CITY HOTELS ANALYSIS

Andro Manukov

Section 1: Introduction:

The goal of this project is to find the optimal set of variables to predict the Average Daily Rate (ADR) of someone staying in a city hotel. This information is useful as it allows people to make sure they're getting a decent price based on the findings in this report. I plan to deep dive into all of the variables and provide sound explanations to whether I should include them or not.

The data in this project comes from the Hotel booking demand dataset from Kaggle. The Hotel booking demand dataset consists of observations from 2015 to 2017 with information regarding various details about the type of booking. There is more information on how this dataset was obtained if you click here. It is an interesting read and will provide more insight as you go through my analysis.

Section 2: Exploratory Data Analysis:

Section 2.1: Given Variables:

- Categorical:
  - **hotel** - City or Resort
  - **is_canceled** - Value indicating if the booking was canceled (1) or not (0)
  - **arrival_date_year** - Year of arrival date
  - **arrival_date_month** - Month of arrival date
  - **arrival_date_week_number** - Week number of year for arrival date
  - **arrival_date_day_of_month** - Day of arrival date
  - **meal** - Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
  - **market_Segment** - Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
  - **reserved_room_type** - Code of room type reserved.
  - **customer_type** - type of customer; split into 4 categories.
- Numerical:
  - **adr** - Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
  - **total_of_special_requests** - Number of special requests made by the customer (e.g. twin bed or high floor)
  - **babies** - Number of babies

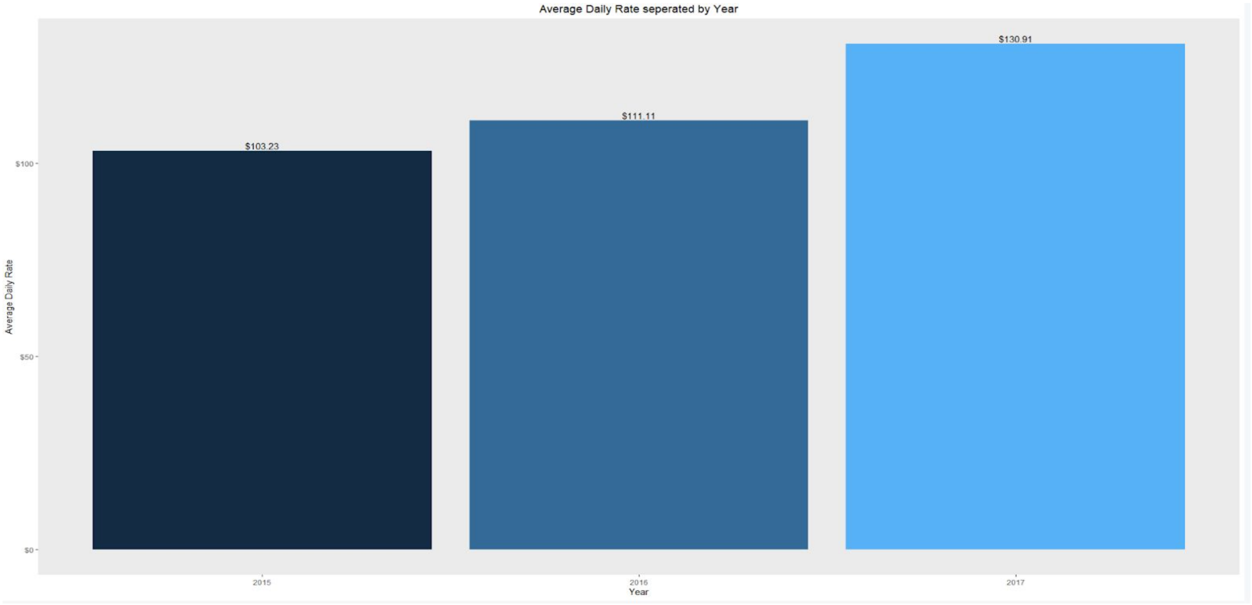
- **children** - Number of children
- **adults** - Number of adults
- **stays_in_week_nights** - Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **stays_in_weekend_nights** - Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **lead_time** - Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

Note: There are some variables that can go either way. This will be addressed later in the report.

Section 2.2/2.3: Time Related Variables:

There are a ton of time related variables in this dataset and prior to doing any analysis, I can certainly say that we will not be keeping all of them.

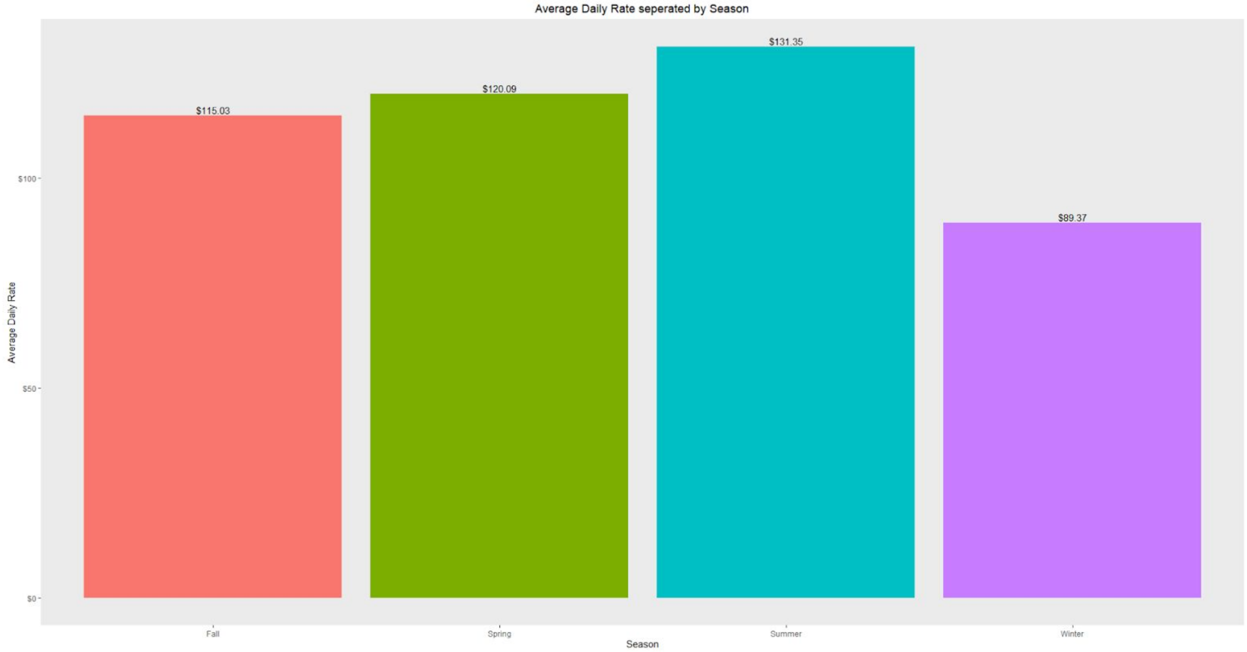
Arrival_Date_Year:



Prior to this bar chart, I didn’t think the year variable would have a huge effect. I’m still skeptical since this is only 3 years of data but maybe the future years will continue to increase ADR. Nonetheless, since the difference between years is significant, I will keep it in the model. However, it is important to take note that this variable could be volatile. (For example 2020 and COVID-19 will probably lower ADR by a lot)

Arrival_Date_Month:

One of the most important aspects of data analysis is succinctness. If your data is clustered, no one will want to look at it. That’s why one of the first things I will do is map the arrival_date_month variable to seasons.



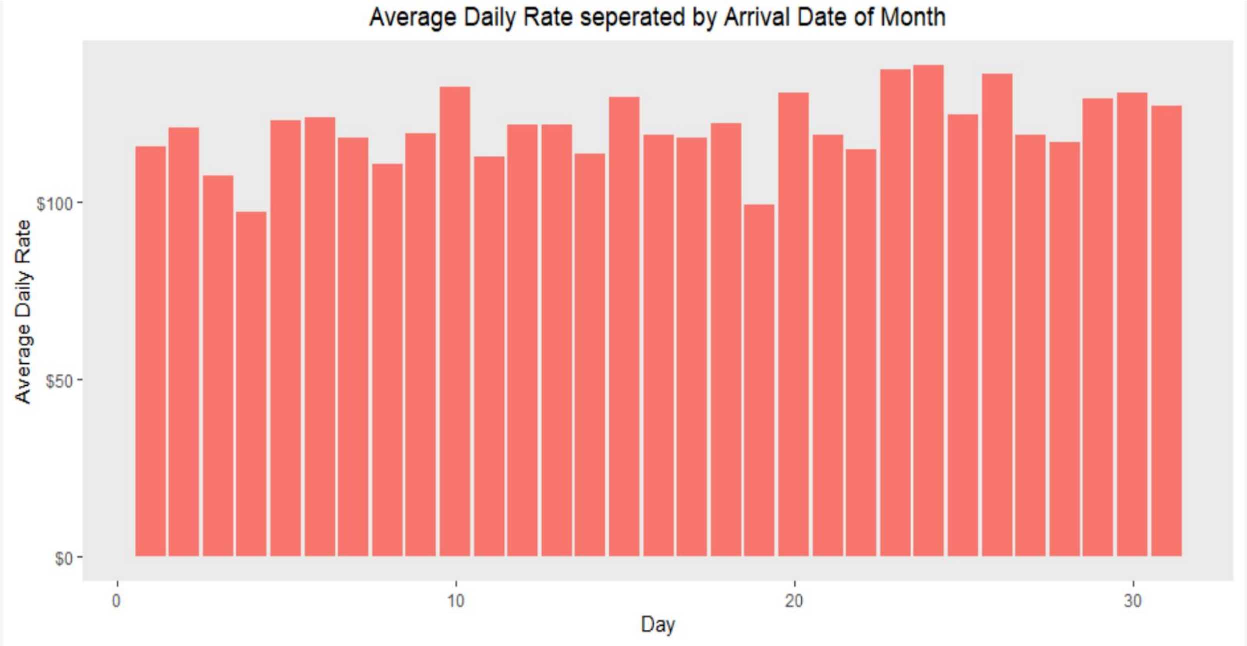
The first thing we see is that there is a clear difference between winter and all of the other seasons. This makes sense as rates are typically cheaper during the winter. I think this will be one of the most significant time variables in this analysis. For that reason, it will be included.

Alternatively, you can create a binary variable that is mapped to 1 for if the season is winter and 0 if it is anything else.

Arrival_Date_Week_Number:

There is no need to do any analysis on this variable since it is basically another way to code month. Since I created the seasons variable, it is fine to leave this variable out of the analysis.

Arrival_Date_Day_Of_Month:

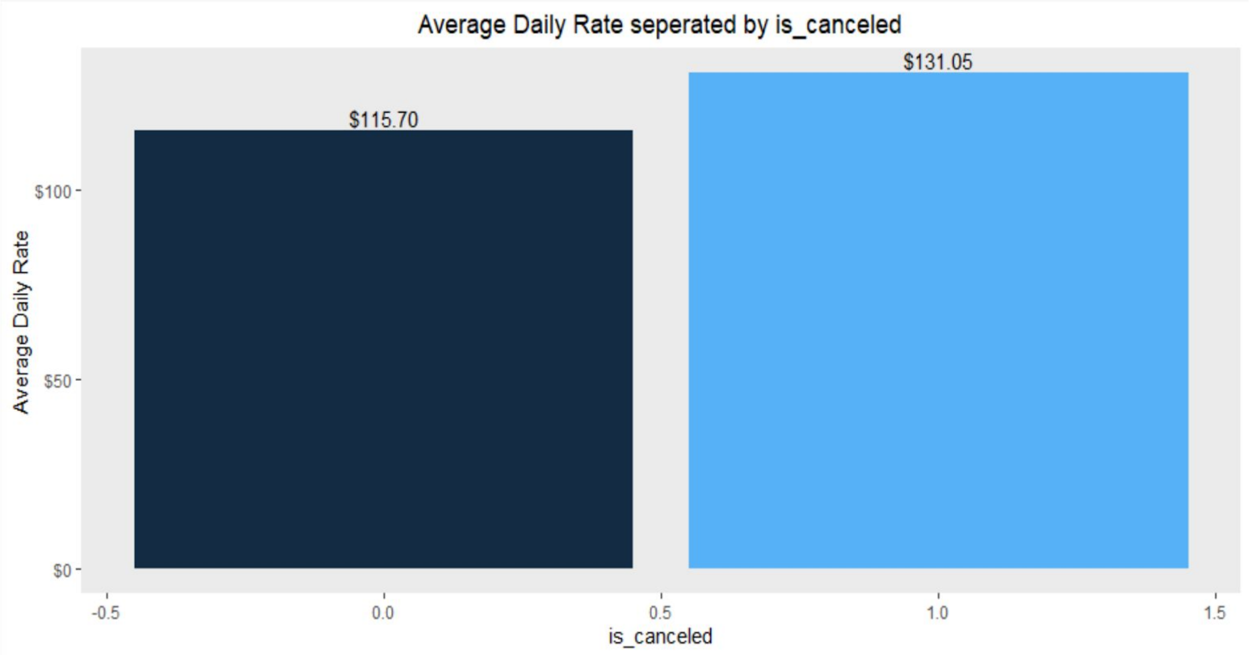


Similar to the previous variable, it’s no surprise that the date of the month doesn’t seem to have an effect. I will leave this out for obvious reasons.

Section 2.3.5: Other Variables

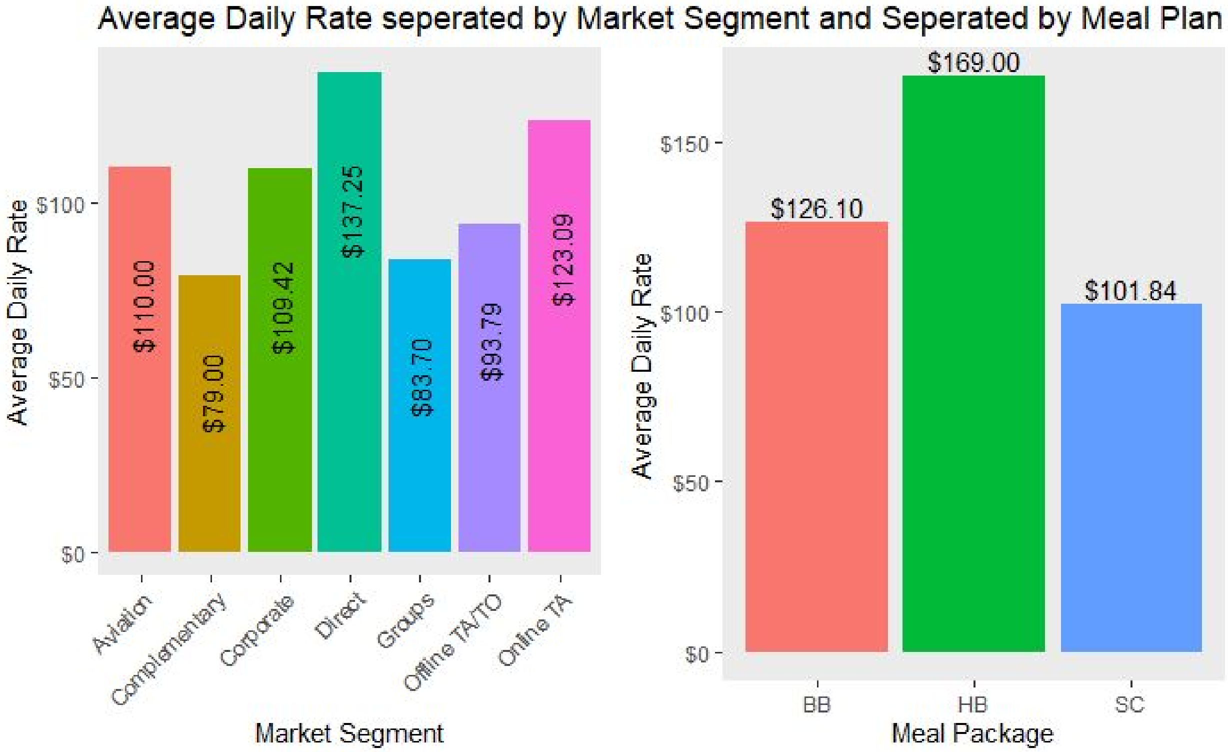
Is_canceled:

The way this variable is tracked is somewhat confused. If a booking is canceled, how can you calculate an accurate ADR? For the purpose of this project, I will assume that ADR is estimated but it is important to note that for cases where a booking is canceled, ADR will be slightly lower than what is should be due to there not being any charges for things such as minibars.



These results seem pretty normal. It makes sense that a booking with a larger ADR is more likely to be canceled. I will include this in the model.

Meal and Market_Segment:

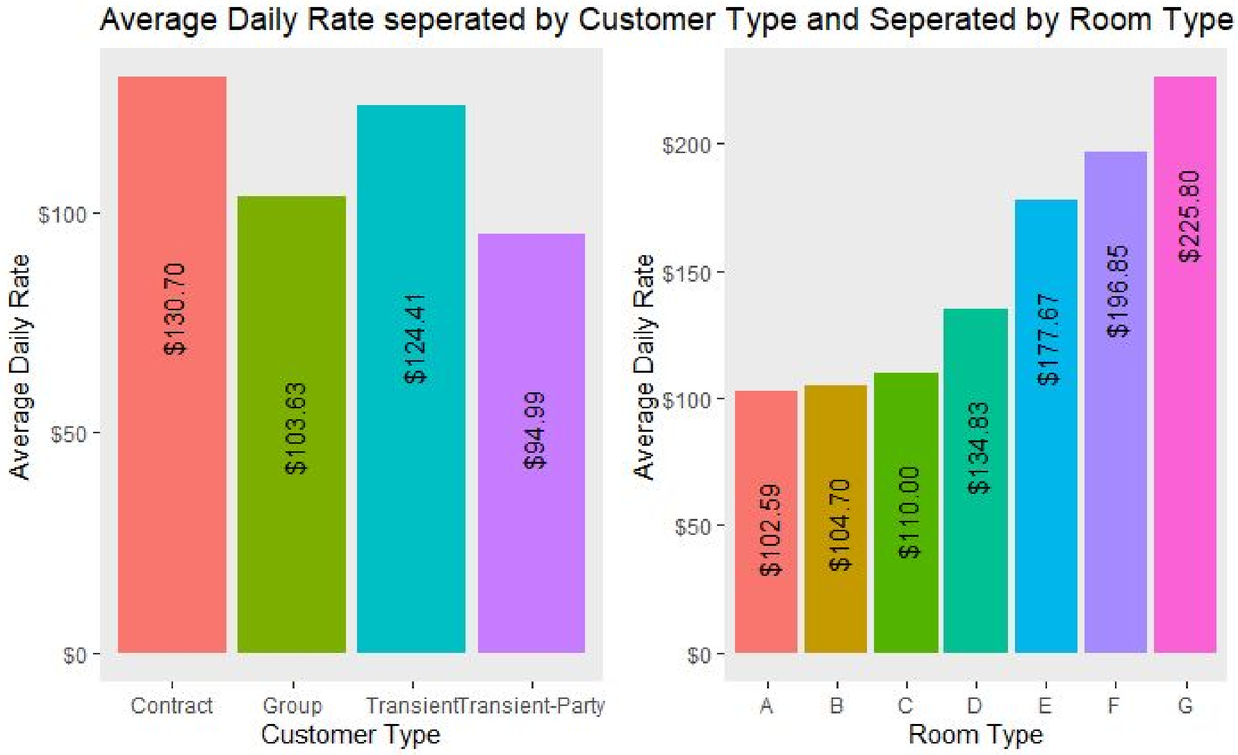


Meal: As mentioned in section 2.1, the meal variable consists of 4 different categories. In order of cheapest to most expensive the meal packages are SC, BB, HB, FB. It's expected that our data will reflect that. The ADR of the different meal packages is as expected. There were no observations with the Full Board package. I believe this is because these are all city hotels and FB is more common in resorts. I will include this in the model.

Market Segment: The most interesting part here is that the ADR for the Direct market segment is the highest. This makes sense as the consumer makes their own booking. It's also interesting that we can form groups here. For example, Aviation and Corporate seem to get some sort of discount. Groups and Complementary seem to get an even bigger discount. This makes sense as it is cheaper to book many hotels at once and complementary rooms are often offered as compensation. It's also good to note that Offline TAs seem to book cheaper rooms than online TAs.

Both of these variables will be included in our model

Reserved_Room_Type and Customer_Type:

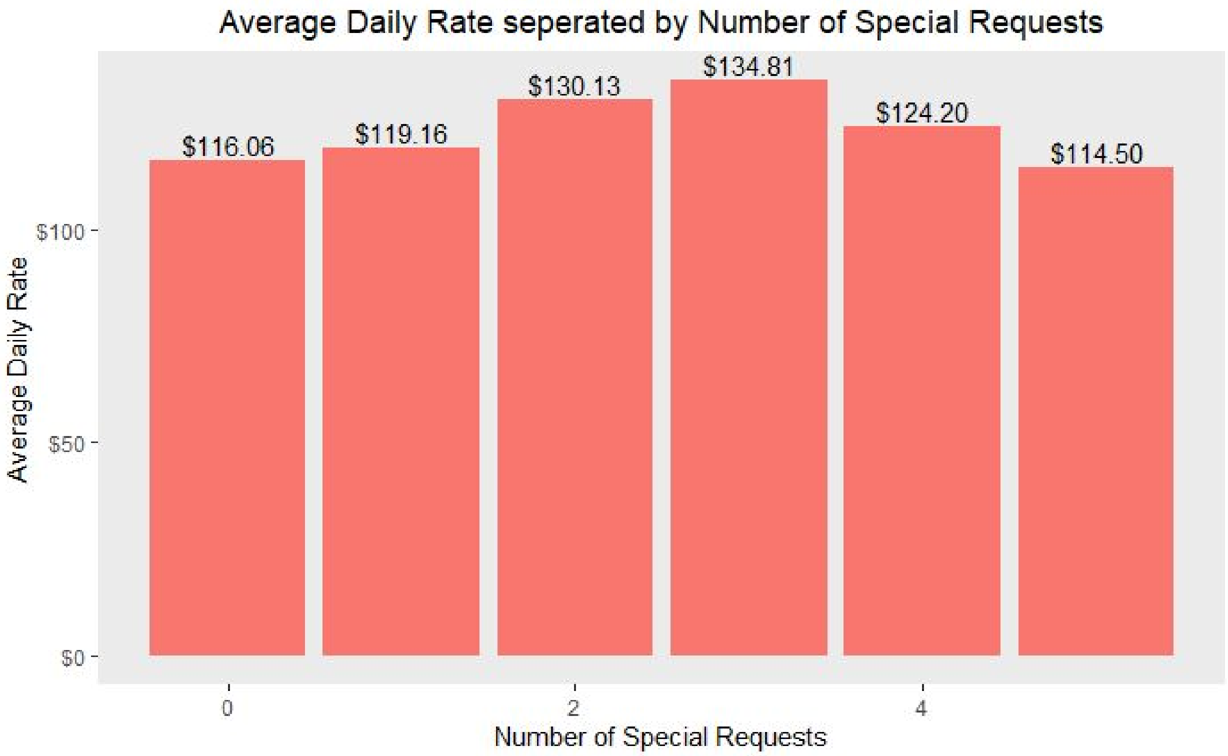


Customer Type: I was not sure what to expect from this variable but it makes sense that the group related categories have a lower ADR. This variable seems important to our model.

Room Type: It's important to note that these labels may vary for different hotels. Nonetheless, there seems to be a general trend that the "higher" the room, the more expensive it is. This variable will be very important in our models.

Total_of_special_requests:

I don’t expect this variable to have an effect on ADR since every hotel I’ve stayed at has done special requests free of charge.



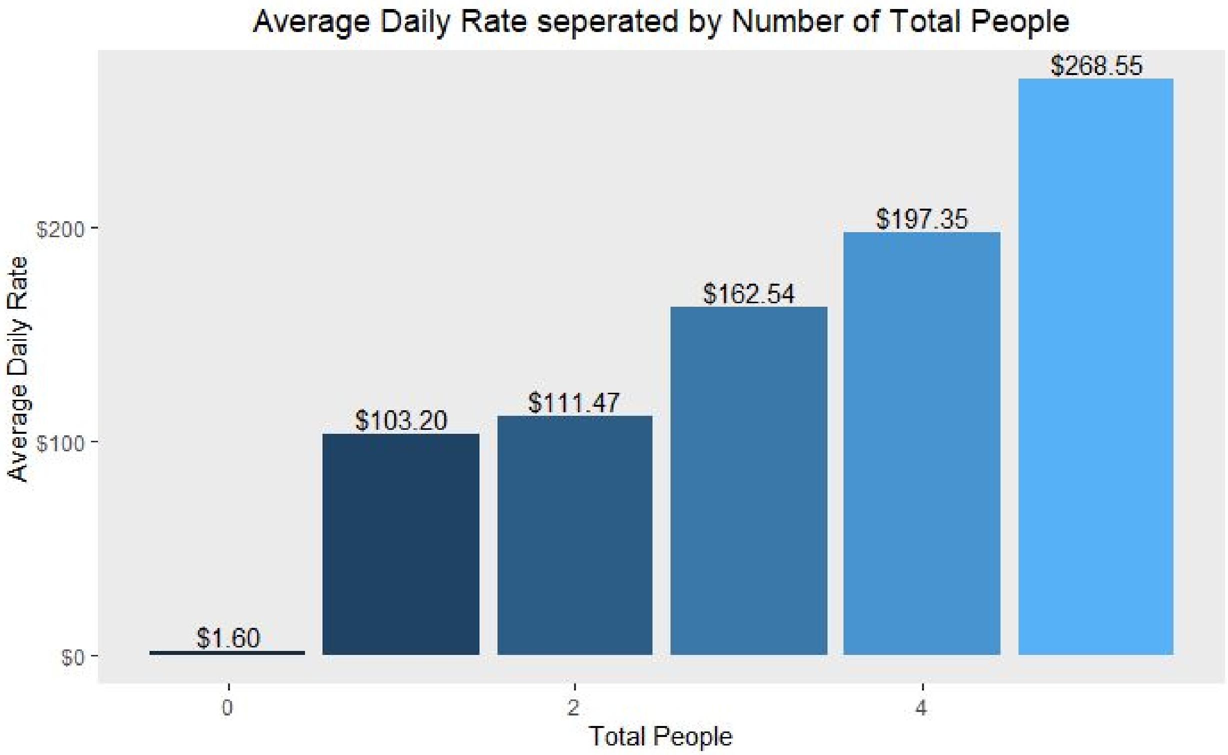
No surprise here, I will be leaving this out of the model

Babies:

Dropping this variable because there are too few observations

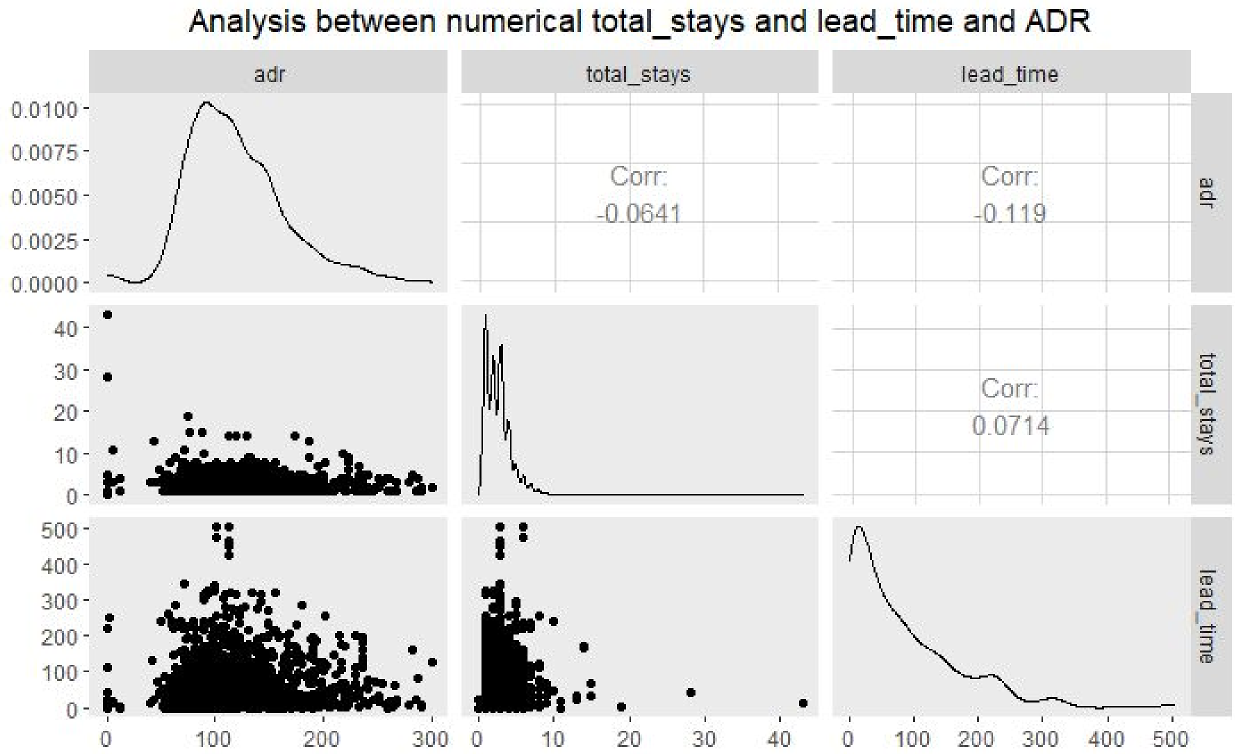
Total People (Adults + Children):

I’ve decided to combine these variables because it doesn't matter if you're an adult or child, you still take up similar space



So obviously the 0 seems like a data entry error so we’re going to ignore that. It makes sense that the ADR for 1 and 2 total people is around the same since it’s easy to share a room. Additionally, it makes sense that there is an increasing trend of ADR as the number of people increases. This will be included in our model

Total_stays and Lead_time:

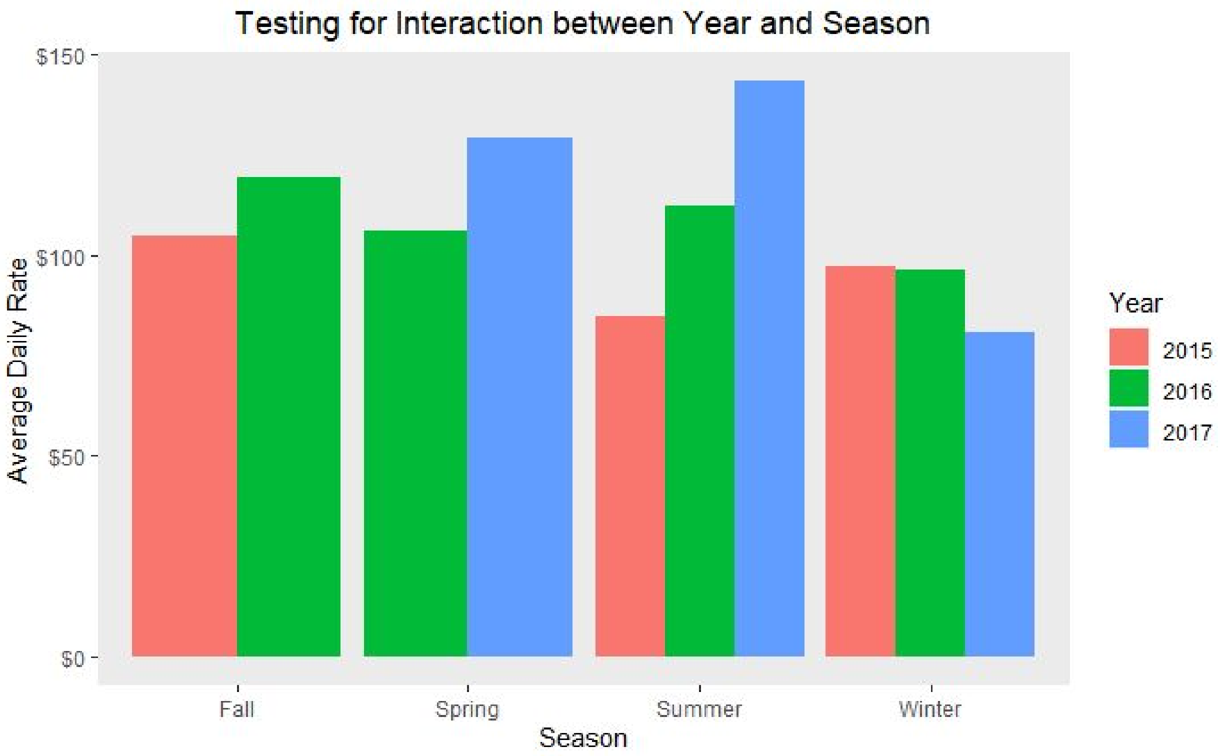


I didn’t realize this earlier but now it makes a lot of sense. It shouldn’t matter how long you stay because ADR is calculating AVERAGE daily rate. I initially thought that the higher the total stays, the lower the ADR but according to the correlation coefficient, total stays and ADR are not really correlated.

The same can be said about lead_time. Neither variables will be included in our model.

Section 2.4/2.5: Interactions and Non-Linear Trends

Season and Year:



Although we are missing a bit of data, there does seem to be some interaction between arrival year and arrival season. I think it’s a good idea to include this in our models in order to cover our bases.

Lead Time and Total Stays:

There are clear non-linear trends with these two variables but since we are not including either in our model, it shouldn’t matter too much.

Section 3: Models

Our final model will consist of the following variables:

- Categorical:
  - **is_canceled** - Value indicating if the booking was canceled (1) or not (0)
  - **arrival_season** - Season of arrival date
  - **arrival_date_year** - Year of arrival date
  - **meal** - Type of meal booked.
  - **market_Segment** - Market segment designation.
  - **reserved_room_type** - Code of room type reserved.

    - **customer_type** - type of customer; split into 4 categories.
- Numerical:
    - **total_people** - Number of people, excluding babies

Section 3.1: Simple Model

Our simple model will consist of all the variables we've selected but without any transformations and interaction terms. In other words, this is the simplest model we could possibly have at this point. This implementation is rather trivial.

After fitting this model we have the following results:

Adjusted R-squared: 0.6337

BIC: 15490.89

The Adjusted R-squared value is certainly a good sign, considering this model is very simple. I noticed that all of the market_segment values are not significant in the slightest so I will consider removing them for the final linear regression model. Nonetheless, this simple model is a good baseline.

Section 3.2: Linear Regression Model

Similar to the model in section 3.1, this model's implementation is also somewhat trivial. However, I will still explain it.

The main components of a linear regression model are that it's based on 4 assumptions.

1. Linearity
2. Homoscedasticity
3. Independence
4. Normality

So what this means is that we assume our model will have a linear relationship between the predictors and response, have constant error variance throughout the values of the predictors, have independent observations, and have a normal distribution of residuals.

Now I will be creating a more complex model to predict with. This model will use all of the variables from the previous model **except market_segment** and it will also contain an interaction term between Year and Season.

There were two methods that I used to choose the predictors for my model. The first of which was just creating plots and testing if there was a significant difference between predictors. The second of which was using the simple model to screen out the variables that should not be in the final model (market segment).

Our new results are:

Adjusted R-squared: 0.6392

BIC: 11642.17

Immediately we notice that our Adjusted R-squared and BIC are better than the simple model. This leads me to believe that removing the market segment predictor and adding an interaction term between season and year was a good idea.

Now, as for predicting, we can use the test data set we set up earlier to predict the ADR. Our results were an average of $118.92 daily rate.

Section 3.3: Random Forest

In our case, we will be “growing” 2000 decision trees and using the mean of all of the predictions to predict what our ADR will be. The benefit to growing all of these trees is that we will improve our level of accuracy and also limit the sensitivity to data change. The goal of this method is to decorrelate all of the trees so we can eventually come up with one predicted value.

So, after growing our forest, we come to a prediction of an average $120.50 ADR using the same test data as before.

It’s also interesting to note how the random forest considered the importance of different variables

Predictor	Importance
Reserved_room_type	736625
total_people	450918
arrival_season	172309
Is_canceled, meal, customer_type, arrival_date_year	All at or below 120000

If we take a step back, it makes sense that the type of room, amount of people, and season would be the most important

Section 4: Discussion

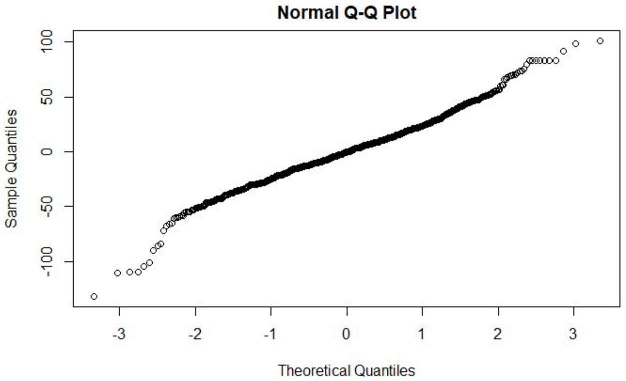
Now that we have both predictions, can we decide which model is better? One way we can compare them is to use the RMSE.

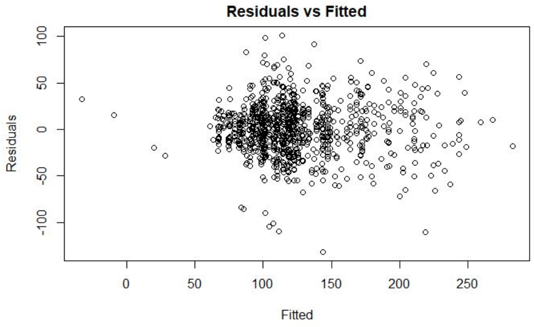
Model	RMSE
Model from 3.2	23.42904
Random Forest	27.40389

The standard deviation of ADR in our test set is 44.02. This means both of our model's RMSE are good.

Off the bat, we can see that the linear regression model has a lower RMSE than the random forest. This is not what I expected but it's also not an unusual result. Maybe growing more trees would lower the RMSE but I think 2000 is a good amount. From these results, we have almost no reason to use the random forest model over the linear regression model. The MLR model is easier to compute and easier to understand.

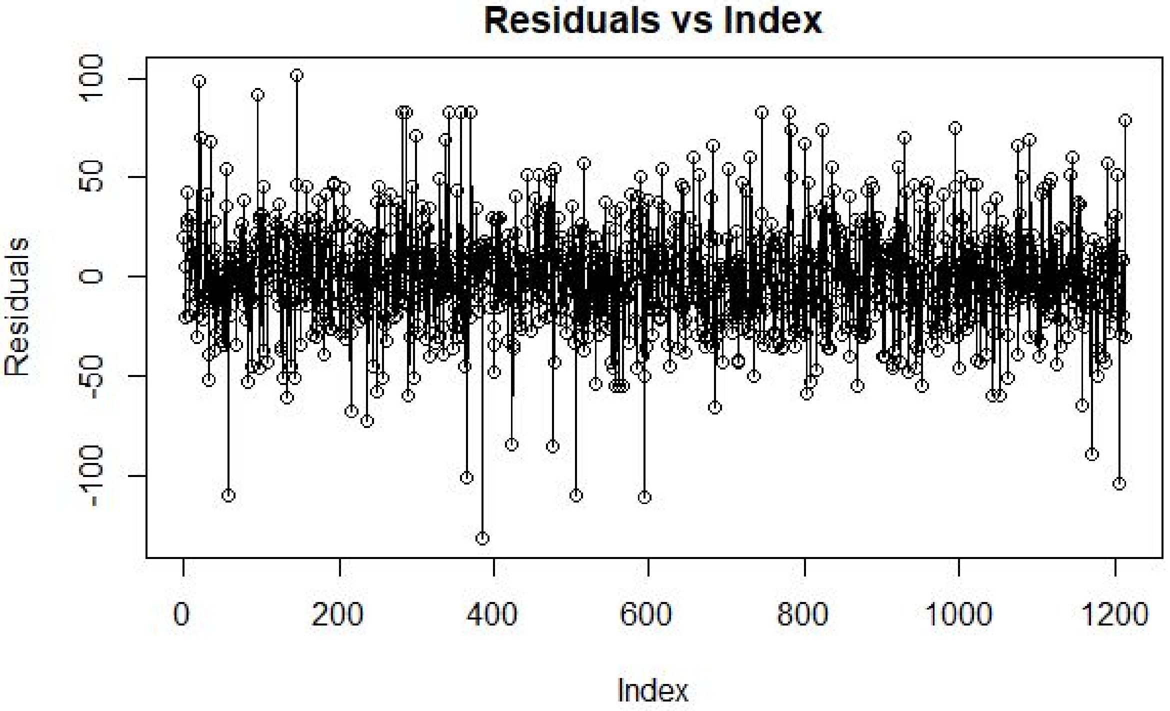
Let's see if the MLR satisfies it's assumptions. To begin with, let's create a normal qq plot and a residuals vs fitted plot.





We can see some fluctuation towards the tails but with a dataset this big, I think our model satisfies the normality assumption.

The same reasoning can be applied to the Residuals vs Fitted plot. Yes, it's not perfect but it is enough to pass the constant variance assumption.



Finally, our Residuals vs Index plot helps us confirm that the errors are not correlated. To help strengthen our argument, I've also conducted a Durbin-Watson test which returned a P-value of 0.75, thus confirming our results.

So, after 14 pages of analysis, we've concluded that the linear regression model is slightly better than the RF. Maybe if I used some other variables in my analysis it would have led to a different result. Specifically, I could have included the lead_time variable but ultimately decided to leave it out for the sake of simplicity.

Overall, however, I am satisfied with the variables I've chosen. I think most of them have a solid explanation behind them and if I were to present them to an employer, they would not be challenged. I think someone can realistically use this model to get a strong prediction of ADR of a city hotel.