Hotel Booking Demand: STA 210 Project Report

Mean Girls 2: Martha Aboagye, Raymond Chen, Maria Henriquez, Calleigh Smith

April 29, 2020

Introduction

Motivation

When talking about our group interests, it became clear that each of us has a passion for travel and other cultures. A lot of our co-curricular activities at Duke have been travel related, whether that be through DukeEngage, study abroad, athletics, or other club activities. We identified early on that we wanted to choose a data set that involved the travel industry and were interested in gaining insights that would lead us to be more informed explorers. We identified two priorities that hotels have:

1) They need to maximize revenue by optimizing the amount of rooms that are occupied.
2) They need to ensure that they have sufficient resources, such as staff members, food, and security, to meet the demand of hotel guests.

In order to achieve these goals, it is important that hotels proactively use the information they have available from their reservations to predict which guests are likely to cancel.

As a result, we decided to investigate a hotel data set, which has information about hotel booking demand. More specifically, this data set contains information regarding two hotels in Portugal. More information on the data can be found below.

Research Question and Relevant Variables

Our research question is:

> Can we predict whether a customer is going to cancel their hotel booking?

To answer this question, we will build a logistic regression model and see which variables are significant in predicting cancellations. The response variable for our logistic regression model is `is_canceled`, which is a binary categorical variable that has values of either 0 or 1. 1 corresponds to a cancelled booking and 0 corresponds to non-cancelled bookings. The explanatory variables we will use for our logistic regression model are the following:

- `adr`, the average daily rate of a booking
  - We predict that a booking is less likely to be cancelled if it has a higher daily rate.
- `adults`, the number of adults per booking
  - We expect that if there are a lot of adults on a booking, then the probability of a cancellation is likely greater because there could be more conflicts.
- `arrival_date_month`, the month of an arrival (we will use this to make a new variable, `season`)
  - We expect that season will affect whether a booking is cancelled.
- `babies` + `children`, the number of babies and children per booking, respectively (we will use these to make a new variable `infants`, which indicates if a booking included babies or children)
  - We expect that a booking with more infants on the reservation will increase chances of cancellation, as infants are unpredictable and could get sick suddenly.
- `booking_changes`, the number of changes made to the booking before check-in (we will use this to make a new variable, `changed`, which indicates whether a booking had 1 or more `booking_changes` or was unchanged)
  - We expect the probability of cancellation to be higher for more booking changes, as the customer might be indecisive or have concerns about traveling.

- `country`, the country that a customer is from (we will make a variable `origin` that classifies a customer as being a domestic or international traveler)
  - We expect that whether a customer is from Portugal or has to travel internationally will affect cancellation.
- `hotel`, whether the hotel is the city or resort hotel
  - We expect that the city hotel will have more cancellations, as people travel to the city for business or events, which are likely to change more frequently.
- `is_repeated_guest`, whether a guest is a repeated guest or not
  - We expect that repeated guests will be more loyal to the hotel and not cancel their bookings.
- `lead_time`, the amount of days between when a booking was made and when the guest checked-in
  - We expect that there will be more cancellations for bookings with high lead times, as unexpected circumstances might arise that would lead to a booking needing to be cancelled.
- `market_segment`, the market segment designation of a guest
  - We expect that the market segment of a guest might impact cancellation.
- `previous_cancellations`, the number of previous cancellations a guest has made (we will use this to create a new variable, `prior_cancellation`, that indicates whether the client has ever cancelled a booking at the hotel prior to the current booking)
  - We expect guests with previous cancellations to be serial cancelers or more likely to cancel than guests who have never cancelled.
- `stays_in_weekend_nights` + `stays_in_week_nights`, the number of weekend nights and week nights, respectively, in a booking (we will use these to make a new variable, `length_stay`)
  - We expect that longer stays will be less likely to be cancelled, as these bookings probably have more planning that go into them.

Dataset Description

`hotels` contains hotel booking data. There are 119,390 observations, each representing a single hotel booking. The bookings were retrieved from one city hotel in Lisbon, Portugal, and one resort hotel in Algarve, Portugal. Both data sets include bookings due to arrive between July 1, 2015, and August 31, 2017, including bookings that effectively arrived and bookings that were canceled.

The data set includes 32 variables of interest, including information about each booking, such as whether the booking was cancelled (`is_cancelled`), the customer's meal plan (`meal`), room type (`assigned_room_type`), when the booking was made (`arrival_date_year`, `arrival_date_month`, `arrival_date_day`), the length of the stay (`stay_weekend_nights`, `stay_week_days`), the number of adults, children, and/or babies (`adults`, `children`, `babies`), and the number of required parking spaces (`required_car_parking_spaces`), among others. A detailed description of all 32 variables can be found in the code book.

The data was originally from the article "Hotel Booking Demand Data sets," written by Nuno Antonio, Ana Almeida, and Luis Nunes for the journal *Data in Brief*, Volume 22, which was published in February 2019. The data set we will be using was cleaned by Thomas Mock and Antoine Bichat during the week of February 11th, 2020 for #TidyTuesday.

In the article "Hotel Booking Demand Datasets," the authors state that the data was extracted from the hotels' public Property Management System (PMS) databases' servers "by executing a TSQL query on SQL Server Studio Manager, the integrated environment tool for managing Microsoft SQL databases." The article assured that there was no missing data in the data sets constructed. Since this is real hotel data, all identifying data elements about the hotel or the customer were deleted.

While the primary table used to compile the data set was "Bookings," the researchers joined the "Bookings" table with other tables, including "Bookings change log," "Meals," "Distribution Channels," "Transactions," "Customer Profiles," "Nationalities," and "Market Segments," to get a more complete picture of the variables that affect bookings.

The documentation for the original data set can be found here:
https://www.sciencedirect.com/science/article/pii/S2352340918315191

Data Wrangling

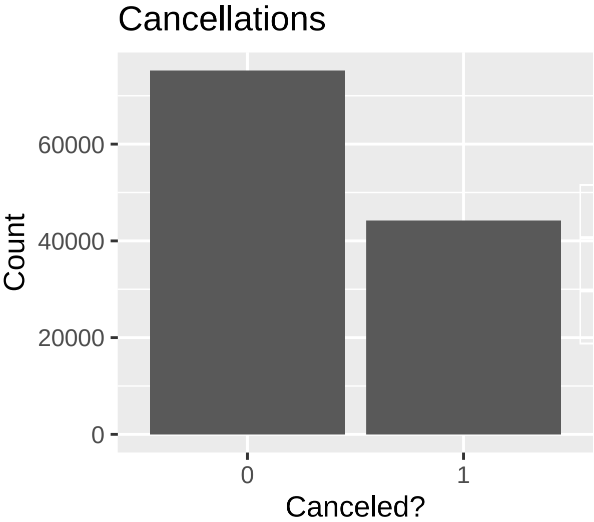
As discussed above, we needed to create a few variables to use for our analysis, specifically `infants`, `season`, `length_stay`, `prior_cancellation`, `origin`, and `changed`. We ensured that our variables are of the correct type, specifically that binary categorical variables are made factors (`is_canceled`, `infants`, `is_repeated_guest`, `prior_cancellation`, `hotel`). We removed the market segment “Undefined,” since it only belongs to 2 observations in the data set. After making these changes, we created a smaller subset of the data that included only the variables of interest for our analysis (as specified above).

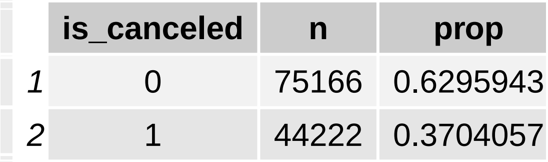
Exploratory Data Analysis

Before we continue with finding a model, we will complete an exploratory data analysis to understand the variables we are dealing with.

Univariate Analysis

`is_canceled`

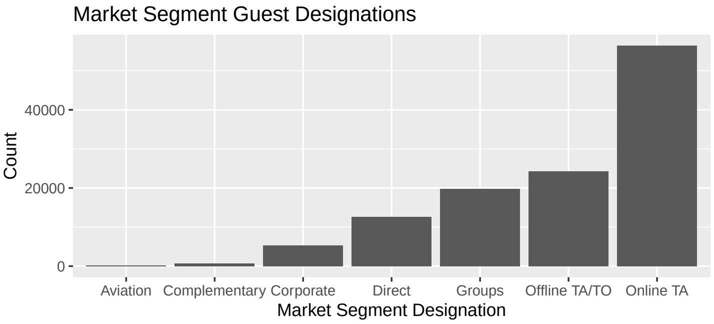


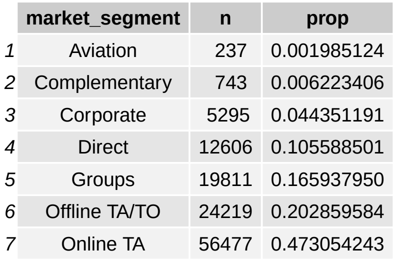


	is_canceled	n	prop
1	0	75166	0.6295943
2	1	44222	0.3704057

`is_canceled` is the response variable for our logistic regression model. From the distribution and the table above, we see that approximately 37% of the booking observations in our data sample were canceled.

`market_segment`

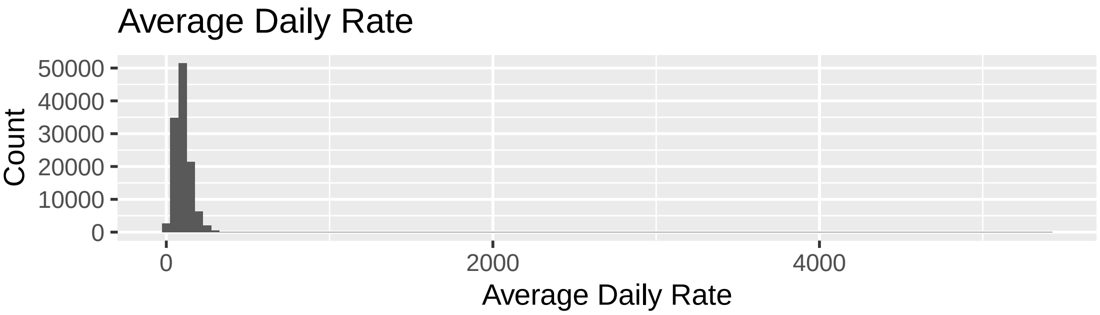




	market_segment	n	prop
1	Aviation	237	0.001985124
2	Complementary	743	0.006223406
3	Corporate	5295	0.044351191
4	Direct	12606	0.105588501
5	Groups	19811	0.165937950
6	Offline TA/TO	24219	0.202859584
7	Online TA	56477	0.473054243

From the bar chart of `market_segment`, which is the market segment designation for a given booking, we can see that the vast majority of bookings in the data set (47.30%) have a market segment designation of Online TA (Travel Agent), followed by Offline TA/TO, then Groups, then Direct, then Corporate, then Complementary, and finally Aviation (0.20%).

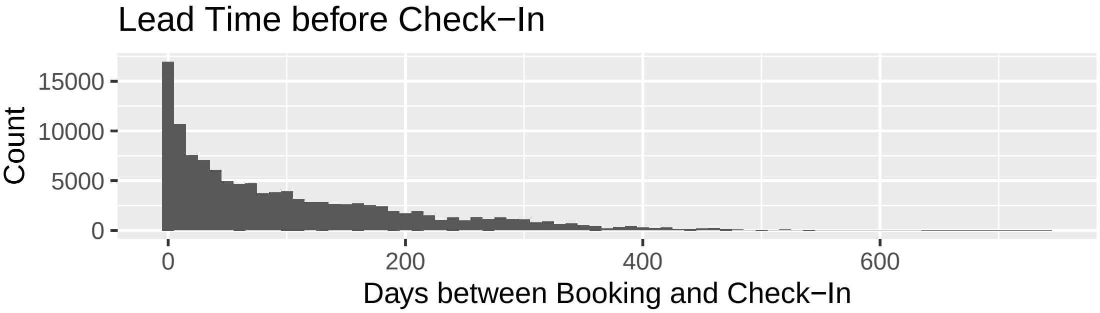
adr



	min	q1	median	q3	max	IQR	loweroutlier	upperoutlier
1	−6.38	69.29	94.59	126	5400	56.71	−15.775	211.065

From the graph of `adr`, the average daily rate, we can see that the distribution is skewed right. The center (median) is at 94.59 euros. The spread is about 56.71 euros, which is not that large. However, it appears that there are outliers towards the higher end of the spectrum (above 211.07 euros). The maximum for the distribution is 5,400 euros, which is well above the normal range of average daily rates. For our analysis, we will only consider average daily rates that are not outliers (< 211.07 euros), as we are interested in predicting cancellation for the average client.

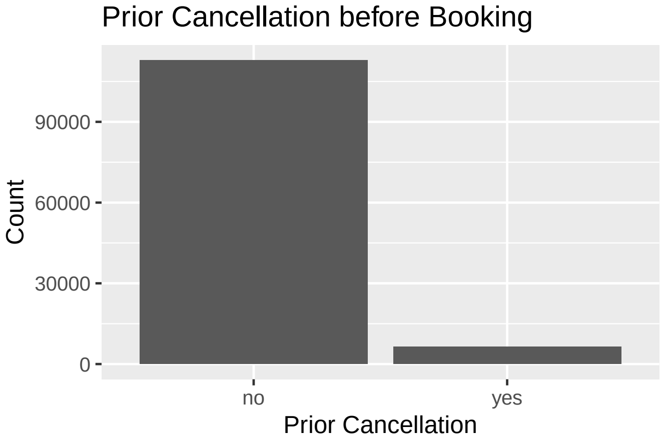
lead_time



	min	q1	median	q3	max	IQR	loweroutlier	upperoutlier
1	0	18	69	160	737	142	−195	373

From graph of `lead_time`, the number of days between booking the hotel and checking in, we can see that the distribution is skewed right. The center (median) is at 69 days, meaning that people normally book their hotels a little over 2 months in advance. The spread is 142 days, which means that there is a decent amount of variability in terms of how far in advance a client reserves a space at the hotel. However, it appears that there are outliers towards the higher end of the spectrum (above 373 days (approximately a year)). Furthermore, it's a bit surprising to see that a great number of bookings were made on the same day as the check-in. This is probably not unusual for a city hotel, but for a resort, it might be. The maximum for the distribution is 737 days (approximately 2 years). We would only like to consider bookings with a lead-time of a year, or 365 days.

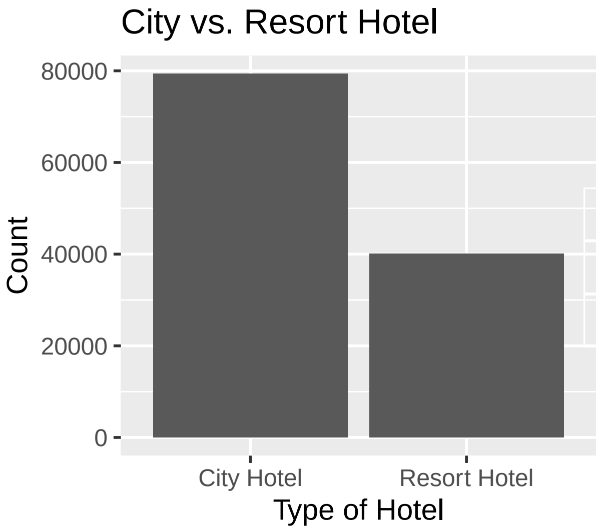
prior_cancellation



	prior_cancellation	n	prop
1	no	112904	0.94568968
2	yes	6484	0.05431032

From graph of `prior_cancellation`, an indicator of whether the client who reserved a booking had previously cancelled a booking before the current booking, we can see that the vast majority of bookings (94.57%) were reserved by clients who had never cancelled a booking at the hotel before.

`hotel`

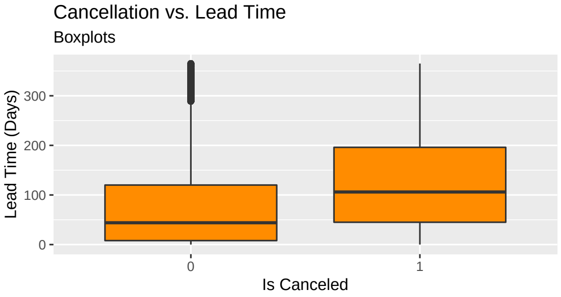


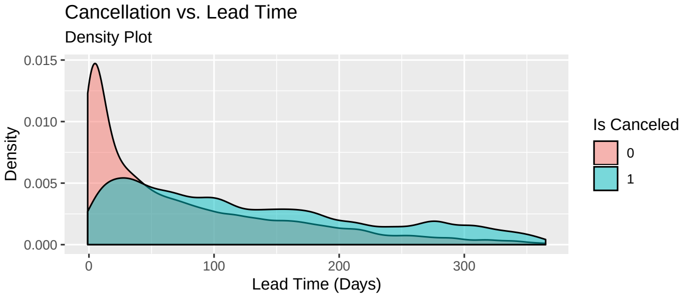
	hotel	n	prop
1	City Hotel	79328	0.6644554
2	Resort Hotel	40060	0.3355446

From the bar chart of `hotel`, the type of hotel the booking was made for, we can see that the vast majority of bookings in the data set (66.45%) were made for the city hotel, while the remaining bookings (33.55%) were made for the resort hotel.

Bivariate Analysis

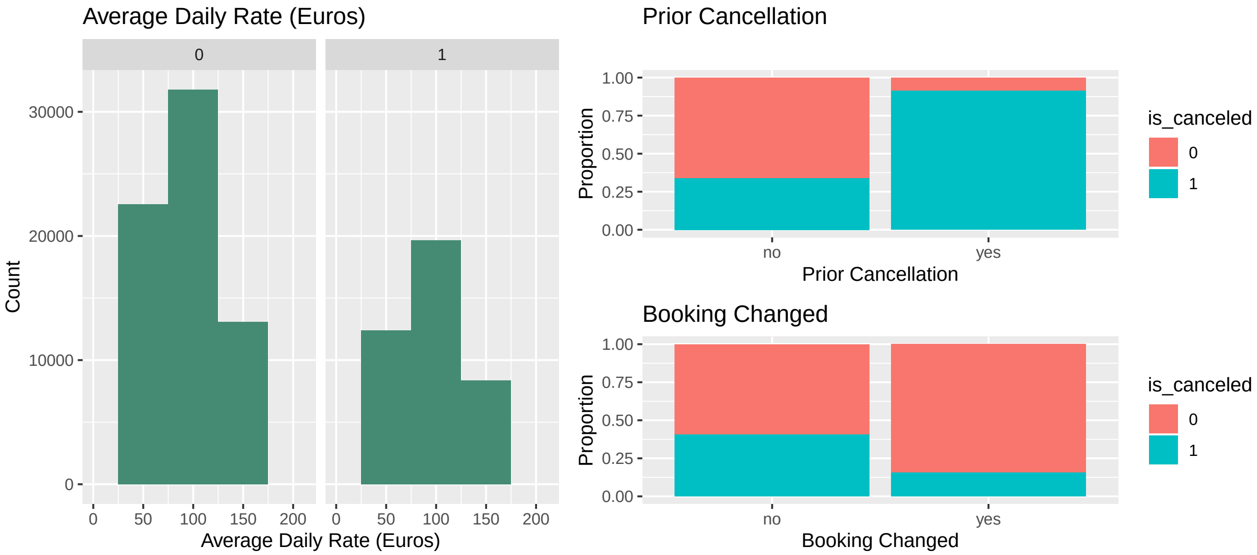
`lead_time vs. is_cancelled`





The boxplot of `lead_time` vs. `is_canceled` shows that bookings that were canceled had a larger median `lead_time`. Canceled bookings also had a larger IQR. The histogram and density plot of `lead_time` by `is_canceled` is right skewed for both canceled and non-canceled bookings. The distribution for `lead_time` for non-canceled bookings is more right skewed skewed and has a smaller spread than the `lead_time` for non-canceled submissions. Most of the `lead_time` for non-canceled bookings is between 0 and 200 days while for canceled submissions, the `lead_time` is usually between 0 and 400 days.

`adr`, `prior_cancellation`, and `changed` vs. `is_canceled`

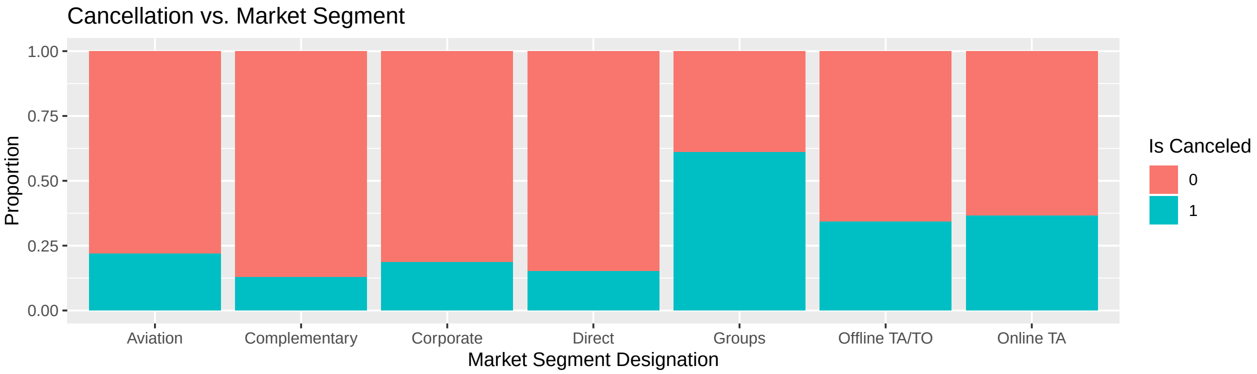


The histogram of average daily rate by `is_canceled` shows that the distribution of average daily rate is roughly the same across canceled and non-canceled reservations. For both canceled and non-canceled reservations, the average daily rate is typically between 0 and 250 euros with the most frequent average daily rate being about 100 euros.

The bar chart of `prior_cancellation` by `is_canceled` shows that bookings reserved by clients who have previously cancelled a reservation at the hotel are significantly more likely to be canceled compared to bookings reserved by clients who have never canceled a reservation before.

The bar chart of `changed` by `is_canceled` shows that a booking is more likely to be canceled if a booking has never been changed compared to if a booking was modified.

`market_segment` vs. `is_canceled`



The bar graph of `is_canceled` vs. `market_segment` shows that the market segment that has the highest proportion of canceled reservations are Groups. Next, Online TA and Offline TA/TO have the third highest proportion of canceled reservations, although for those market segments, reservations are still more likely to not be canceled than canceled. Direct, Corporate, Complementary and Aviation market segments have proportions of about a 25% cancellation rate.

Clean Data Further

We filtered the observations to exclude unnecessary outliers in the following variables: `adr < 211.06`, `1 <= adults <= 10`, and `lead_time <= 365`. We chose to fit our model only on these filtered observations because we want our model to best apply to the most "likely" case. We encourage future work to consider building a model for only extreme cases in variables.

We also mean-centered three variables for interpretation purposes: `adults`, `adr`, and `length_stay`.

Our data set is also extremely large (about 120,000 observations). Therefore, we took a random sample of 10,000 observations from the original data set to make our analysis more efficient and generalizable. We will use the random sample of 10,000 bookings to build our model.

Regression Analysis

Model

Due to the fact that we have a categorical response variable with two levels (`is_canceled`), we will fit a binary logistic regression model.

We began by fitting a full model with 12 variables and 4 interaction variables. We considered the following 4 interaction effects:
`lead_time` x `changed`
`season` x `adr`
`hotel` x `adr`
`season` x `lead_time`

We chose to explore these four interactions because we believed from previous knowledge that out of all the variables we were exploring, these variables seemed the most likely to relate to one another.

We then performed a backwards selection with BIC as the criterion. We chose BIC as the selection criterion because we have a lot of variables and want to strictly penalize for any variables that are not truly necessary. Since our model is intended to be used by hotel personnel, we wanted to keep it as simple and efficient as possible.

The backwards selection using BIC removed `adultsCent`, `infants`, and all of the interaction variables.

The variable `lead_time` was not removed in the backwards selection, but the estimated coefficient was extremely small and had a negligible impact on cancellations. Because we did not believe it was practically significant, we decided to remove the variable from the model. `adr` also had a small coefficient, but we are more interested in exploring this variable's impact on predicting cancellations, so we made the executive decision to include it in the model.

Below is the final model after BIC selection and removing `lead_time`.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.098	0.677	-0.145	0.884	-1.624	1.112
adrCent	0.008	0.001	11.139	0.000	0.007	0.010
seasonSummer	-0.206	0.066	-3.121	0.002	-0.335	-0.077
seasonFall	-0.274	0.069	-3.938	0.000	-0.410	-0.138
seasonWinter	-0.081	0.077	-1.054	0.292	-0.232	0.070
changedyes	-1.225	0.086	-14.251	0.000	-1.396	-1.059
origininternational	-1.855	0.060	-31.048	0.000	-1.973	-1.739

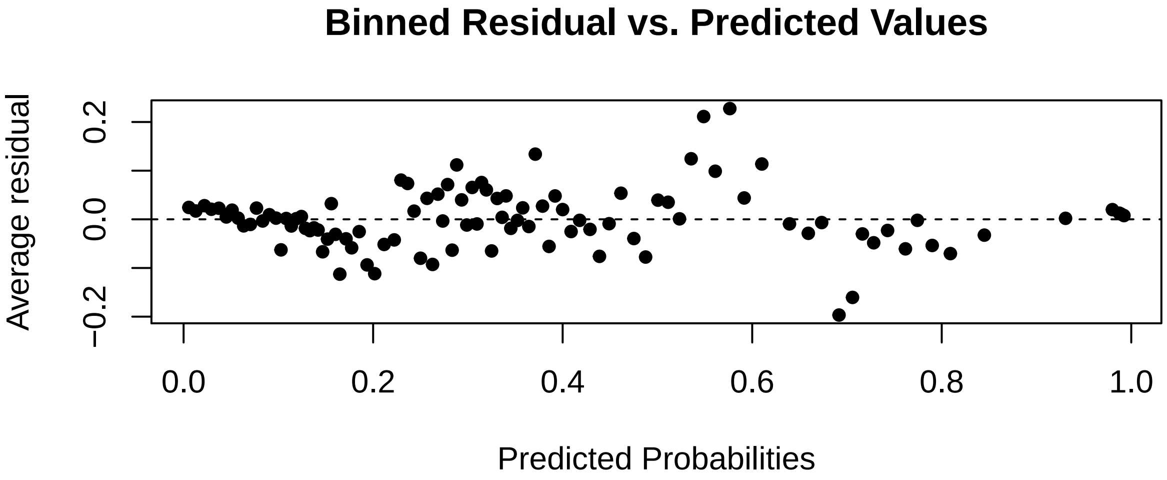
term	estimate	std.error	statistic	p.value	conf.low	conf.high
hotelResort Hotel	-0.658	0.059	-11.123	0.000	-0.774	-0.542
is_repeated_guest1	-2.744	0.313	-8.773	0.000	-3.389	-2.159
market_segmentComplementary	-0.529	0.813	-0.651	0.515	-2.081	1.191
market_segmentCorporate	-0.272	0.692	-0.393	0.694	-1.518	1.277
market_segmentDirect	-0.169	0.683	-0.248	0.804	-1.394	1.366
market_segmentGroups	1.462	0.679	2.154	0.031	0.247	2.992
market_segmentOffline TA/TO	0.534	0.678	0.788	0.431	-0.679	2.062
market_segmentOnline TA	1.283	0.677	1.895	0.058	0.072	2.810
prior_cancellationyes	3.793	0.267	14.229	0.000	3.301	4.352
length_stayCent	0.110	0.011	10.450	0.000	0.090	0.131

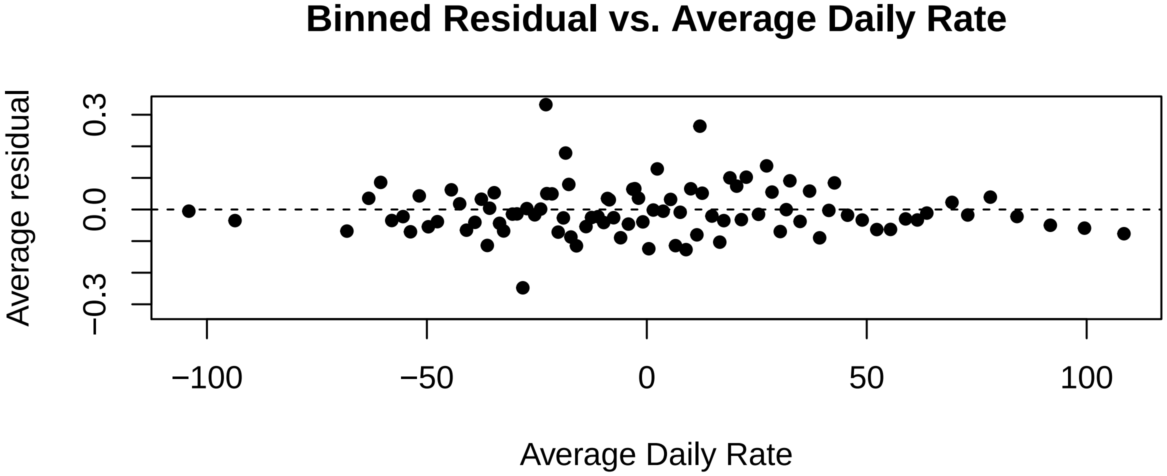
Model Assumptions

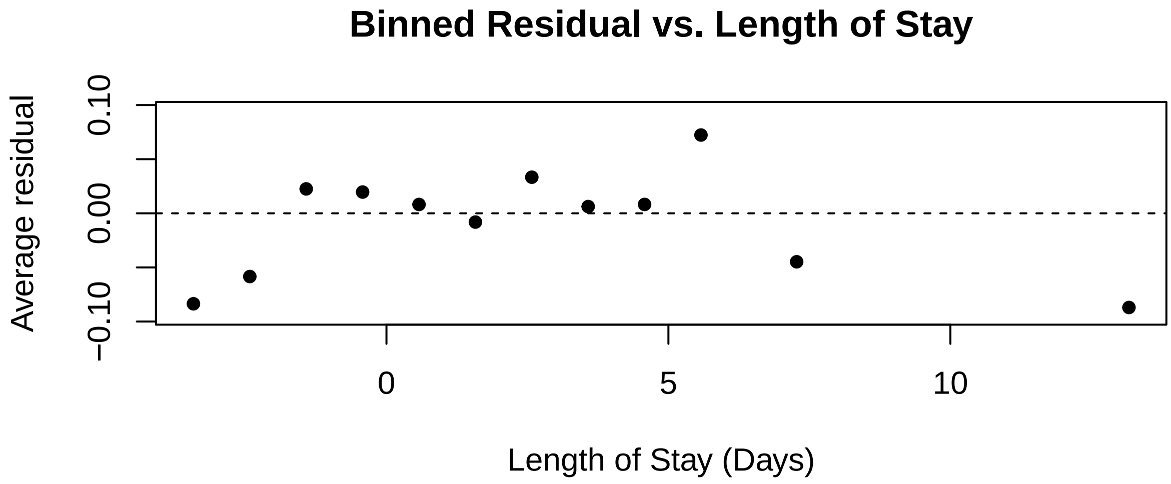
Now we will check the model assumptions for logistic regression, which include linearity, randomness, and independence.

Linearity

In order to make a conclusion about linearity, we must look at the binned residuals vs. the predicted probabilities, the binned residuals vs. the quantitative predictor variables, and the mean residuals for the categorical variables. We will start with the binned residual plots.







Based on the binned residual plots, it seems evident that the linearity condition is satisfied in all of the plots. None of the plots show signs of any obvious pattern or shape, and the points are all scattered randomly about the horizontal line where `average residual = 0`.

Now let's look at the average residuals for the categorical variables.

season	mean_resid
Spring	0
Summer	0
Fall	0
Winter	0

origin	mean_resid
domestic	0
international	0

hotel	mean_resid
City Hotel	0
Resort Hotel	0

market_segment	mean_resid
Aviation	0
Complementary	0
Corporate	0
Direct	0
Groups	0
Offline TA/TO	0
Online TA	0

prior_cancellation	mean_resid
no	0

prior_cancellation	mean_resid
yes	0

changed	mean_resid
no	0
yes	0

We can see that for our categorical variables all of the mean residuals are extremely close to 0, which is what we want for linearity to be satisfied.

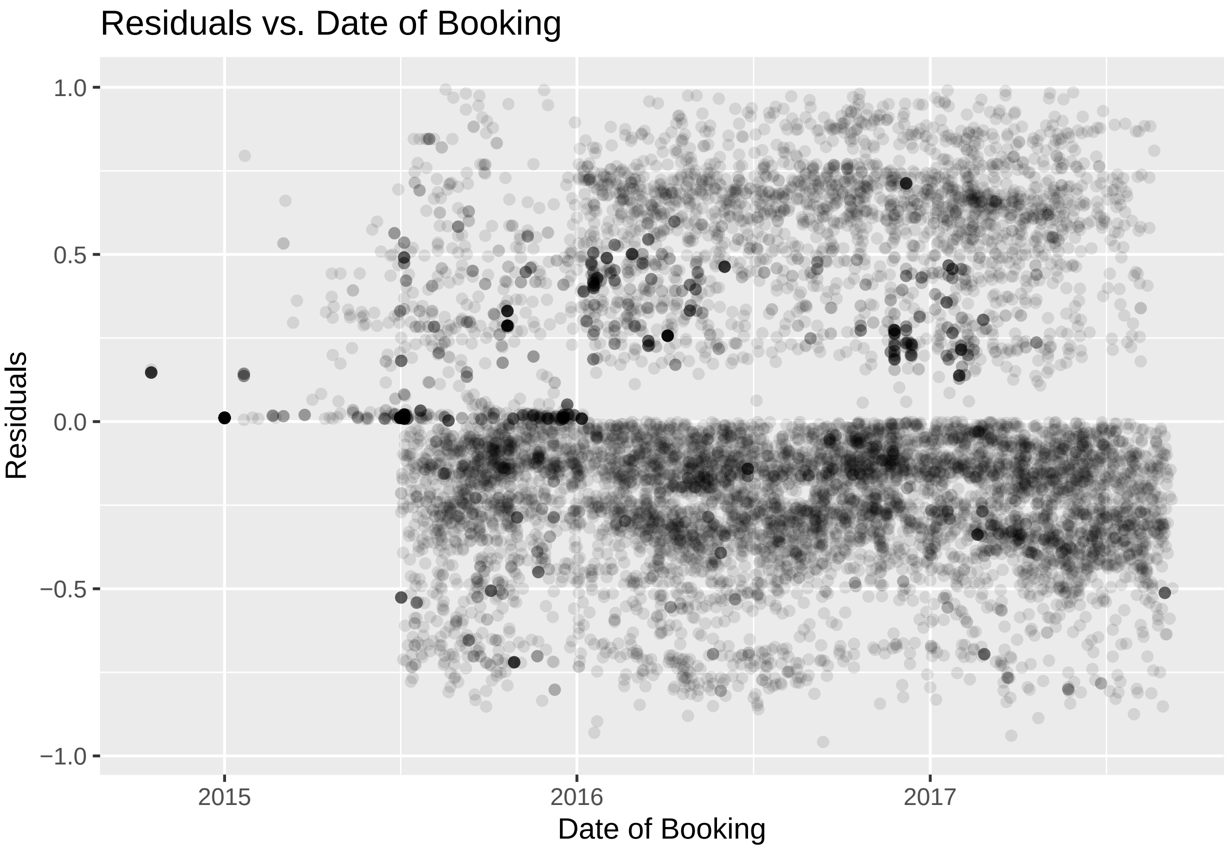
Since the binned residual plots and the tables of the average residuals generally show no departures from linearity, we believe it is safe to assume that the linearity condition is satisfied.

Randomness

In order to assess randomness, we must look at how the original data was collected. We know that all of the bookings come from only 2 hotels in Portugal, 1 resort hotel and 1 city hotel. This certainly is a specific subset of hotels. Due to this, our model is not as generalizable as we would like it to be and can really only be generalized to hotels in Portugal. However, the bookings that are included in the data set are ALL bookings at the two hotels between July 1st, 2015 and August 31st, 2017. We took a simple random sample of 10,000 bookings from the original "population" data set. Since our model does not isolate or exclude bookings pertaining to any given group and is built from a simple random sample, the randomness condition is satisfied for our data set.

Independence

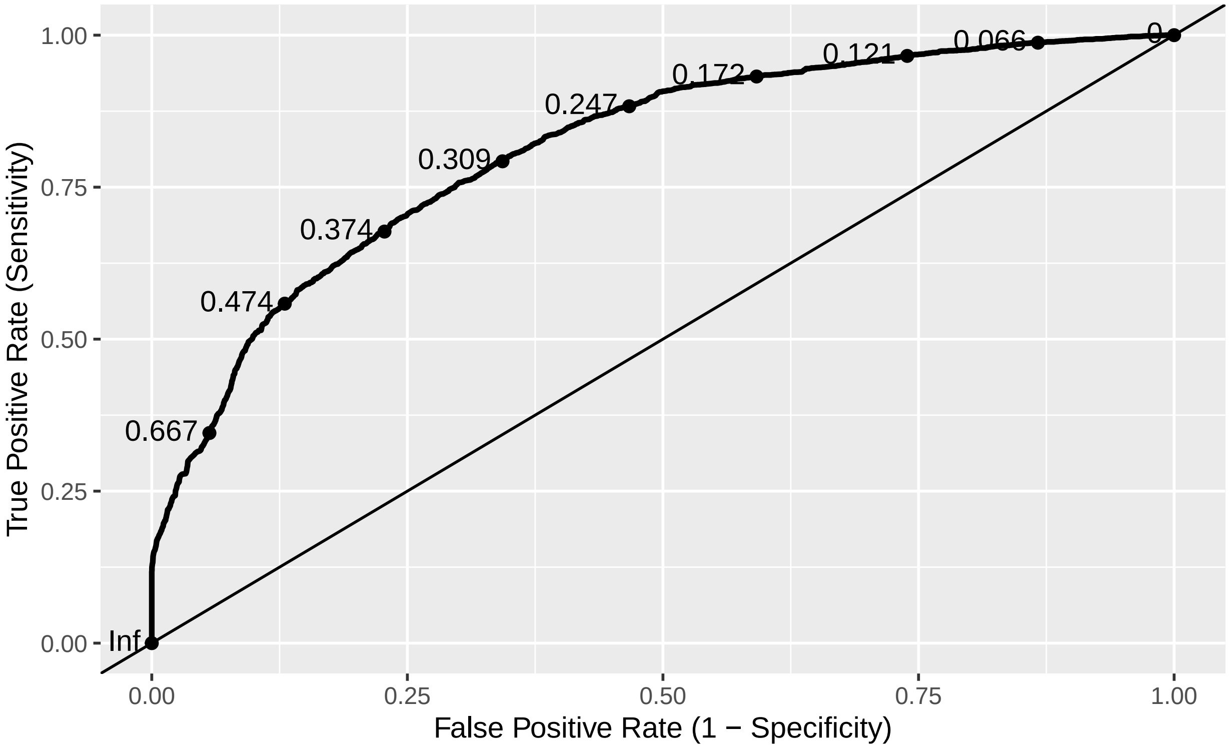
We believe that the hotel bookings in our data set are independent. Since there are a large number of bookings and since the bookings are from all different kinds of guests, we have no reason to believe that the bookings are related to each other or that one booking affects another booking. Since our data set has a time dependency (all bookings have a date associated with them), we chose to explore the scatter plot of the residuals vs. date.



Based on the scatter plot, it does not look like the serial effect is a factor in our data set. The scatter plot does not have any obvious pattern like a sinusoidal shape or trend that would indicate that cancellations are dependent on time. It is clear from the plot, however, that the volume of bookings increases over time and that our model tends to over-predict the probabilities of cancellations for bookings since the residuals (observed - predicted) are concentrated below `residuals = 0`.

Model Fit Statistics

Below is the ROC curve and the calculated AUC for the final model.



```
## [1] 0.810783
```

Discussion

The AUC for our model was 0.810783. This means that about 81.0782996% of the time, our model is able to correctly identify whether a booking will be canceled or not. Our AUC value is close to 1, and is thus a good indication that our model can sufficiently predict if a hotel booking will be cancelled.

Based ROC curve, we believe that the best threshold value for predicting whether a reservation will be cancelled or not is about `0.31`. In choosing our threshold, the most important factor we considered was the practical financial implications of having large scale unanticipated cancellations. We assume that for hotels, being able to roughly estimate the amount customers who might cancel will allow them to better fill vacancies and maximize profits. In addition, being able to determine bookings that will be cancelled will allow hotels to build infrastructure that allows them to process refunds or update their booking systems. At the .31 threshold, our model accurately identifies over 75% of the bookings that end up being canceled. It also has about a 25% false positive rate at that threshold, which means that about 25% of the time, it incorrectly identifies bookings which won't be canceled as canceled. If a hotel wants to identify which individual reservations might be canceled, a lower threshold would be more appropriate so that the true positive rate increases. Of course, that does come with the sacrifice of more false cancellation flags. Regardless, we assume hotels are primarily concerned with estimating the amount of reservations that might get canceled. Towards that end, a `0.31` threshold likely provides an accurate result a majority of the time while minimizing the false positive rate.

The coefficient for our intercept is `-1.496`. This means that we expect the odds of a booking being cancelled for a city hotel booking made the day of (`lead_time` = 0) in the Spring with an average daily rate of 97.80 euros made for 3.42 days and for 2 adults by someone who is not a repeated guest, has made no booking changes, is from Portugal, and has made no previous cancellations to be 0.2240245.

For our final model, predictors that increased the probability of a reservation being cancelled are all levels of `market_segment` compared than the baseline (Aviation), if a guest had previously cancelled a reservation at the hotel (`prior_cancellationyes`) compared to if they had not, a longer length of the guest's stay (`length_stayCent`), and a higher average daily rate (`adrCent`). The predictors that decreased the probability of a reservation being cancelled were all `season` levels compared to the baseline (Spring), if the guest was a returning guest (is_repeated_guest1) versus if they were not a returning guest, if the booking was made for the resort hotel (hotelResort Hotel) versus if it was made for a city hotel, if the guest's origin was international (origininternational) versus if it was domestic, and if a guest changed their booking prior to checking in (changedyes) versus if they didn't change. The predictors that weren't statistically significant predictors of whether a reservation would be cancelled were infants and adults. We initially hypothesized that as the quantity of adults and infants increase, the probability of a reservation being cancelled would also increase. Although BIC selection did not remove lead_time from our model, it had a coefficient of nearly 0, which we decided was not practically significant in predicting the odds of a cancellation.

The results of our model also do not support our initial hypothesis for adr. Our model predicts that for every additional increase in adr by 1 euro, the odds of a reservation being cancelled will multiply by a factor of 1.0090406, making reservations with higher daily rates more likely to be cancelled. This might be because a guest found a better deal at different hotel, and so they chose to cancel their original booking.

For the categorical variables, we hypothesized that the levels hotelCity Hotel, prior_cancellationyes, origininternational, changedyes, and is_repeated_guest_no would increase the probability of a reservation being cancelled. For hotel, we hypothesized as more people travel to the city for business or events, so their plans are more likely to change. For prior_cancellation, we thought that if a customer has cancelled a reservation before, they would be more likely to cancel a reservation in general. For origin, we thought that international reservations would have a lower probability (compared to domestic) of being cancelled because their bookings had to have been more thought out. Finally, for changed, we thought that a reservation being changed multiple times served as a good indication of a customer's plan being uncertain. We hypothesized that if details of a reservation had been changed, the reservation would have a higher probability of being cancelled.

Our model provided evidence to support our initial hypothesis for hotelCity Hotel, is_repeated_guest_yes, prior_cancellationyes, and origininternational. With city hotel as the baseline level, the coefficient for hotelresort Hotel has a coefficient of -0.580. This means that holding all else constant, the odds of cancellation for a resort hotel is 0.5598984 times the odds of cancellation for a city hotel. is_repeated_guest_yes has a coefficient of -3.795. This means that holding all else constant, the odds for a reservation being canceled for a repeat guest is exp(-3.795) times the odds of a booking being cancelled for a non-repeat guest. prior_cancellationyes had a coefficient 4.242. This means that holding all else constant, the odds of cancellation for guests with previous cancellations is 69.5468065 times the odds of cancellation for guests with no previous cancellations. The odds of an international reservation being cancelled is 0.1398759 times lower than the odds of a domestic reservation.

Our model surprisingly did not provide any evidence to support our initial hypothesis that changedyes would increase the probability for a reservation being cancelled. changedyes had a negative coefficient of -1.049, which means that changed reservations are less likely to be cancelled. Specifically, the odds of a reservation being cancelled is 0.3502879 times the odds of cancellation for a reservation that hasn't been changed. We thought that changed reservations were an indicator of a guest's uncertainty in their plans, but perhaps changed reservations are actually a sign that the guest is committed to staying at the hotel, as people who have concerns about their reservation might just cancel it rather than making amendments.

Limitations

Through our analysis, we have generated a logistic regression model to predict hotel cancellations. Again, it is important to note that our data was sourced from limited locations, particularly one city hotel and resort, both located in Portugal. Thus, our model should only be used to generalize to hotels with Portugal at most. The model will be best applicable to just the two hotels that our data came from. In fact, we progressed with our analysis with this at mind. Specifically, we took this in as a major consideration when deciding a threshold for our logistic model. We wanted to take into account the financial implications of hotel cancellations in order to weigh the costs and benefits of utilizing our model.

Given that our model is based on data from only two hotels, we would have liked to source data from more locations. For example, if we had data from more hotels of various types within Portugal, we could likely extend our model to create cancellation predictions for hotels in Portugal as a whole. And beyond that, if we had data from hotels around the world, we could further extend our model beyond just Portugal too.

In addition, some changes or improvements we would like to consider include utilizing more interaction effects. Interaction effects are difficult to judge and intuition in itself might be insufficient to decide on these effects. Thus, a future direction could include testing various, if not all, the possible interaction effects (assuming sufficient processing capabilities) to improve our model. We would of course then again perform a backwards selection process to ultimately decide on what terms may or may not be statistically significant.

It's also interesting to note that backwards selection using BIC removed `infants` from our model. It appears that having children or babies truly didn't have a statistically significant impact on cancellations, as the test has shown us. However, we also must consider that we did construct this variable ourselves from the data. It is a combination of the count of children and babies in the booking, and thus that union might have an altered significance compared to that of those terms pre-merge.

A reason that BIC might not have taken out many variables is because we had so many observations in the data set. We know that if we have a very large number of observations and we run hypothesis testing with the null hypothesis that the slope of a particular variable is 0, then no matter how small the slope truly is and no matter what significant level we assign, when n is sufficiently large, we will always reject the null hypothesis.

Conclusion

Through our analysis, we believe that we have created a sufficient logistic regression model that can be utilized to predict hotel cancellations. Specifically, our model can be generalized to two specific unnamed hotels in Portugal, one of which is a resort and the other is a hotel.

We initially isolated 12 predictor variables and 4 interaction terms of interest that we believed would be important in possibly predicting cancellations for these hotels. With these variables, we constructed a logistic linear model and performed backwards selection using BIC as the selection criterion to determine which variables we would ultimately decide to keep in our finalized model. A few predictor variables and all interaction terms were removed. Furthermore, we decided another term was too small to be practically significant. Thus, we ultimately landed on a model with 9 predictor variables and no interaction terms.

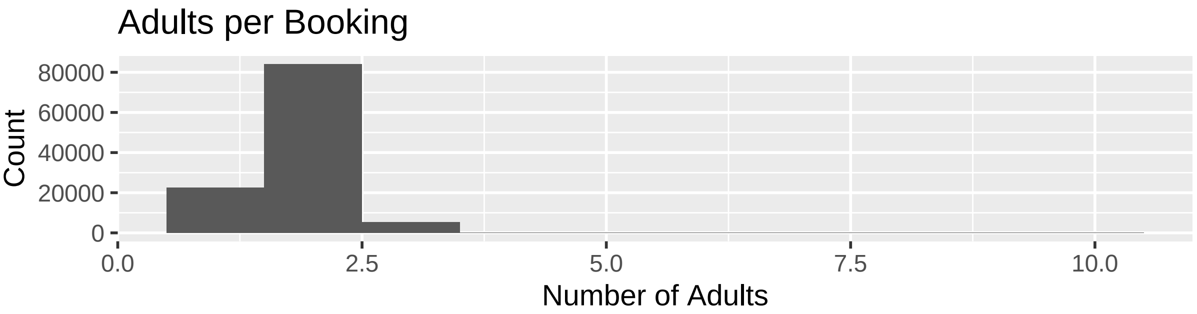
We were unsure what to expect when determining the model's accuracy, but we were pleasantly surprised after constructing a ROC curve and calculating the AUC. We found that our model was able to correctly differentiate cancellations about 81.08% of the time. Therefore, we believe the model we have devised could definitely be of practical use to those hotels, helping them to predict and prepare for hotel cancellations that they could lose revenue from. But it is again critical to note, as we've done in the limitations section, that this model is very specific and should not be generalized further than those two hotels. In the future though, we hope to be able to devise a model that may extend beyond those two hotels, perhaps into all of Portugal, but we will need a more expansive data set for that.

For additional details and computations, please see the following section.

Additional Work

Additional EDA

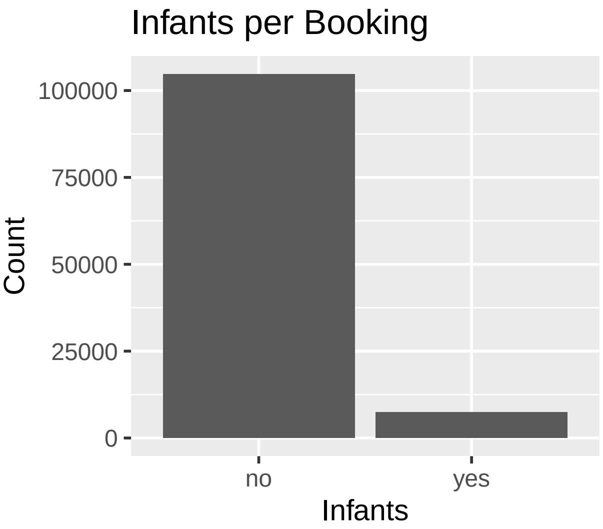
`adults`



	min	q1	median	q3	max	IQR	loweroutlier	upperoutlier
1	1	2	2	2	10	0	2	2

From graph of `adults`, the number of adults per booking, we can see that the distribution is skewed right. The center (median) is at 2 adults. The spread is 0, which means that the vast majority of bookings only are for 2 adults. However, it appears that there are outliers towards the higher end of the spectrum (above 2 adults). The maximum for the distribution is 55 adults, which is very large and may represent a group booking. The minimum number of adults for a booking is 0. This seems unusual, as it makes no sense to have a booking that has no adults (an empty room). We are only interested in being able to predict cancellations for bookings with a reasonable number of adults on the booking, so we will only consider observations with `0 < adults <= 10`.

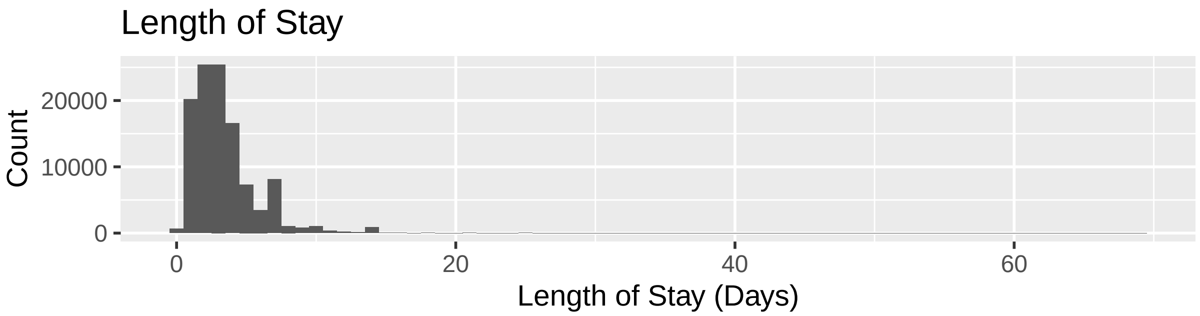
`infants`



	infants	n	prop
1	no	104642	0.93400336
2	yes	7394	0.06599664

From graph of `infants`, an indicator of whether a booking includes infants (children or babies) or not, we can see that the vast majority of bookings (92.18%) have 0 infants (are adults only).

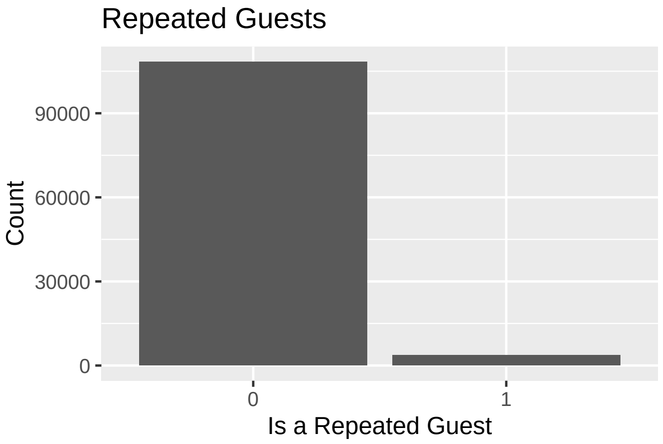
`length_stay`



	min	q1	median	q3	max	IQR	loweroutlier	upperoutlier
1	0	2	3	4	69	2	−1	7

From graph of `length_stay`, the number of nights stayed for a given booking, we can see that the distribution is skewed right. The center (median) is at 3 nights, meaning that most bookings include 3 nights as part of their stay. The spread is 2 days, which means that there is not a lot of variability in the number of nights stayed per reservation and that the length of a stay is generally on the short-term side. However, it appears that there are outliers towards the higher end of the spectrum (above 7 nights). The maximum for the distribution is 69 nights, a little over 2 months, which probably belongs to a long-term hotel stay. We will consider all lengths of stay and not filter out any outliers.

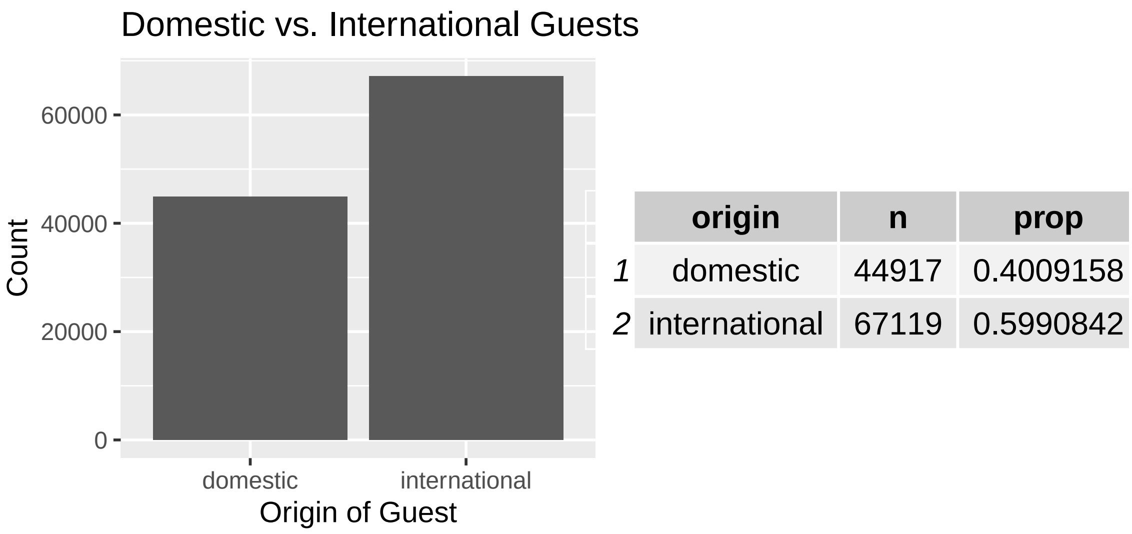
`is_repeated_guest`



	is_repeated_guest	n	prop
1	0	108333	0.96694812
2	1	3703	0.03305188

From the bar chart of `is_repeated_guest`, which is an indicator of whether a client is a repeated guest at the hotel, we can see that the vast majority of bookings in the data set (96.81%) were made by first-time guests, while the remaining bookings (3.19%) were made by repeated guests.

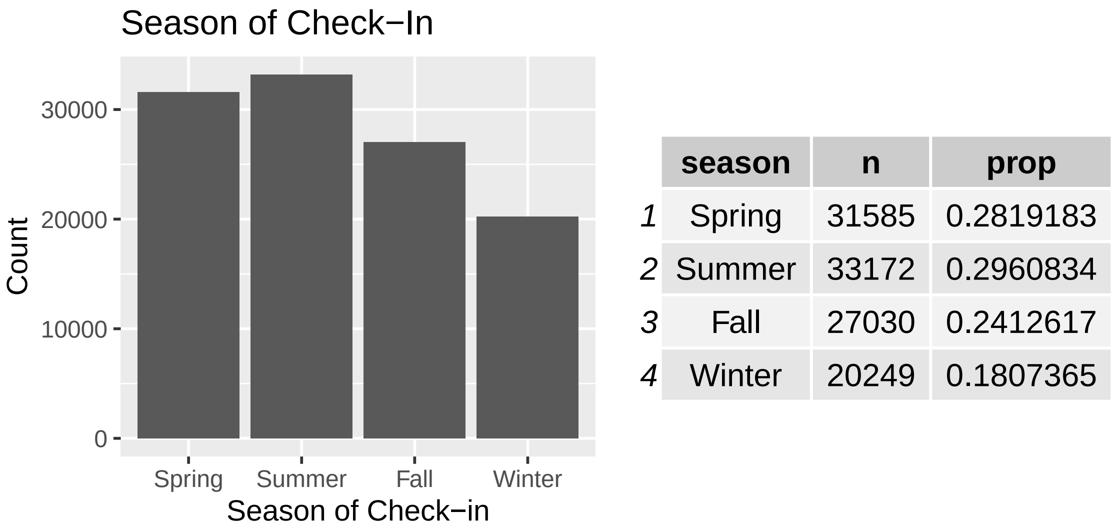
`origin`



	origin	n	prop
1	domestic	44917	0.4009158
2	international	67119	0.5990842

From the bar chart of `origin`, which provides information on whether a client is a domestic traveler (within Portugal) or a international traveler, we can see that the majority of clients in the data set (59.30%) are international travelers, while the remaining clients (40.70%) are domestic travelers (from Portugal).

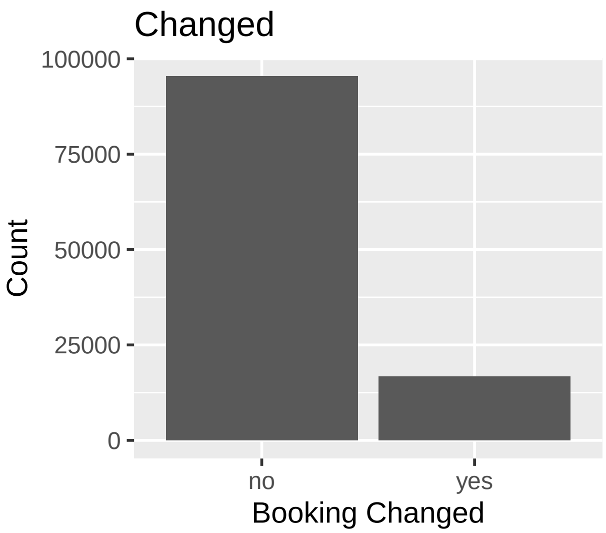
`season`

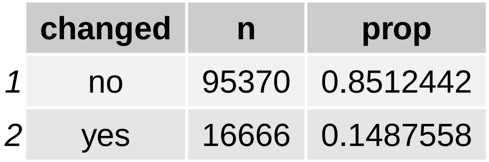


	season	n	prop
1	Spring	31585	0.2819183
2	Summer	33172	0.2960834
3	Fall	27030	0.2412617
4	Winter	20249	0.1807365

From the bar chart of `origin`, which provides information on what season the check-in for a booking occurred, we can see that the majority of bookings in the data set (31.39%) were made for the summer, followed by spring, then fall, and then finally winter (17.40%). This makes sense, as the warmer months tend to have more observations and people normally prefer to travel in nice weather.

`changed`

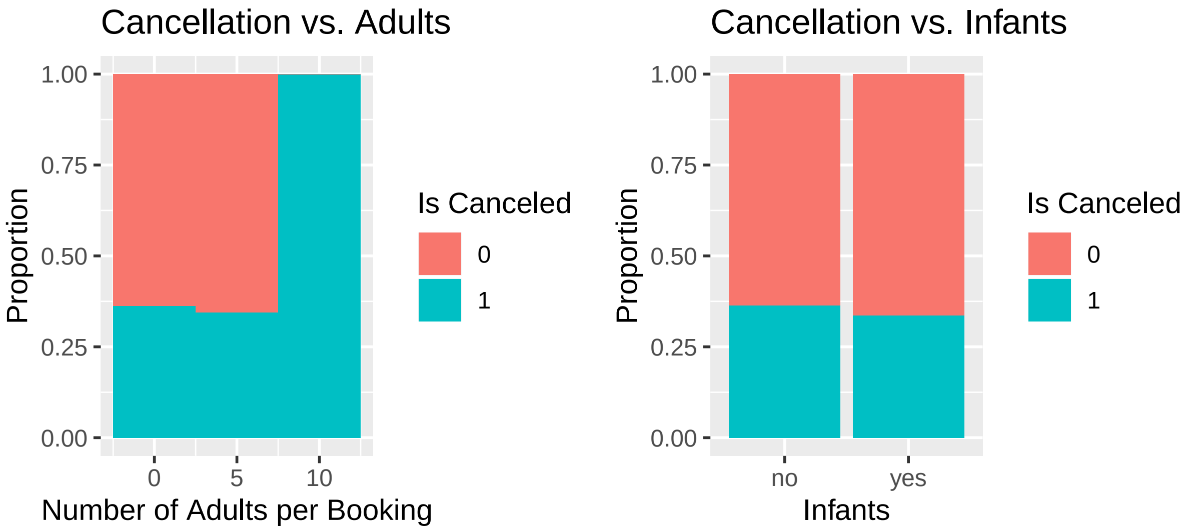




	changed	n	prop
1	no	95370	0.8512442
2	yes	16666	0.1487558

From graph of `changed`, an indicator of whether a booking has been modified or not, we can see that the vast majority of bookings (84.86%) have never been changed from their original booking.

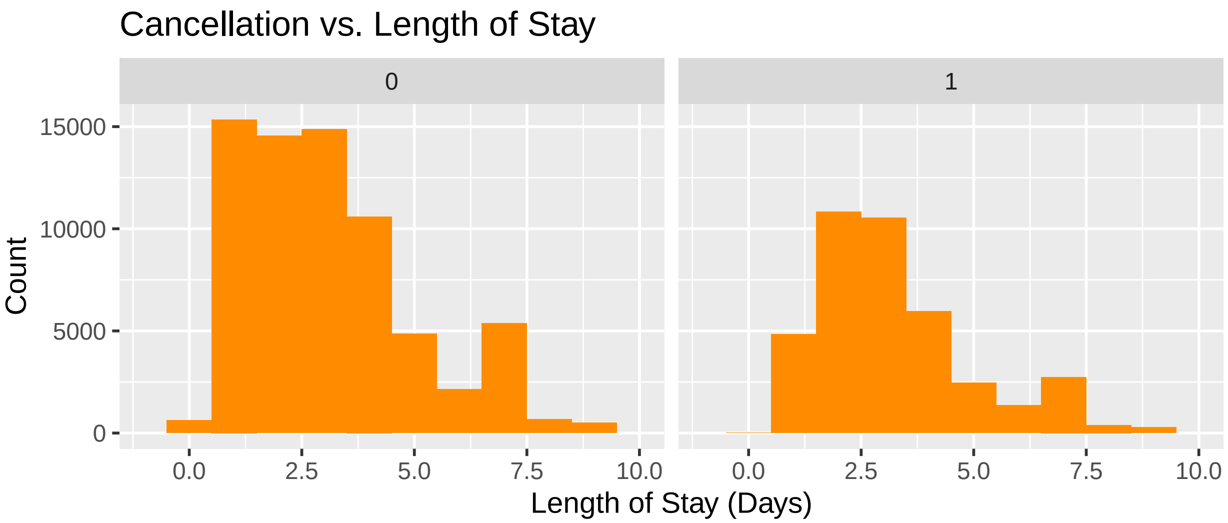
`adults` and `infants` vs. `is_cancelled`



The bar chart for `is_canceled` by `adults` shows that for bookings that have between 0 and 3 adults, the larger proportion of bookings are not cancelled. On the other hand, as the number of adults increase, especially for number of adults between 20 and 60, all of these bookings are canceled.

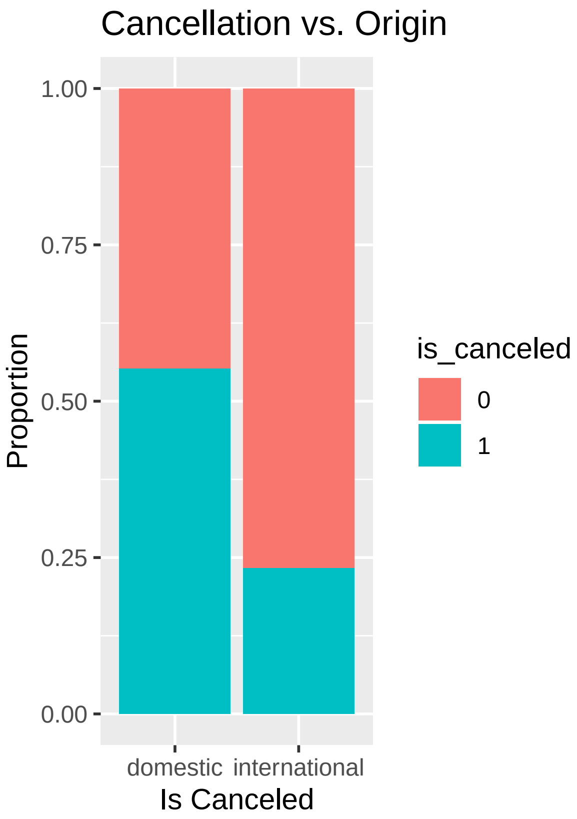
The bar chart for `is_ canceled` by `infants` shows that the number of infants does not look like it has a big effect on cancellations. We can see that proportion of cancellations are about the same for bookings that have infants and bookings that do not have infants.

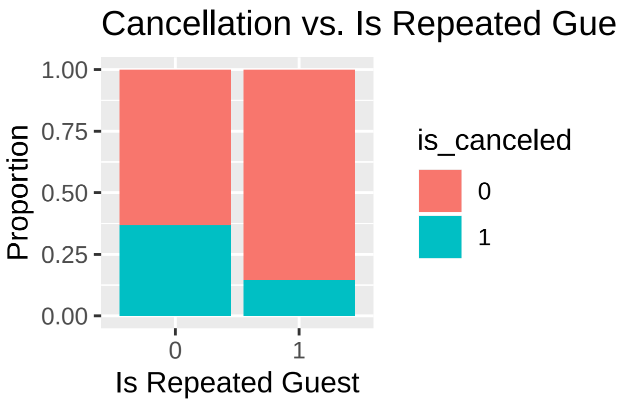
`length_stay` vs. `is_canceled`

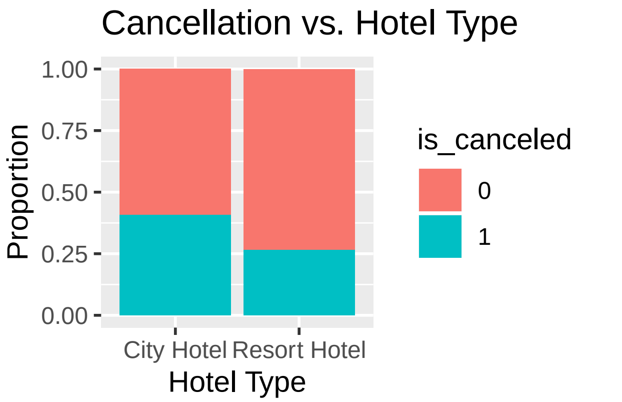


The histogram of `length_stay` by `is_canceled` shows that there is no significant difference in the distributions of lead time for canceled reservations vs. non-canceled reservations. For both canceled reservations and non-canceled reservations, the length of stay with the highest frequency is between 2 and 3 nights. The average length of stay is usually between 1 and 4 nights.

`origin`, `repeated_guest`, and `hotel` vs. `is_canceled`





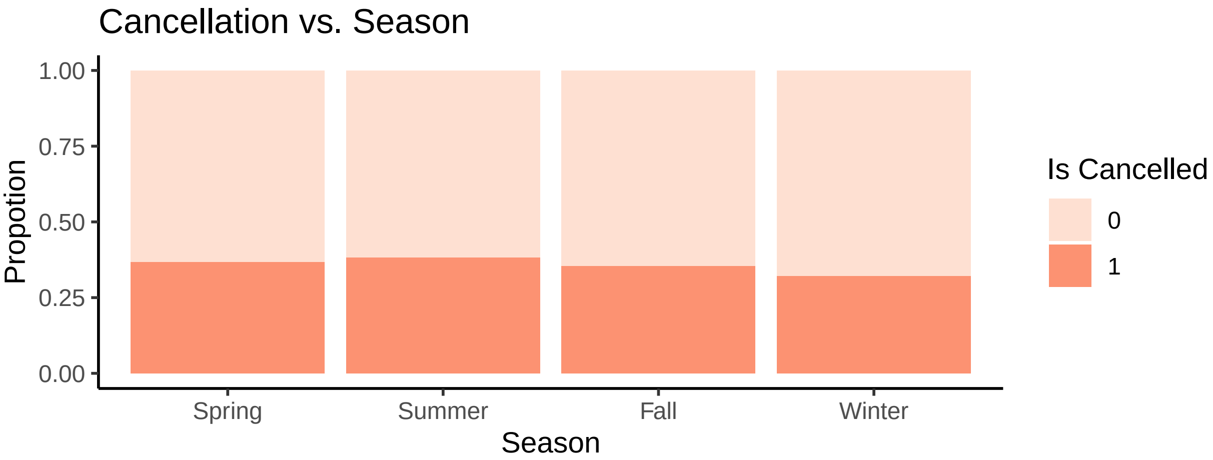


The bar plot `is_canceled` by `origin` shows that for reservations where the guest is domestic, the proportion of cancellations is greater than the proportions for non cancellations. For reservations where the guest is international, the proportion of non cancellations if far bigger than the proportion of cancellations.

The bar plot `is_canceled` by `is_repeated_guest` shows that for both repeat and non-repeat guests, the proportion of non-cancellations is far higher than the proportion of cancellations, although repeat guests have a lower cancellation rate than non-repeat guests.

The bar plot `is_canceled` by `hotel`shows that for both city and resort hotels, the proportion of non-cancellations is far higher than the proportion of cancellations, although resort hotels have a lower cancellation rate than city resorts.

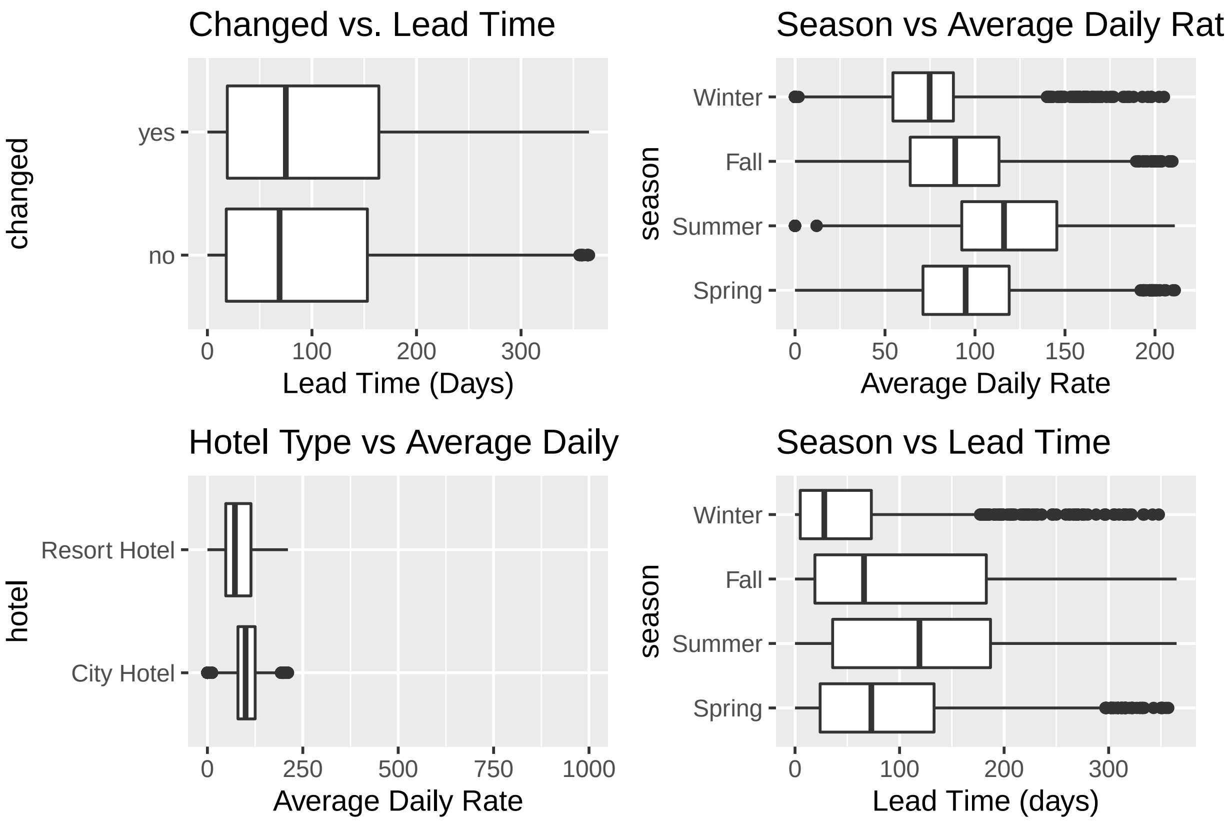
`season vs is_canceled`



The bar graphs for `season` by `is_canceled` shows that the distribution of `seasons` is roughly the same across canceled and non canceled reservations. Every season, the proportion of canceled reservations vs. non-canceled reservations stays roughly the same across the seasons. Summer has a slightly higher rate of cancellations than other seasons and winter has a slightly lower rate of cancellation than other seasons.

Interaction Analysis

Next, we'll plot the interaction variables to determine if there is any possible relationship or collinearity.



There does not seem to exist a strong association between `lead_time` and `changed`, as the distribution looks exactly the same for changed and unchanged reservations.

There may exist an interaction between the numerical variables `season` and `adr` because it seems that on average, the average daily rate for hotel bookings in the summer is larger than the other average daily rates. Furthermore, it seems that average daily rate for winter bookings are lower than other average daily rates.

There does not seem to exist as strong as an association between `hotel_type` and `adr`.

There also may exist an interaction between `season` and `lead_time`. According to the box plot (bottom right), people in the data set tended to make their summer bookings more ahead of time than winter bookings.

Model Fitting Process

Below we fit the full model with all four of the interactions we were interested in exploring earlier.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.154	0.681	-0.226	0.821	-1.686	1.065
adrCent	0.012	0.001	7.782	0.000	0.009	0.015
adultsCent	0.001	0.059	0.021	0.983	-0.113	0.116
seasonSummer	-0.226	0.091	-2.480	0.013	-0.405	-0.048
seasonFall	-0.258	0.085	-3.022	0.003	-0.425	-0.091
seasonWinter	-0.074	0.097	-0.765	0.444	-0.266	0.116
infantsyes	-0.143	0.103	-1.379	0.168	-0.347	0.059
changedyes	-1.057	0.106	-9.931	0.000	-1.269	-0.851
origininternational	-1.903	0.061	-31.086	0.000	-2.024	-1.784
hotelResort Hotel	-0.624	0.064	-9.758	0.000	-0.749	-0.499
is_repeated_guest1	-2.556	0.312	-8.201	0.000	-3.200	-1.974
lead_time_squared	0.000	0.000	8.319	0.000	0.000	0.000

term	estimate	std.error	statistic	p.value	conf.low	conf.high
market_segmentComplementary	-0.488	0.816	-0.598	0.550	-2.046	1.237
market_segmentCorporate	-0.317	0.693	-0.457	0.648	-1.565	1.234
market_segmentDirect	-0.245	0.686	-0.357	0.721	-1.475	1.294
market_segmentGroups	1.081	0.682	1.584	0.113	-0.142	2.615
market_segmentOffline TA/TO	0.231	0.681	0.339	0.735	-0.989	1.763
market_segmentOnline TA	1.122	0.680	1.650	0.099	-0.095	2.653
prior_cancellationyes	3.540	0.272	13.002	0.000	3.036	4.109
length_stayCent	0.090	0.011	8.503	0.000	0.069	0.111
adrCent:seasonSummer	-0.002	0.002	-0.935	0.350	-0.005	0.002
adrCent:seasonFall	0.000	0.002	-0.208	0.835	-0.004	0.003
adrCent:seasonWinter	-0.004	0.002	-1.791	0.073	-0.009	0.000
seasonSummer:lead_time_squared	0.000	0.000	-1.511	0.131	0.000	0.000
seasonFall:lead_time_squared	0.000	0.000	-0.772	0.440	0.000	0.000
seasonWinter:lead_time_squared	0.000	0.000	2.456	0.014	0.000	0.000
adrCent:hotelResort Hotel	-0.002	0.001	-1.589	0.112	-0.005	0.000
changedyes:lead_time_squared	0.000	0.000	-3.218	0.001	0.000	0.000

Using the `step` function, we will perform a backward selection on multiple linear regression models with BIC as the selection criteria. We do this by setting k = log(n), in which k is the degrees of freedom. We chose BIC as the selection criterion because we have a lot of variables and want to strictly penalize for any variables that are not truly necessary. Since our model is intended to be used by hotel personnel, we want to keep it as simple and efficient as possible.

```
## Start:  AIC=9980.19
## is_canceled ~ adrCent + adultsCent + season + infants + changed +
##     origin + hotel + is_repeated_guest + lead_time_squared +
##     market_segment + prior_cancellation + length_stayCent + (season *
##     adrCent) + (season * lead_time_squared) + (hotel * adrCent) +
##     (changed * lead_time_squared)
##
##                             Df Deviance     AIC
## - adrCent:season             3   9726.0  9956.3
## - season:lead_time_squared   3   9735.8  9966.0
## - adultsCent                 1   9722.3  9971.0
## - infants                    1   9724.2  9972.9
## - adrCent:hotel              1   9724.8  9973.5
## <none>                           9722.3  9980.2
## - changed:lead_time_squared  1   9733.3  9981.9
## - length_stayCent            1   9797.8 10046.5
## - is_repeated_guest          1   9826.6 10075.3
## - market_segment             6  10098.0 10300.6
## - prior_cancellation         1  10089.2 10337.9
## - origin                     1  10822.6 11071.3
##
## Step:  AIC=9956.27
## is_canceled ~ adrCent + adultsCent + season + infants + changed +
##     origin + hotel + is_repeated_guest + lead_time_squared +
##     market_segment + prior_cancellation + length_stayCent + season:lead_time_squared +
##     adrCent:hotel + changed:lead_time_squared
##
##                             Df Deviance     AIC
## - season:lead_time_squared   3   9738.7  9941.4
## - adultsCent                 1   9726.0  9947.1
## - infants                    1   9728.3  9949.4
## - adrCent:hotel              1   9729.6  9950.6
## <none>                           9726.0  9956.3
## - changed:lead_time_squared  1   9737.2  9958.2
## - length_stayCent            1   9800.7 10021.7
## - is_repeated_guest          1   9830.0 10051.0
## - market_segment             6  10101.6 10276.6
## - prior_cancellation         1  10093.2 10314.2
## - origin                     1  10830.7 11051.8
##
## Step:  AIC=9941.36
## is_canceled ~ adrCent + adultsCent + season + infants + changed +
##     origin + hotel + is_repeated_guest + lead_time_squared +
##     market_segment + prior_cancellation + length_stayCent + adrCent:hotel +
##     changed:lead_time_squared
##
##                             Df Deviance     AIC
## - adultsCent                 1   9738.7  9932.2
## - infants                    1   9741.2  9934.6
## - adrCent:hotel              1   9741.9  9935.3
## <none>                           9738.7  9941.4
## - changed:lead_time_squared  1   9751.2  9944.6
## - season                     3   9777.3  9952.3
## - length_stayCent            1   9814.6 10008.1
## - is_repeated_guest          1   9843.6 10037.0
## - market_segment             6  10117.3 10264.6
## - prior_cancellation         1  10103.3 10296.7
## - origin                     1  10842.9 11036.3
##
## Step:  AIC=9932.15
## is_canceled ~ adrCent + season + infants + changed + origin +
##     hotel + is_repeated_guest + lead_time_squared + market_segment +
##     prior_cancellation + length_stayCent + adrCent:hotel + changed:lead_time_squared
##
##                             Df Deviance     AIC
## - infants                    1   9741.2  9925.4
## - adrCent:hotel              1   9741.9  9926.1
## <none>                           9738.7  9932.2
## - changed:lead_time_squared  1   9751.2  9935.5
## - season                     3   9777.3  9943.1
## - length_stayCent            1   9814.7  9998.9
## - is_repeated_guest          1   9843.7 10027.9
## - market_segment             6  10122.7 10260.9
## - prior_cancellation         1  10103.4 10287.6
## - origin                     1  10847.1 11031.3
##
## Step:  AIC=9925.4
## is_canceled ~ adrCent + season + changed + origin + hotel + is_repeated_guest +
##     lead_time_squared + market_segment + prior_cancellation +
##     length_stayCent + adrCent:hotel + changed:lead_time_squared
##
##                             Df Deviance     AIC
## - adrCent:hotel              1   9744.2  9919.2
## <none>                         9741.2  9925.4
## - changed:lead_time_squared  1   9753.6  9928.6
## - season                     3   9778.8  9935.4
## - length_stayCent            1   9817.5  9992.5
## - is_repeated_guest          1   9845.9 10020.9
## - market_segment             6  10125.9 10254.8
## - prior_cancellation         1  10105.9 10280.9
## - origin                     1  10847.7 11022.7
##
## Step:  AIC=9919.23
## is_canceled ~ adrCent + season + changed + origin + hotel + is_repeated_guest +
##     lead_time_squared + market_segment + prior_cancellation +
##     length_stayCent + changed:lead_time_squared
##
##                             Df Deviance     AIC
## <none>                         9744.2  9919.2
## - changed:lead_time_squared  1   9756.7  9922.5
## - season                     3   9784.2  9931.5
## - length_stayCent            1   9818.9  9984.7
## - is_repeated_guest          1   9849.1 10014.8
## - hotel                      1   9862.9 10028.7
## - adrCent                    1   9895.4 10061.2
## - market_segment             6  10132.7 10252.5
## - prior_cancellation         1  10107.2 10272.9
## - origin                     1  10848.9 11014.7
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.115	0.678	-0.170	0.865	-1.643	1.097
adrCent	0.009	0.001	12.205	0.000	0.008	0.011
seasonSummer	-0.340	0.068	-5.032	0.000	-0.473	-0.208
seasonFall	-0.298	0.071	-4.210	0.000	-0.436	-0.159
seasonWinter	0.042	0.078	0.537	0.591	-0.111	0.195
changedyes	-1.065	0.106	-10.052	0.000	-1.276	-0.860
origininternational	-1.892	0.061	-31.152	0.000	-2.011	-1.773
hotelResort Hotel	-0.640	0.060	-10.711	0.000	-0.758	-0.523
is_repeated_guest1	-2.557	0.311	-8.209	0.000	-3.201	-1.976
lead_time_squared	0.000	0.000	14.189	0.000	0.000	0.000
market_segmentComplementary	-0.493	0.813	-0.607	0.544	-2.044	1.227
market_segmentCorporate	-0.317	0.693	-0.458	0.647	-1.564	1.233
market_segmentDirect	-0.260	0.684	-0.380	0.704	-1.486	1.277
market_segmentGroups	1.081	0.681	1.589	0.112	-0.138	2.613
market_segmentOffline TA/TO	0.222	0.680	0.327	0.744	-0.994	1.752
market_segmentOnline TA	1.119	0.678	1.650	0.099	-0.094	2.647
prior_cancellationyes	3.499	0.271	12.898	0.000	2.998	4.067
length_stayCent	0.089	0.010	8.470	0.000	0.068	0.109
changedyes:lead_time_squared	0.000	0.000	-3.427	0.001	0.000	0.000

The backwards selection using BIC removed the following variables: `adultsCent`, `infants`, and all of the interaction variables.

We can see that BIC did not remove the variable `lead_time_squared`, but the estimated coefficient is approximately 0. Therefore, although this variable is statistically significant due to the large sample size, it is not practically significant and thus we will remove it from the model.

Interpretation of Intercept

We expect the odds of a domestic City hotel booking made the day of (lead_time = 0) in the Spring with an average daily rate of 97.80 dollars made for 3.42 days and for 2 adults by someone who is not a repeated guest, has made no booking changes, and has made no previous cancellations to be 1.1888661.

Interpretation of Coefficients

We will interpret the coefficient of one categorical variable (`origin`), one numerical variable (`adrCent`), and one interaction variable (`is_repeated_guest1:previous_canellations`).

`origin`: The odds of a booking with an international origin being canceled is expected to be 0.1412816 times the odds of a booking with a domestic origin, holding all else constant. In other terms, domestic bookings have larger odds of being canceled, on average and holding all else constant.

`adrCent`: Holding all else constant, for every dollar increase in the average daily rate, odds of a booking being canceled multiplies by a factor of 1.0090406.

`adrCent:hotelResort Hotel`: Holding all else constant, for every dollar increase in the average daily rate, we expect the odds of a booking being canceled to multiply by a factor 0.9970045 if a hotel is a Resort Hotel versus if it is a City Hotel.

AIC Model Check

We can also perform a backwards selection using AIC as our criterion instead to determine if there's any difference from our current model selection.

```
## Start:  AIC=9778.3
## is_canceled ~ adrCent + adultsCent + season + infants + changed +
##     origin + hotel + is_repeated_guest + lead_time_squared +
##     market_segment + prior_cancellation + length_stayCent + (season *
##     adrCent) + (season * lead_time_squared) + (hotel * adrCent) +
##     (changed * lead_time_squared)
##
##                             Df Deviance     AIC
## - adrCent:season             3   9726.0  9776.0
## - adultsCent                 1   9722.3  9776.3
## - infants                    1   9724.2  9778.2
## <none>                           9722.3  9778.3
## - adrCent:hotel              1   9724.8  9778.8
## - season:lead_time_squared   3   9735.8  9785.8
## - changed:lead_time_squared  1   9733.3  9787.3
## - length_stayCent            1   9797.8  9851.8
## - is_repeated_guest          1   9826.6  9880.6
## - market_segment             6  10098.0 10142.0
## - prior_cancellation         1  10089.2 10143.2
## - origin                     1  10822.6 10876.6
##
## Step:  AIC=9776
## is_canceled ~ adrCent + adultsCent + season + infants + changed +
##     origin + hotel + is_repeated_guest + lead_time_squared +
##     market_segment + prior_cancellation + length_stayCent + season:lead_time_squared +
##     adrCent:hotel + changed:lead_time_squared
##
##                             Df Deviance     AIC
## - adultsCent                 1   9726.0  9774.0
## <none>                           9726.0  9776.0
## - infants                   1   9728.3  9776.3
## - adrCent:hotel             1   9729.6  9777.6
## - season:lead_time_squared   3   9738.7  9782.7
## - changed:lead_time_squared  1   9737.2  9785.2
## - length_stayCent           1   9800.7  9848.7
## - is_repeated_guest         1   9830.0  9878.0
## - market_segment            6  10101.6 10139.6
## - prior_cancellation        1  10093.2 10141.2
## - origin                    1  10830.7 10878.7
##
## Step:  AIC=9774.02
## is_canceled ~ adrCent + season + infants + changed + origin +
##     hotel + is_repeated_guest + lead_time_squared + market_segment +
##     prior_cancellation + length_stayCent + season:lead_time_squared +
##     adrCent:hotel + changed:lead_time_squared
##
##                            Df Deviance     AIC
## <none>                          9726.0  9774.0
## - infants                   1   9728.3  9774.3
## - adrCent:hotel             1   9729.6  9775.6
## - season:lead_time_squared   3   9738.7  9780.7
## - changed:lead_time_squared  1   9737.2  9783.2
## - length_stayCent           1   9800.7  9846.7
## - is_repeated_guest         1   9830.1  9876.1
## - prior_cancellation        1  10093.2 10139.2
## - market_segment            6  10106.7 10142.7
## - origin                    1  10835.2 10881.2
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.143	0.678	-0.211	0.833
adrCent	0.011	0.001	10.654	0.000
seasonSummer	-0.250	0.084	-2.988	0.003
seasonFall	-0.265	0.084	-3.137	0.002
seasonWinter	-0.023	0.090	-0.257	0.797
infantsyes	-0.156	0.103	-1.517	0.129
changedyes	-1.057	0.106	-9.935	0.000
origininternational	-1.906	0.061	-31.186	0.000
hotelResort Hotel	-0.625	0.060	-10.386	0.000
is_repeated_guest1	-2.547	0.311	-8.192	0.000
lead_time_squared	0.000	0.000	8.314	0.000
market_segmentComplementary	-0.413	0.813	-0.508	0.611
market_segmentCorporate	-0.316	0.693	-0.456	0.649
market_segmentDirect	-0.247	0.684	-0.360	0.719
market_segmentGroups	1.077	0.681	1.581	0.114
market_segmentOffline TA/TO	0.227	0.680	0.334	0.738
market_segmentOnline TA	1.121	0.679	1.652	0.099
prior_cancellationyes	3.526	0.271	12.991	0.000
length_stayCent	0.089	0.011	8.462	0.000
seasonSummer:lead_time_squared	0.000	0.000	-1.355	0.176
seasonFall:lead_time_squared	0.000	0.000	-0.706	0.480
seasonWinter:lead_time_squared	0.000	0.000	2.476	0.013
adrCent:hotelResort Hotel	-0.002	0.001	-1.886	0.059
changedyes:lead_time_squared	0.000	0.000	-3.244	0.001

```
## [1] 9991.866
```

```
## [1] 9774.025
```

```
## [1] 10114.44
```

```
## [1] 9947.073
```

We can see that backwards selection using AIC as the criterion produces the same model as the model produced when BIC is used as the criterion.

K-Fold Validation

The k-fold cross validation method involves splitting the data set into k-subsets. In this instance, we will use 10 subsets, as our data set includes close to 100,000 observations. One subset is used as the testing set, while the model is trained on all other subsets. This process is completed until accuracy is determined for each of the 10 rounds of validation, and an overall accuracy estimate is provided. Therefore, k-fold validation is a robust method for estimating model accuracy. As the goal of our analysis is to make accurate predictions about whether a booking will or will not be cancelled, we certainly care about accuracy.

Below, we use 10-fold cross validation to estimate the generalized linear model on the `hotels_small` data set. For this analysis, we will use functions from the `caret` package.

```
## Generalized Linear Model
##
## 112036 samples
##     12 predictor
##      2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 100832, 100832, 100832, 100832, 100833, 100833, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7613893  0.451714
```

```
##
## Call:  NULL
##
## Coefficients:
##                     (Intercept)                               adr
##                      -1.105e+00                         9.303e-03
##                      infantsyes                            adults
##                      -9.315e-05                         6.279e-02
##                    seasonSummer                        seasonFall
##                      -3.880e-01                        -1.154e-01
##                    seasonWinter                        changedyes
##                       6.518e-02                        -1.021e+00
##             origininternational              `hotelResort Hotel`
##                      -1.938e+00                        -4.255e-01
##              is_repeated_guest1                 lead_time_squared
##                      -2.904e+00                         1.709e-05
##      market_segmentComplementary           market_segmentCorporate
##                      -5.285e-01                        -4.783e-01
##            market_segmentDirect              market_segmentGroups
##                      -6.174e-01                         7.424e-01
##    `market_segmentOffline TA/TO`         `market_segmentOnline TA`
##                       -9.546e-02                         8.322e-01
##            prior_cancellationyes                       length_stay
##                        3.634e+00                         9.206e-02
##   `changedyes:lead_time_squared`                `adr:seasonSummer`
##                       -7.530e-06                         4.921e-04
##                 `adr:seasonFall`                `adr:seasonWinter`
##                       -2.283e-03                        -2.459e-03
##          `adr:hotelResort Hotel`  `seasonSummer:lead_time_squared`
##                       -2.861e-03                        -1.506e-06
##   `seasonFall:lead_time_squared`  `seasonWinter:lead_time_squared`
##                       -8.613e-07                         1.622e-05
## 
## Degrees of Freedom: 112035 Total (i.e. Null);  112008 Residual
## Null Deviance:       146600 
## Residual Deviance: 108700    AIC: 108700
```

```
## 
## Call:
## NULL
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.0369  -0.7832  -0.4450   0.8135   3.6682  
## 
## Coefficients:
##                                   Estimate Std. Error  z value Pr(>|z|)    
## (Intercept)                     -1.105e+00  1.859e-01   -5.947 2.74e-09 ***
## adr                              9.303e-03  4.338e-04   21.445  < 2e-16 ***
## infantsyes                      -9.315e-05  3.058e-02   -0.003 0.997569    
## adults                           6.279e-02  1.734e-02    3.622 0.000293 ***
## seasonSummer                    -3.880e-01  6.771e-02   -5.731 1.00e-08 ***
## seasonFall                      -1.154e-01  6.316e-02   -1.828 0.067615 .  
## seasonWinter                     6.518e-02  6.582e-02    0.990 0.322027    
## changedyes                      -1.021e+00  3.079e-02  -33.162  < 2e-16 ***
## origininternational             -1.938e+00  1.830e-02 -105.899  < 2e-16 ***
## `hotelResort Hotel`             -4.255e-01  4.286e-02   -9.928  < 2e-16 ***
## is_repeated_guest1              -2.904e+00  9.567e-02  -30.352  < 2e-16 ***
## lead_time_squared                1.709e-05  6.456e-07   26.465  < 2e-16 ***
## market_segmentComplementary     -5.285e-01  2.180e-01   -2.424 0.015337 *  
## market_segmentCorporate         -4.783e-01  1.843e-01   -2.595 0.009464 ** 
## market_segmentDirect            -6.174e-01  1.824e-01   -3.385 0.000712 ***
## market_segmentGroups             7.424e-01  1.812e-01    4.096 4.20e-05 ***
## `market_segmentOffline TA/TO`   -9.546e-02  1.809e-01   -0.528 0.597616    
## `market_segmentOnline TA`        8.322e-01  1.805e-01    4.610 4.03e-06 ***
## prior_cancellationyes            3.634e+00  8.579e-02   42.360  < 2e-16 ***
## length_stay                      9.206e-02  3.242e-03   28.398  < 2e-16 ***
## `changedyes:lead_time_squared`  -7.530e-06  8.842e-07   -8.517  < 2e-16 ***
## `adr:seasonSummer`               4.921e-04  5.436e-04    0.905 0.365398    
## `adr:seasonFall`                -2.283e-03  5.692e-04   -4.011 6.05e-05 ***
## `adr:seasonWinter`              -2.459e-03  7.013e-04   -3.507 0.000454 ***
## `adr:hotelResort Hotel`         -2.861e-03  3.993e-04   -7.165 7.80e-13 ***
## `seasonSummer:lead_time_squared` -1.506e-06  8.104e-07   -1.858 0.063135 .  
## `seasonFall:lead_time_squared`   -8.613e-07  8.333e-07   -1.034 0.301300    
## `seasonWinter:lead_time_squared`  1.622e-05  1.417e-06   11.444  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 146610  on 112035  degrees of freedom
## Residual deviance: 108667  on 112008  degrees of freedom
## AIC: 108723
##
## Number of Fisher Scoring iterations: 6
```

Accuracy is the proportion of accurate predictions, which in this case is 0.7611571. This means that the model produced by k-fold cross validation correctly predicts for about 76.127% of bookings, which is pretty good! We can see that cross validation removed `infants`, which is the same result from our BIC test above.