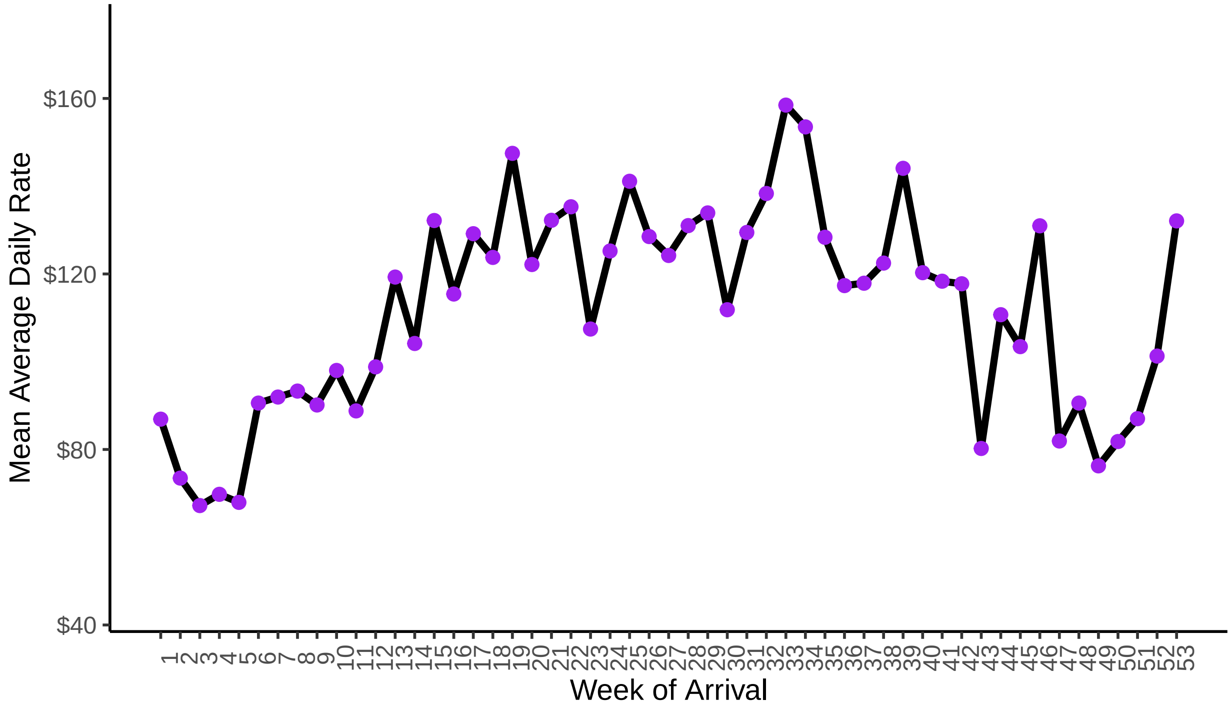

Figure 3: Mean Average Daily Rate by Week of Arrival

Figure 3 above shows bar plots containing information on the mean average daily rate by week of arrival.

As expected, I see similar results to the month plot. This tells me that keeping this variable is not a good idea, so it will be dropped in favor of the month variable (also much less categories).

Overall, through these plots I have determined that keeping both year and month as predictors and dropping week of arrival and day as month will be the best plan for modeling the average daily rate for hotels. So, I will not be keeping all time variables in my final model. As mentioned above, I will be keeping month in favor of week due to similar patterns and less categories to be considered which will lead to more generalization in the model.

Next, I want to consider an interaction between month and year.

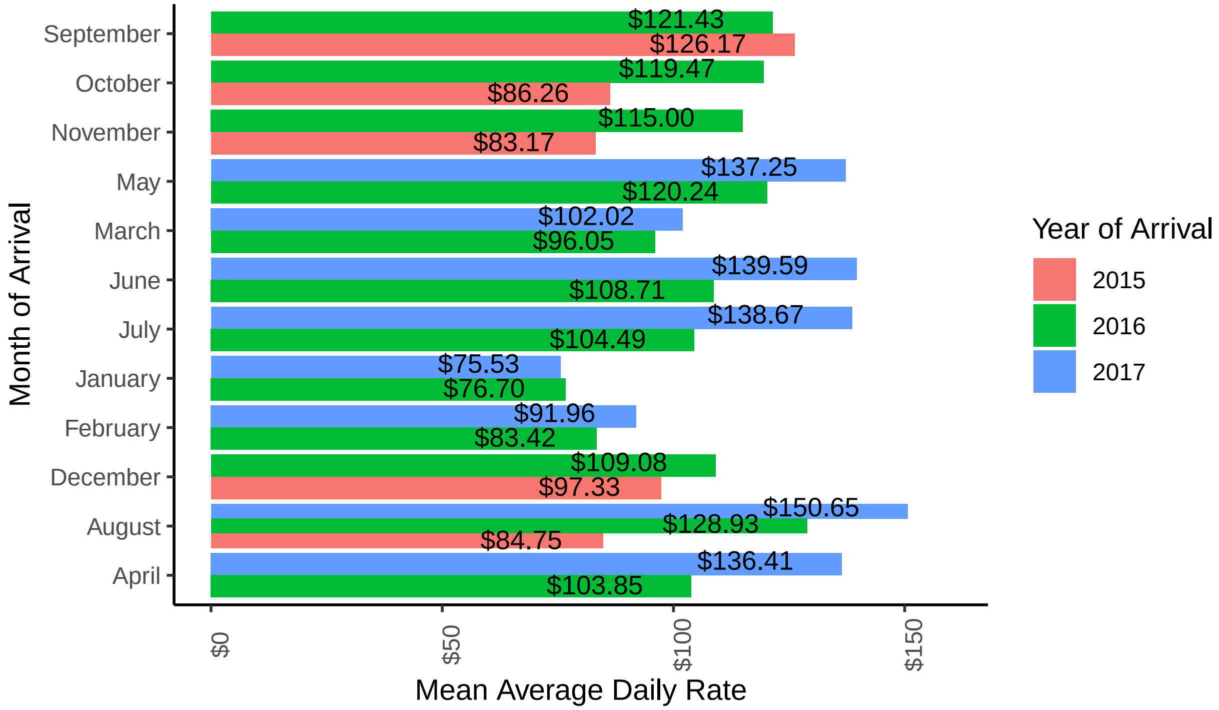

Figure 4: Mean Average Daily Rate by Month of Arrival − Grouped by Year of Arrival

Figure 4 above shows bar plots containing information on the mean average daily rate by Month of Arrival - Grouped by Year of Arrival.

Looking at the plot above, I am able to see if there is any type of interaction occuring between year of arrival and month of arrival on the average daily rate. Note that some months do not have all three years included, but may only include data from 2015 & 2016 or 2016 & 2017 (excluding August, which includes all three years).

Judging by the output, there does seem to be *some* interaction occuring between these two variables for some months. This may be due to yearly price increases on top of seasonal price changes. With that said, I will include an interaction between year and month in my model.

With time aside, I want to look more into the remaining variables.

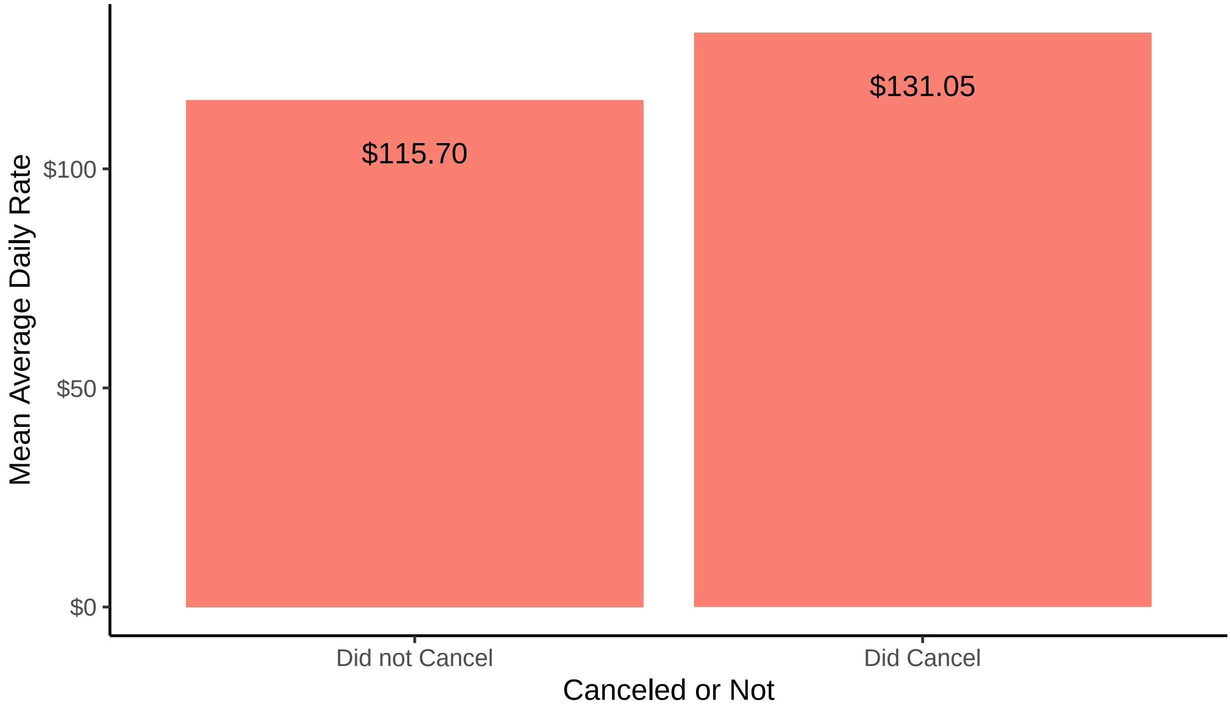

Figure 5: Mean Average Daily Rate by Cancellations

Figure 5 above shows bar plots containing information on the mean average daily rate by cancellations.

Looking at the plot above, there does seem to be a sizeable difference between the mean average daily rate of those who cancelled versus those that did not. This tells me that people were more likely to cancel if the average daily rate was too high.

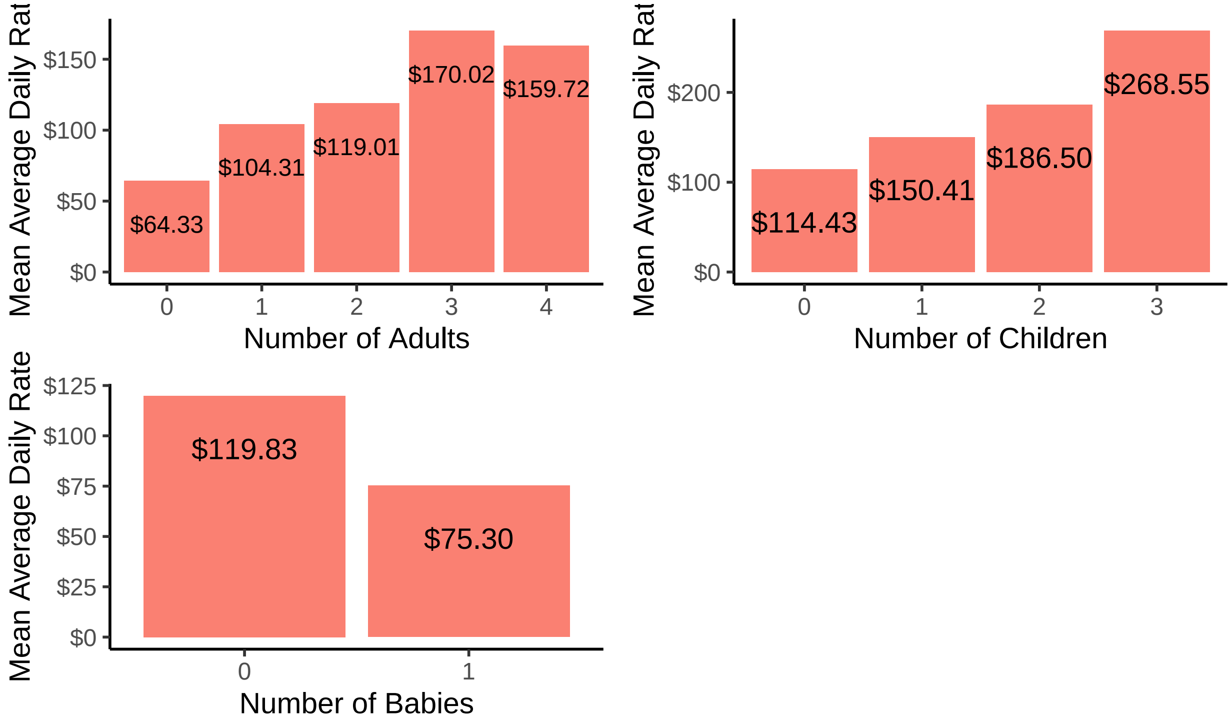

Figure 6: Mean Average Daily Rate by Number of Adults, Children, and Babies

Figure 6 above shows bar plots containing information on the mean average daily rate by number of adults, children, and babies.

Looking at the grid above, I am able to analyze three categorical variables pertaining to the number of individuals staying in the hotel room. These individuals are split up into adults, children, and babies.

For adults, The mean average daily rate shows an increasing trend as number of adults increases up until three and then decreases at four adults. This possibly demonstrates that three adults may need a hotel room with two rooms, while four adults may be cramming into one room with two beds.

For children, there is a clear linear trend that as the number of children increases so does the mean average daily rate. This is very understandable as more children means more space required.

For babies, we see a decreasing linear trend such that having a baby in the hotel room leads to a cheaper average daily rate. This may be due to having a baby means less children and therefore less space required in the hotel room.

I want to also look at a possible interaction between these variables.

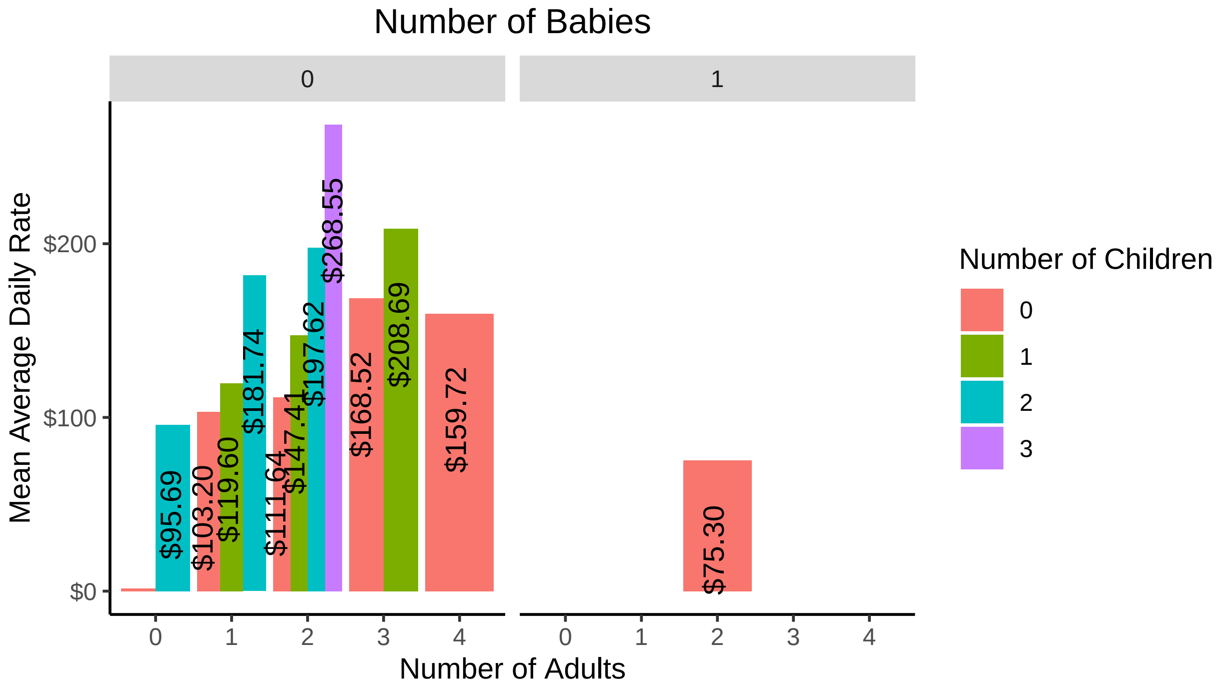

Figure 7: Mean Average Daily Rate by Number of Adults − Grouped by Number of Children Faceted by Number of Babies

Figure 7 above shows bar plots containing information on the mean average daily rate against number of adults - grouped by number of children and faceted by number of babies.

Looking at the plot above, I can visualize the interactions between number of adults, children, and babies.

First thing that I notice is that there are almost no observations in the 1 babies category. This tells me that having a baby is an outlier, and should be removed as it will cause complications in the modeling process.

For adults and children, there is definitely some interaction at play. There seems to be a trend that as number of adults increase, so does the mean average daily rate. This can also be said for the number of children. When looking at adults and children combined, there is a linear trend that as both increase so does the mean average daily rate.

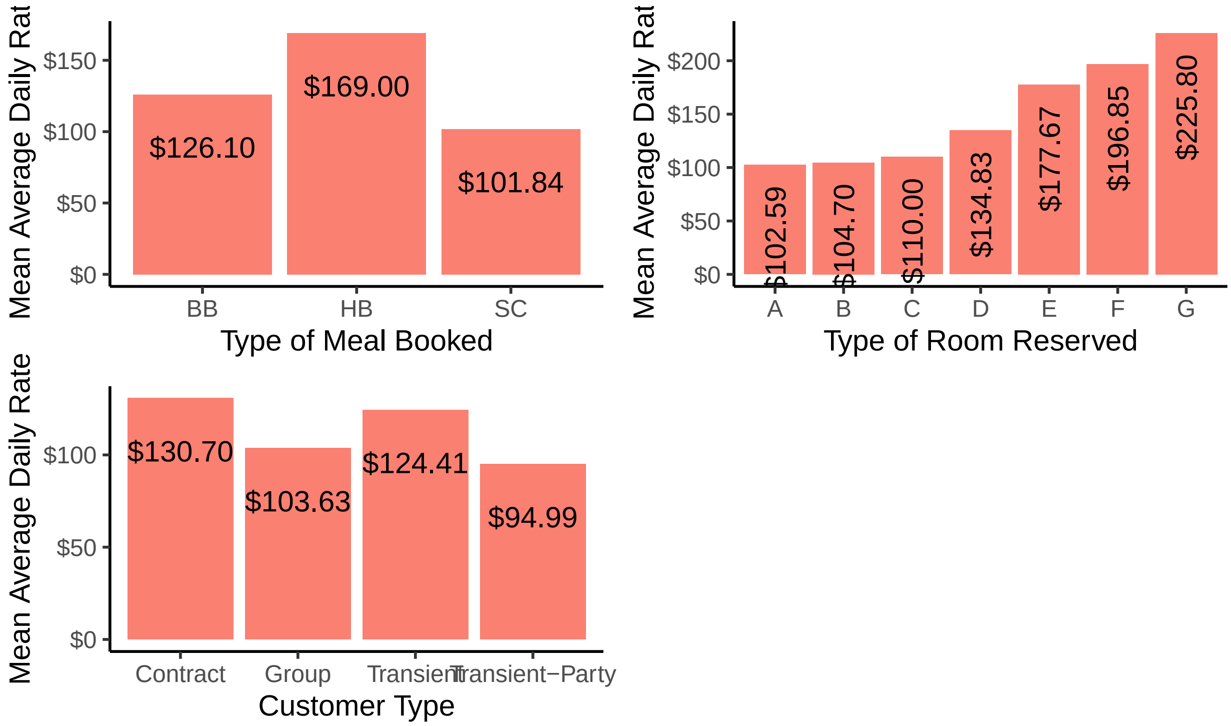

Figure 8: Mean Average Daily Rate by Type of Customer, Meal Booked, and Room Reserved

Figure 8 above shows bar plots containing information on the mean average daily rate by type of customer, meal booked, and room reserved.

The plot above looks at the relationships between mean average daily rate and customer type, type of meal booked, and type of room reserved.

For type of meal booked, the `HB` meal results in the most expensive average daily rate. There is definitely a difference between average daily rates and the type of meal booked, so this predictor will be kept for modeling.

For type of room reserved, there seems to be a linear trend with room type and average daily rate beginning with room type `A` as the cheapest, and ending with room type `G` as the most expensive. It should be noted that there is not a great difference between room types `A`, `B`, and `C` which may reflect that these rooms are all the same size but might include a different number of beds.

For customer type, there is also a noticeable difference between mean average daily rates. `Contract` customers have the highest mean average daily rate, which is understandable to me as contract customers may have a pre-negotiated rate that does not have seasonal fluctuation. `Transient` type have the second highest rate, which is also understandable as these are typically urgent and short stay guests. `Group` type customers have the third highest, due to the nature of Group bookings getting a deal on the rate due to booking a mass amount of rooms. Lastly, `Transient-Party` have the lowest rate which may possibly be due to these transient customers being regular and therefore having a deal.

I am not interested in interactions between these variables as they are not really related to one another.

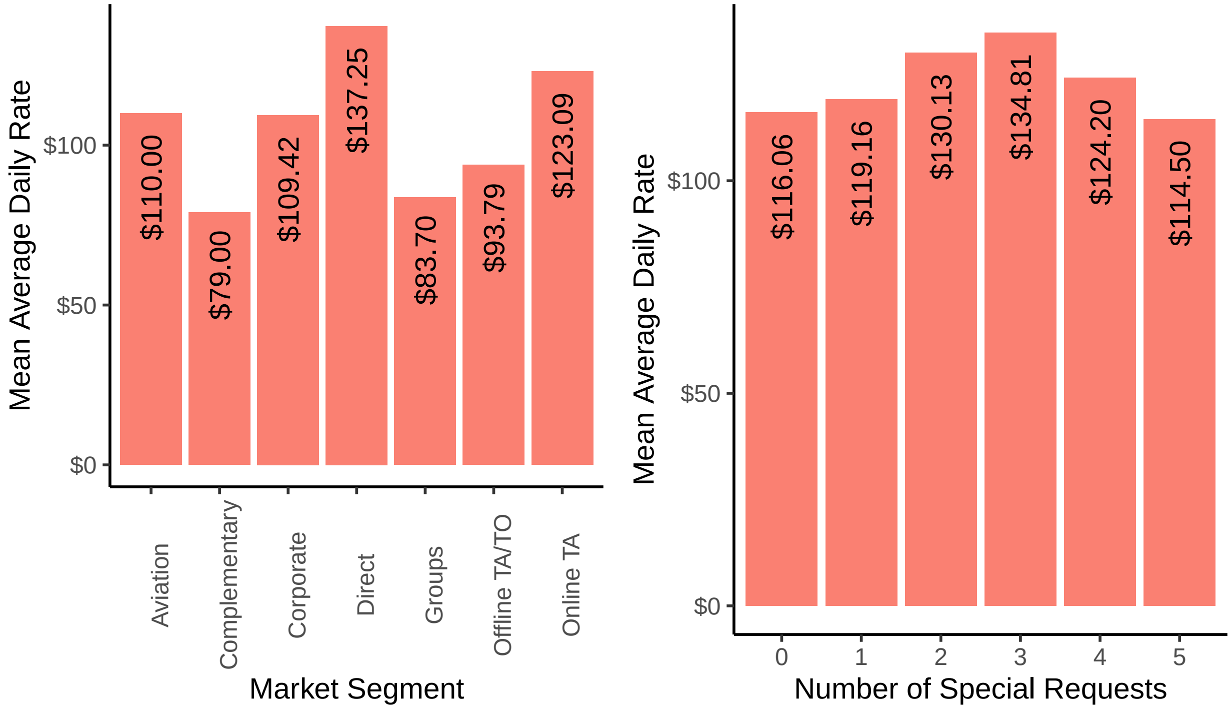

Figure 9: Mean Average Daily Rate by Market Segment and Number of Special Requests

Figure 9 above shows bar plots containing information on the mean average daily rate by market segment and number of special requests.

The plot above looks at the relationships between mean average daily rate and market segment and number of special requests.

For number of special requests, it is quite obvious that it does not really effect the average daily rate for the room. This makes sense to me as most special requests in hotels are free of charge outside of laundry or such. Therefore, I may want to look into removing this variable.

For the market segment, there is a noticeable difference between most segments and the mean average daily rate. The `Direct` segment, or typically those that book a hotel themselves in person, have the highest mean average daily rate. This makes sense as there are usually no deals for these consumers. For the `Online TA` segment, they have the second highest mean average daily rate due to also not receiving many deals but getting a cheaper rate due to online booking. The market segment with the lowest mean average daily rate is the `Complementary` segment, which is typically the segment that receives complimentary bookings for extremely low prices due to issues with previous bookings.

Section 2.2 - Numeric Variables

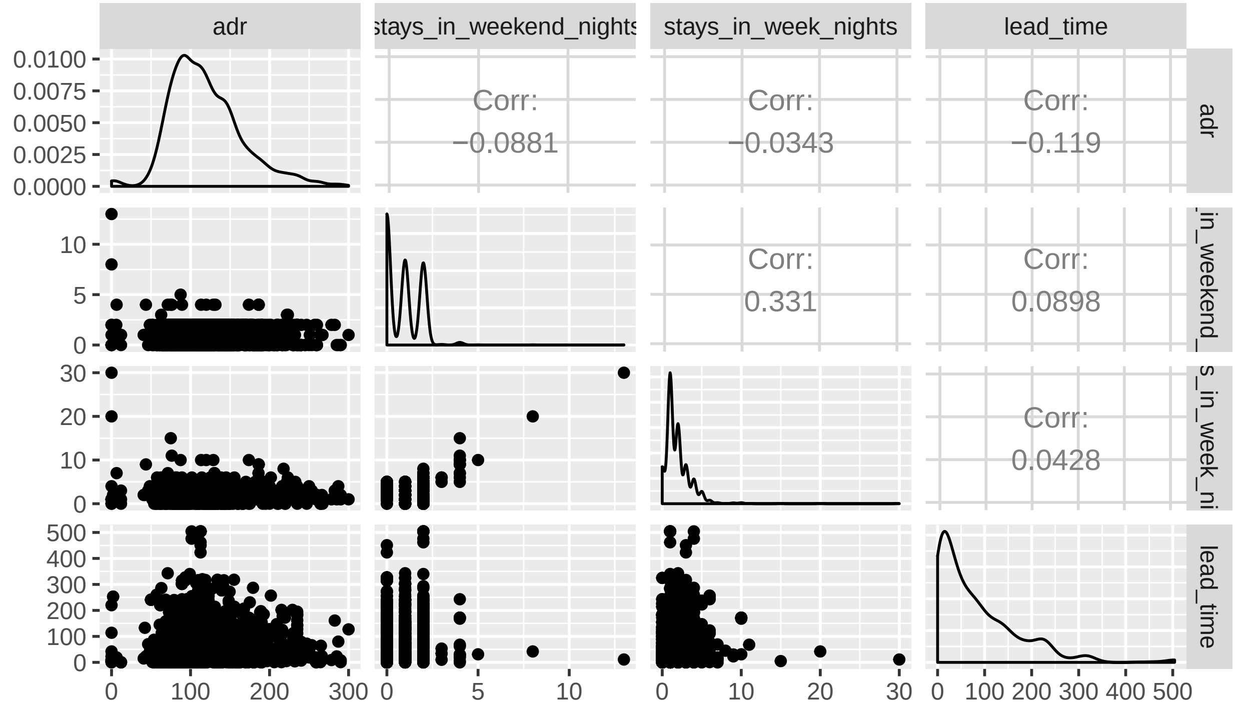

Figure 10: Scatter Matrix of Numeric Variables

Figure 10 contains a scatter matrix on all numeric variables of the data.

The plot above shows the paired correlation and scatter plots between variables `adr`, `stays_in_weekend_nights`, `stays_in_week_nights`, and `lead_time`. Judging by the correlation, the stays in both weekend and week nights are weekly correlated with the average daily rate but have a good amount of positive correlation with each other indicating that as one increases so does the other. This leads me to believe that combining these two variables may prove to be useful. Another observation is that `lead_time` and `adr` have weak negative correlation.

For nonlinear trends, it is obvious that for average daily rate and weekend nights stayed there is no linear trend. Same can be said about average daily rate and week nights stayed. While slightly less nonlinear, average daily rate and lead time also have a mostly nonlinear trend.

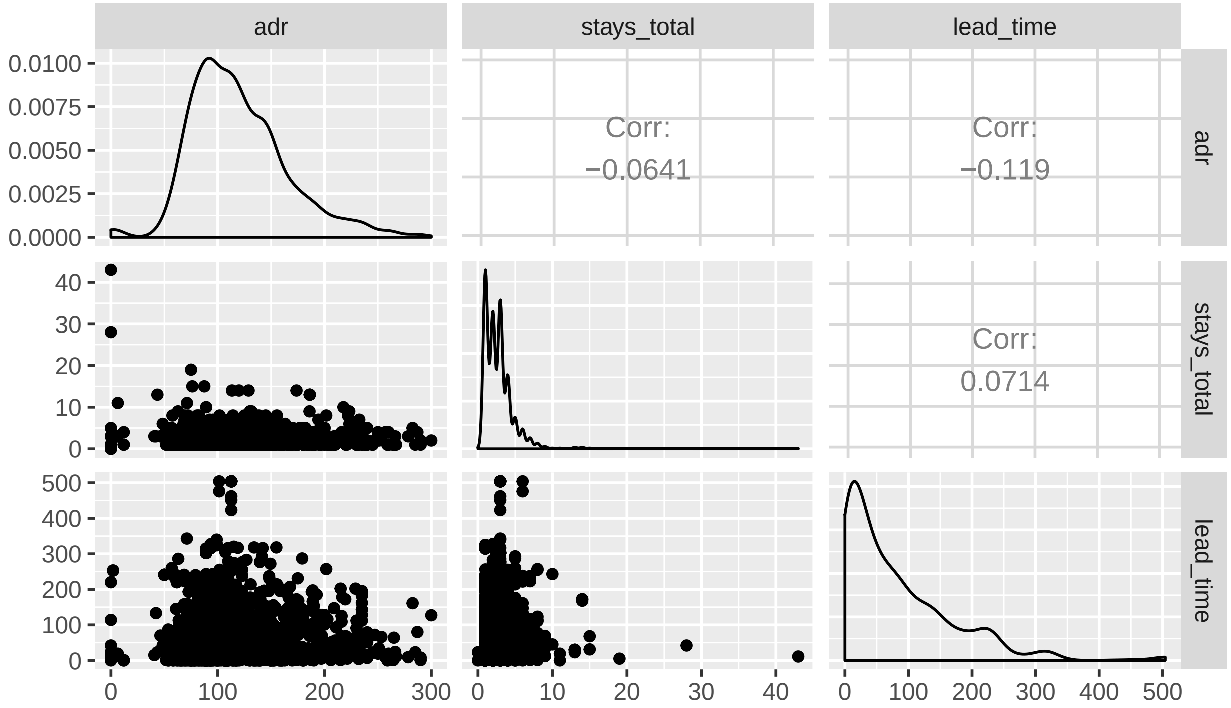

Figure 11: Scatter Matrix of Numeric Variables − Combined Stays

Figure 11 contains a scatter matrix on numeric variables - with newly created `stays_total` variable.

Looking at the output above, there still does not seem to be correlation between this total days stayed variable and average daily rate as well as with lead time. Therefore, I am deciding to not include the total number of stays as a variable in my models as it does not contain much useful information to the average daily rate. I will however keep the lead time variable as I believe it's correlation, albeit weak, may play a role in modeling.

With all variables looked over, it should be stated that the following variables will be kept for the modeling process:

- `lead_time`
- `is_canceled`
- `arrival_date_year`
- `arrival_date_month`
- `adults`
- `children`
- `meal`
- `market_segment`
- `reserved_room_type`
- `customer_type`
- `total_of_special_requests`

With all variables looked over and explained of why I am keeping the variable or not, I am moving onto modeling.

Section 3 - Methods

Section 3.1 - Linear Regression

For the first model, I will be creating a multiple linear regression using all variables mentioned in section 2. The implementation of this model is quite simple, and parameters can be obtained by the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Where X is the design matrix and y is the output variable (average daily rate). The equation above can be seen as obtaining parameters that minimize the sum of squared residuals of the model. Some key assumptions for this model are:

- Residuals are normally distributed
- Independent variables are not highly correlated
- Constant error variance
- Errors are indpendent of one anotherr
- Y is a linear function of the predictors

With implementation and methodology introduced, I will create the model. I generate a few models, and utilize BIC to choose the best model.

Table 1: Comparison of Linear Regression Models

	BIC	NumParameters
Simple	10782.57	79
Subset	11017.33	9
Quadratic	10607.14	37

Table 1 above shows a comparison between linear regression models using BIC/# of Parameters.

The model with the lowest BIC is the quadratic model with `lead_time` having a squared term. I will use this model as the best linear regression. I will also compute the RMSE on this model to compare to further models created.

I want to take a quick look at variables removed in the quadratic model.

Table 2: Removed Variables from Quadratic Model

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1572	932678.2	10625.31
- customer_type	3	8146.609	1575	940824.8	10617.21
- market_segment	6	20134.826	1581	960959.7	10607.14

Table 2 above shows the removed variables and associated details on the quadratic model.

So dropped variables were `market_segment` and `customer_type`. This is understandable to me as we did not see a huge impact on mean average daily rate by these variables when visualizing them beforehand.

Section 3.2 - Predict with Linear Regression Model

For this section, I will compute predictions on the linear regression model I selected.

We see an RMSE from this model of 24.37 which is quite good given a standard deviation of the average daily rate of 45.189.

One example that comes to mind is say you want to take advantage of a vacation after Christmas. So you book a vacation for January of 2017, with a lead time of ~70 days which means you booked the vacation in October. You have 3 kids and 2 adults, so you reserve a room type of G assuming G is the biggest room. You make zero special requests, and you have not cancelled the vacation. You book a hotel that provides a hot breakfast.

With this scenario, you can expect to pay an average daily rate of $234.95.

Section 3.3 - Random Forest

Next up, I want to create a non-parametric type of model which will be a Random Forest. The methodology for this model is to not assume our data follows any functional form, and therefore no assumptions are made. The implementation for this model is to grow mass amounts of decision trees using bootstrap aggregation. This can be boiled down to sampling data with replacement and growing a tree on this sampled data using a random sample of predictors (typically ~1/3 of the predictors used in each tree). With this method, we will have a mass amount of decorrelated trees which we can use to make predictions to aggregate into one mean predicted value.

I will be using 5000 trees in my model.

With this tree, I get an Out-of-Bag Error (Mean-Squared-Error) estimate of 560.872. Taking the square root (RMSE), I get 23.683 which is slightly better than the best linear regression model.

It is also important to me to understand how this model is prioritizing variables.

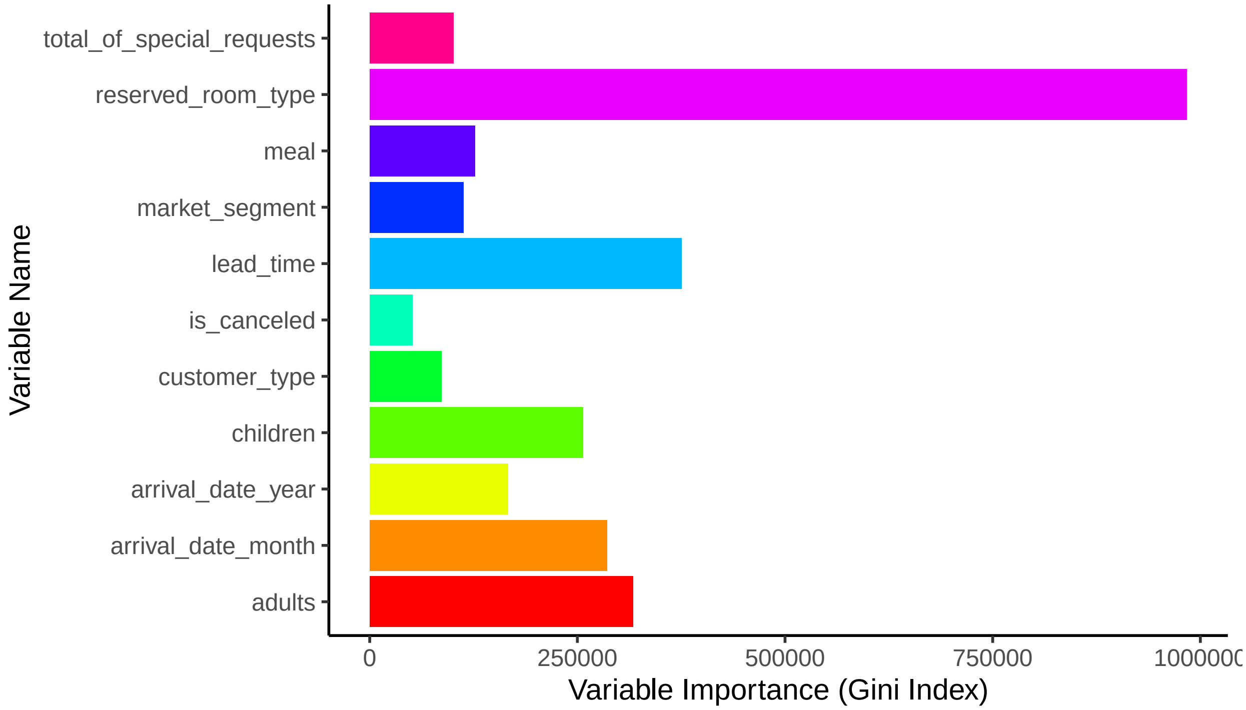

Figure 12: Variable Importance Plot for RF Model

Figure 12 above holds information regarding variable importance in the random forest model.

Using the above plot, I am able to directly visualize the importance of each variable in the random forest model. The most important variable by far is `room_reserved_type`, which was also seen in the linear regression model. The least important variables are `is_canceled`, `customer_type`, and `market_segment`. This is similar to our best regression model, which completely dropped `customer_type` and `market_segment`.

Section 3.4 - Predict with Random Forest Model

We see an RMSE from this model of 15.492 which is quite good given a standard deviation of the average daily rate of 45.189.

Using the same example laid out as above, I will predict with the random forest model. I will include a market segment of `online` and a customer type of `transient`.

With this scenario, you can expect to pay an average daily rate of $71.27. Note this is much smaller than the prediction from the Linear Regression model. Given the lower OOB-Error than the Linear Regressions training error, this prediction seems more accurate to me.

Section 4 - Results

Through this project I have analyzed numerous categorical and numeric variables and then utilized these variables for creating different models such as linear/polynomial regressions and random forests.

Table 3: Comparison of Best Linear Regression and Random Forest

	RMSE
LinearRegression	24.37044
RandomForest	15.49229

The table above (table 3), displays the root-mean-squared-error for between the best regression model and the random forest model. I see that the Random Forest model has a substantially lower RMSE, and should be deemed as the best model.

The results obtained from this data and models created are overall very good, and therefore these models may be deemed useful for understanding and predicting average daily rates when booking a hotel room.