BUSN 41201: Big Data
Spring Quarter 2023 - Veronika Rockova

Hotel Reservation Cancellations:

An Analytical Approach to Understanding and Modeling Customer Behavior

May 28th, 2023

Tina Tong, Jingxuan Zhang, Michael Han, Emil Perdue, Alexander Vattakkattu

Honor Code

We pledge our honor that we have not violated the Booth Honor Code during this assignment.

Table of Contents

1. Executive Summary

This research paper presents a comprehensive analysis of a 'Hotel Reservation Dataset'[1] to gain insights into customer cancellation behavior and its underlying factors. By examining various aspects related to hotel reservations, including market segments, booking preferences, and previous cancellation patterns, we aimed to identify the types of customers more likely to cancel their reservations and explore predictive models to assist hotels in flagging such users.

We first focus on the factors that affect the hotel's pricing. Our second question explores the differences in booking preferences and behaviors among distinct market segments, such as business travelers, families, and couples. Next, we focused on developing predictive models to accurately forecast reservation cancellations and identify potential cancellations in advance. Finally, we looked at which variables in our predictive model might be causally related to hotel cancellation.

[1] Link to the dataset: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

2. Introduction

In this research paper, our group looks to emphasize three unique goals:

1. *Understanding Hotel Strategy*
2. *Knowledge of Customer Preference*
3. *Optimizing Benefits for Hotels*

The Kaggle dataset we are working with presents the following action question: "The online hotel reservation channels have dramatically changed booking possibilities and customers' behavior. A significant number of hotel reservations are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with." With this in mind, our group will look to understand the rationale behind cancellation and if certain customers are more likely to preemptively cancel than others.

In our paper, we will begin by exploring the associated Kaggle dataset before diving into a number of research questions. We will start with data cleaning/data exploration before diving into the following four fundamental questions, which will allow us to draw significant conclusions about customer behavior and inference regarding future hotel reservations.

Our research questions are as follows:

1. *What are the factors that affect the average price per room in the hotel?*
2. *How do booking preferences differ among distinct market segments?*
3. *What reservation characteristics lead to hotel cancellation?*
4. *Can we identify any causal factors for predicting hotel cancellations?*

3. Dataset

The columns of use in our dataset are as follows:

- Booking_ID: unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

4. Data Cleaning

Our dataset does not include any missing values.

For the purposes of modeling, we have re-coded several of our variables.

For the purposes of data cleaning to provide more concrete numerical responses, we relabel entries within columns as follows:

- We convert ‘Room Type’ numbers into their specific number (1, 2, 3, 4, 5, 6, or 7), where ‘Room Type 1’ would be replaced with ‘1’.
- We convert ‘Meal Plan’ numbers into their specific number (1, 2, or 3), where Meal Plan 1’ would be replaced with ‘1’. Additionally, we also replace ‘Not Selected’ with 0.
- We replace ‘'Market Segment Type’: ‘Offline’, ‘Online’, ‘Aviation’, ‘Complementary’, and ‘Corporate’ to 0, 1, 2, 3, or 4 respectively
- We replace ‘Booking Status’ ‘Canceled’ or ‘Not Canceled’ to 0 and 1, respectively

This data relabeling process allows us to better examine and categorize our variables for more in-depth analysis later in our paper.

5. Exploratory Analysis

First, we hope to consider each of the columns that we have access to in this dataset, and think about how they can help solve our overarching problem of trying to predict how likely a customer is to cancel their hotel booking. Picking apart which variables we might initially believe to have importance, we can draw the following initial conclusions:

1. Lead Time: The number of days between the date of booking and the arrival date can provide insights into the customers' commitment level. Longer lead times may indicate a lower probability of cancellation.

2. Type of Meal Plan: The meal plan chosen by the customer may impact their decision to cancel. Some meal plans may be more flexible or have lower cancellation penalties, leading to fewer cancellations. Similarly, high lever meal plans may entice customers to stay, or conversely, be too expensive, and lead a customer astray.

3. Required Car Parking Space: Customers who require a car parking space may have specific needs or preferences that could affect their cancellation behavior, particularly, them being more prone to potential travel conflicts.

4. Room Type Reserved: The type of room reserved by the customer might be a factor in their decision to cancel. Certain room types may be more desirable or have specific features that customers value, making them less likely to cancel.

5. Market Segment Type: The market segment designation of the customer can provide insights into their behavior and preferences. Different market segments may have varying cancellation patterns which we will explore.

6. Repeated Guest: Customers who have stayed at the hotel before (repeated guests) may have less of a tendency to cancel compared to first-time guests. Specifically, repeat guests might be more loyal.

7. Number of Previous Cancellations: If a customer has a history of canceling previous bookings, it might indicate a higher likelihood of cancellation for the current booking.

8. Number of Previous Bookings Not Canceled: Conversely, if a customer has a history of not canceling previous bookings, it might suggest a lower likelihood of cancellation for the current booking.

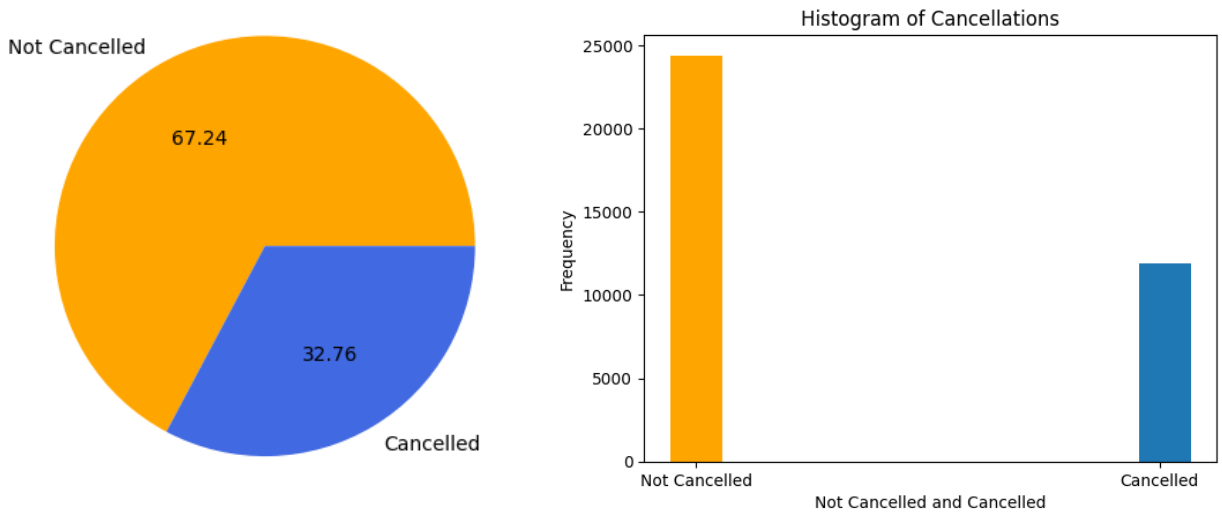
9. Average Price per Room: The average price per room might influence the customers' commitment and financial considerations. Higher-priced reservations might have a lower probability of cancellation as customers may be planning their trips out in advance with budgets already mapped out.

10. Number of Special Requests: The total number of special requests made by the customer can indicate their level of engagement and specific preferences, which might lead them to be less likely to cancel.

It is important to note that the importance and relevance of these variables may vary depending on the specific context and characteristics of the hotel and its customers.

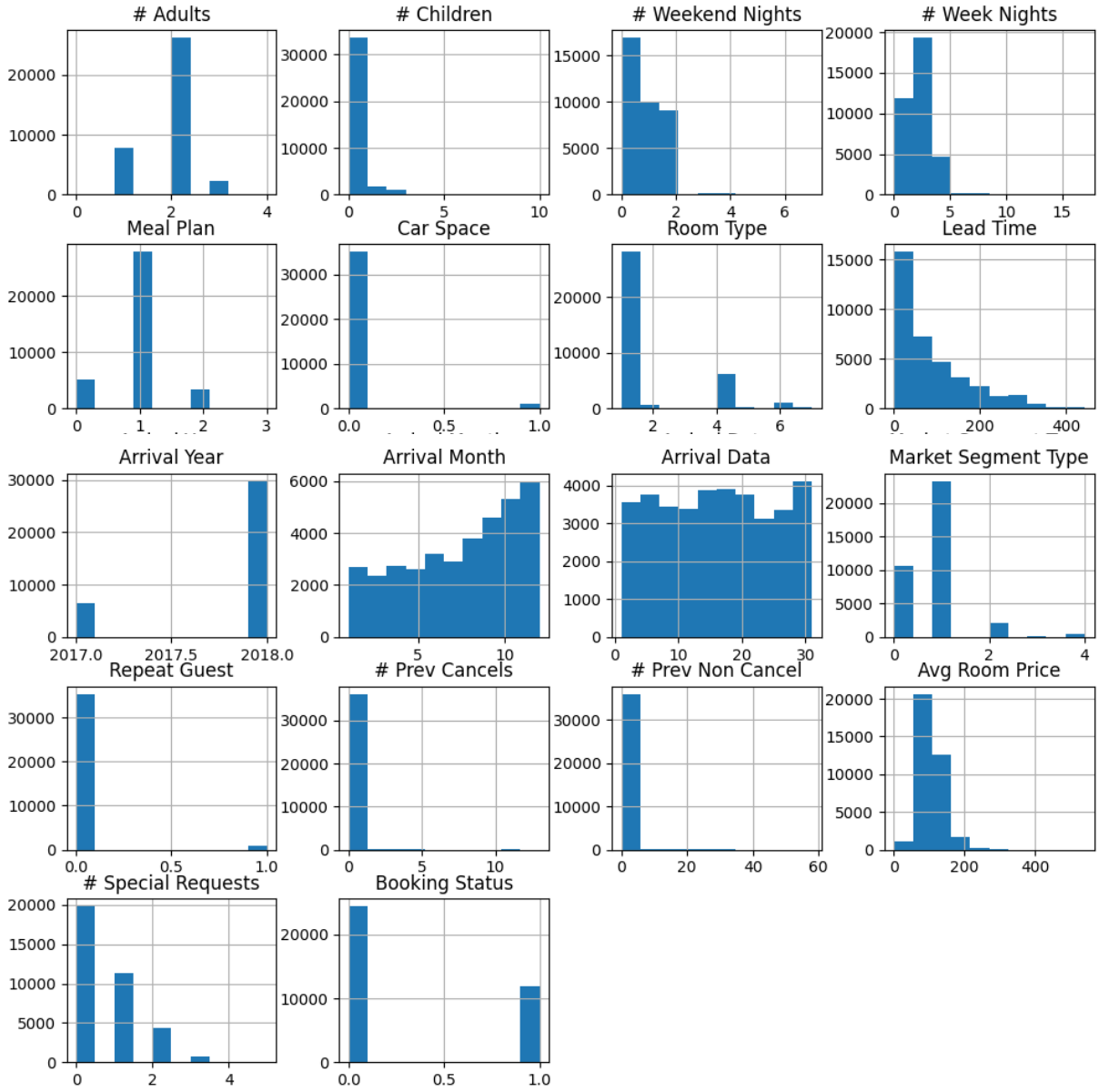
a) Cancellation Histogram

First, we look to understand our dependent variable, the number of cancellations (re-coded to 0 for 'Not Canceled' and 1 for 'Canceled').



For context, we have 36,275 rows of bookings data. We can see from our charts above that out of all bookings, 67.54% were not canceled as opposed to 32.76% being canceled, reflecting a minority in our population of overall bookings. We can see an approximately 2:1 ratio between 'Not Canceled' and 'Canceled'.

b) Overall Histogram Overview



Above we have included histograms for each of our variables when trying to develop an understanding of booking rates. Similarly, for references to independent variable coding, one can navigate to the 'Data Cleaning' section of our paper.

In general, we don't see any distributions that may skew away from what we believe to be normal in any of these variables, relative to a normal hotel going, worldly population. We do want to make note of potential correlations across these histograms as well as the significance of categories of data and the average room price, which we will explore below.

c) Correlation Matrix

	# Adults	# Children	# Weekend Nights	# Week Nights	Meal Plan	Car Space	Room Type	Lead Time	Arrival Year	Arrival Month	Arrival Data	Market Segment Type	Repeat Guest	# Prev Cancels	# Prev Non Cancel	Avg Room Price	# Special Requests	Booking Status
# Adults	1.000	-0.020	0.103	0.106	-0.004	0.011	0.270	0.097	0.077	0.022	0.026	-0.096	-0.192	-0.047	-0.119	0.297	0.189	0.087
# Children	-0.020	1.000	0.029	0.024	0.042	0.034	0.364	-0.047	0.046	-0.003	0.025	0.073	-0.036	-0.016	-0.021	0.338	0.124	0.033
# Weekend Nights	0.103	0.029	1.000	0.180	-0.019	-0.031	0.057	0.047	0.055	-0.010	0.027	-0.020	-0.067	-0.021	-0.026	-0.005	0.061	0.062
# Week Nights	0.106	0.024	0.180	1.000	0.027	-0.049	0.094	0.150	0.033	0.037	-0.009	-0.065	-0.100	-0.030	-0.049	0.023	0.046	0.093
Meal Plan	-0.004	0.042	-0.019	0.027	1.000	-0.015	0.093	0.227	-0.188	0.017	0.016	-0.223	0.010	-0.007	0.006	0.135	-0.090	0.049
Car Space	0.011	0.034	-0.031	-0.049	-0.015	1.000	0.039	-0.066	0.016	-0.016	-0.000	0.117	0.111	0.027	0.064	0.061	0.088	-0.086
Room Type	0.270	0.364	0.057	0.094	0.093	0.039	1.000	-0.108	0.103	-0.006	0.033	0.163	-0.026	-0.008	-0.008	0.470	0.145	0.023
Lead Time	0.097	-0.047	0.047	0.150	0.227	-0.066	-0.108	1.000	0.143	0.137	0.006	-0.312	-0.136	-0.046	-0.078	-0.063	-0.102	0.439
Arrival Year	0.077	0.046	0.055	0.033	-0.188	0.016	0.103	0.143	1.000	-0.340	0.019	0.082	-0.018	0.004	0.026	0.179	0.053	0.180
Arrival Month	0.022	-0.003	-0.010	0.037	0.017	-0.016	-0.006	0.137	-0.340	1.000	-0.043	-0.027	0.000	-0.039	-0.011	0.054	0.111	-0.011
Arrival Data	0.026	0.025	0.027	-0.009	0.016	-0.000	0.033	0.006	0.019	-0.043	1.000	0.008	-0.016	-0.013	-0.001	0.018	0.018	0.011
Market Segment Type	-0.096	0.073	-0.020	-0.065	-0.223	0.117	0.163	-0.312	0.082	-0.027	0.008	1.000	0.298	0.074	0.197	-0.043	0.203	-0.049
Repeat Guest	-0.192	-0.036	-0.067	-0.100	0.010	0.111	-0.026	-0.136	-0.018	0.000	-0.016	0.298	1.000	0.391	0.539	-0.175	-0.012	-0.107
# Prev Cancels	-0.047	-0.016	-0.021	-0.030	-0.007	0.027	-0.008	-0.046	0.004	-0.039	-0.013	0.074	0.391	1.000	0.468	-0.063	-0.003	-0.034
# Prev Non Cancel	-0.119	-0.021	-0.026	-0.049	0.006	0.064	-0.008	-0.078	0.026	-0.011	-0.001	0.197	0.539	0.468	1.000	-0.114	0.027	-0.060
Avg Room Price	0.297	0.338	-0.005	0.023	0.135	0.061	0.470	-0.063	0.179	0.054	0.018	-0.043	-0.175	-0.063	-0.114	1.000	0.184	0.143
# Special Requests	0.189	0.124	0.061	0.046	-0.090	0.088	0.145	-0.102	0.053	0.111	0.018	0.203	-0.012	-0.003	0.027	0.184	1.000	-0.253
Booking Status	0.087	0.033	0.062	0.093	0.049	-0.086	0.023	0.439	0.180	-0.011	0.011	-0.049	-0.107	-0.034	-0.060	0.143	-0.253	1.000

We can see from our correlation matrix below that the among the clear correlations we can see are:

1. '# Previous Non Cancellations' and 'Repeat Guest' have a correlation of .539 which may stem from the fact that those likely to not cancel are more loyal customers, being repeat guests in nature. Similarly, '# Previous Cancellations' and 'Repeat Guest' has a .391 correlation

2. We also see '# of Previous Cancels' has a .468 correlation with '# of Previous Non Cancels' which reflects the binary relationship between the 2 variables

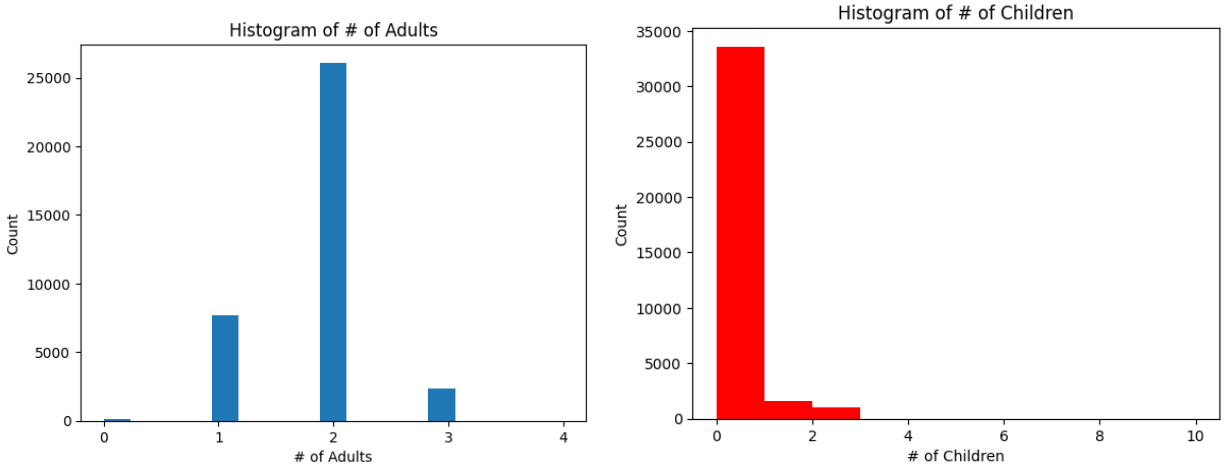
3. We can also see a ~.3 correlation between Market Segment Type and Repeat Guest, implying that certain hotel goers may be more prone to repeat than others

4. Average Room Price and Room Type has a .470 correlation which makes sense as we would assume a linear relationship between these two variables

5. Number of Children and Average Room Price has a .338 correlation which means we might have some relationship with more expensive rooms being necessary for larger families

d) Examining Relationship of Adults vs. Children

Next, we want to see how many children/adults are in each booking on average, by comparing the histograms for our variables '# Adults' and '# Children' below:

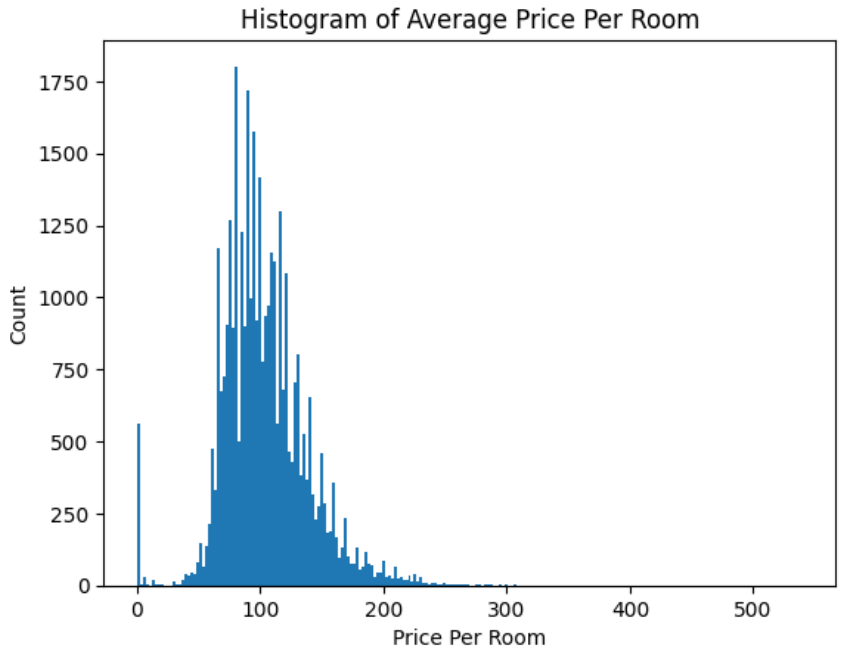


As can be seen from our '# of Children Histogram', we have a number of outliers, which extends the width of our plot to the right. Specifically, we have the following number of instances of children:

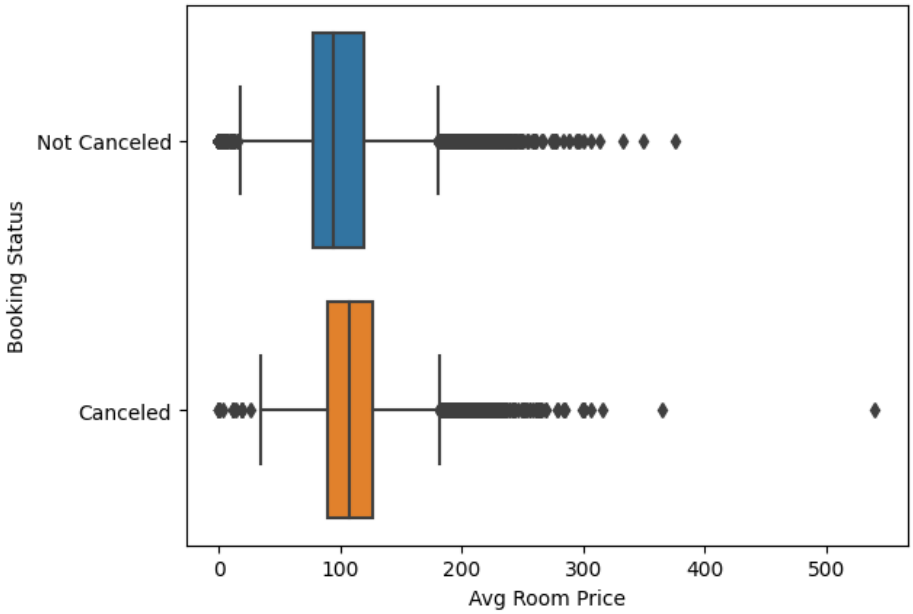
# of Children	*# of Occurrences in Dataset*
1	1618
2	1058
3	19
9	2
10	1

In general, we can assume an average of 2 adults and 1-2 children across all of the booking types, showing that our dataset isn't too skewed from national averages that we would assume to be ordinary (our hotel goers are not different from average).

e) Understanding of Average Price Per Room



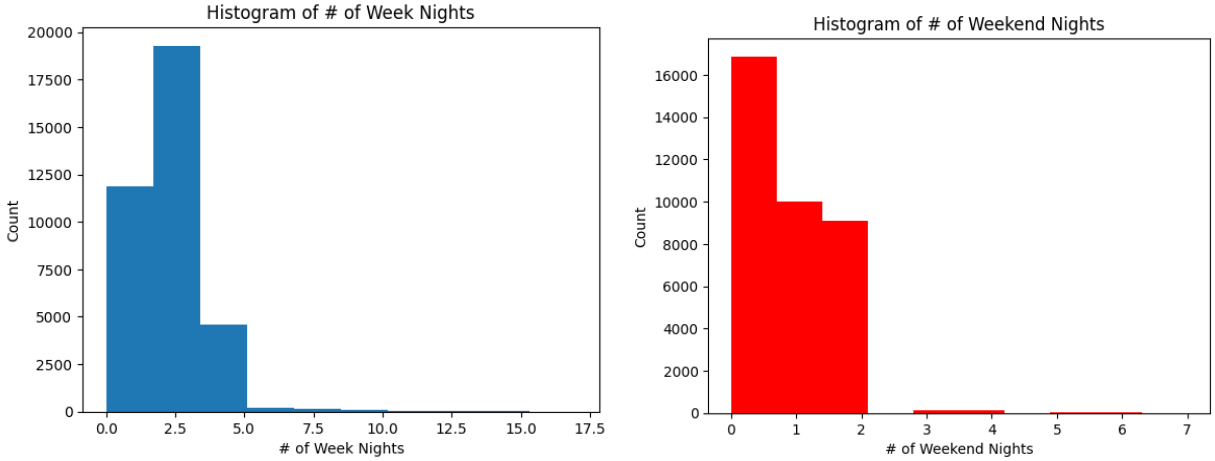
Above, we have shown a histogram of price per room, which shows an approximately normal distribution of around $100 per room reserved. This normal distribution approximation will allow us to better test for certain hypotheses surrounding average prices, using this assumption.



Similarly, we have shown that it appears that the average room price of canceled reservations is slightly higher than non canceled rooms. However, this finding is not significant because the entire 25th to 75th percentile of 'Canceled is not above that of 'Not Canceled'.

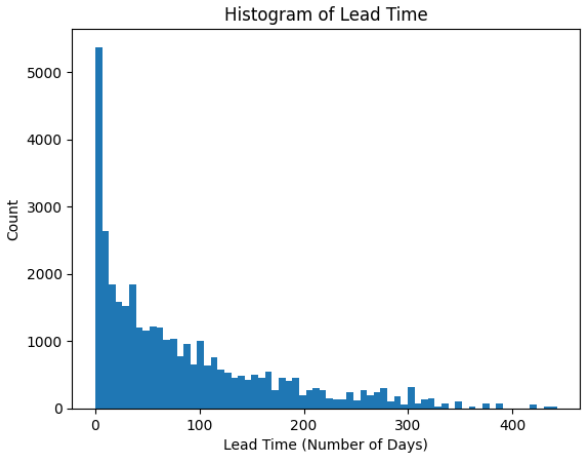
f) Understanding Week Nights vs. Weekend Nights

Next, we want to see how many week nights / weekend nights are in each booking on average, by comparing the histograms for our variables '# Week Nights and '# Weekend Nights below:



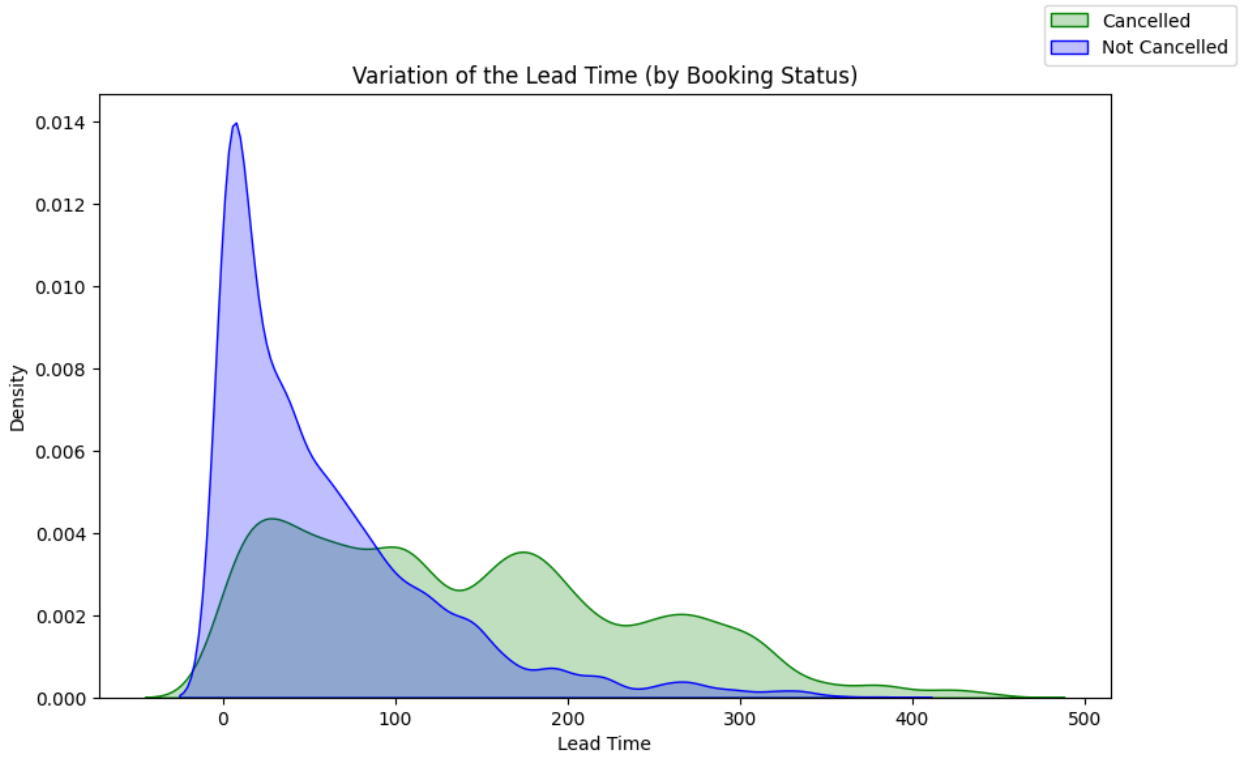
From the graph above, we can infer that most bookings are under a week long, and there is a very small proportion of bookings for over one week or multiple weekends. We can see that generally, ~3 week nights are booked on average, in addition to ~1 weekend night.

g) Understanding Lead Time



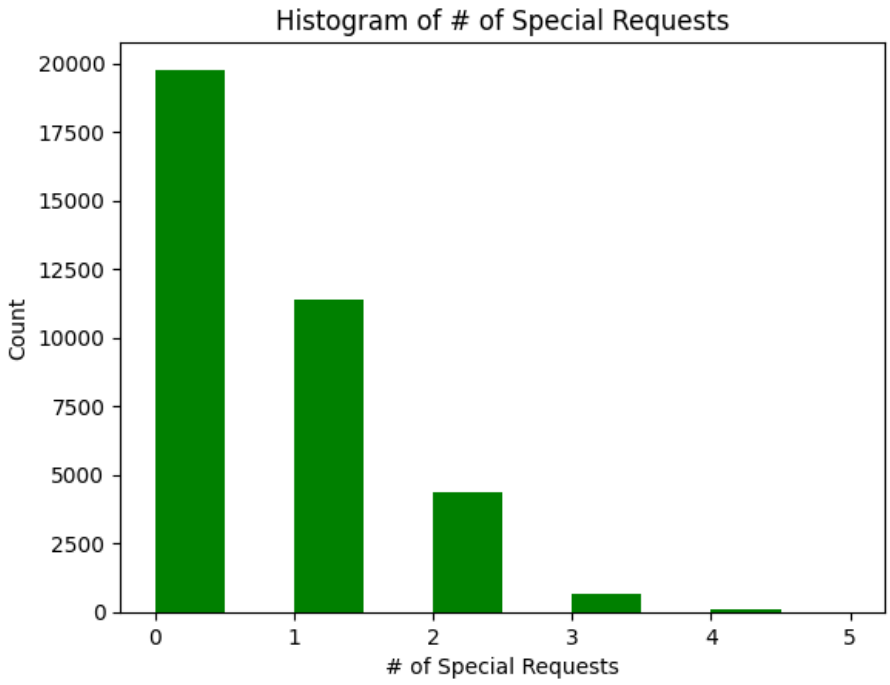
From our closer look at the lead time distribution above, we can see that our distribution is right skewed where most of the reservations are not booked that far in advance (lead time is measured in the number of days in advance)

Through our histogram of lead time above, we can see that as we have an increase in lead time, the total number of reservations continues to decrease and eventually diminishes entirely at a lead time of ~400 days.



Above, with our density plot of lead time (by booking status), we can similarly see once again that when there are smaller lead times, we have a generally higher chance of no cancellation. On the other hand, we can once again see in the case of lead time increasing, this does also lead to a higher chance of cancellation. These general observations will be explored in much more statistical detail later in our paper.

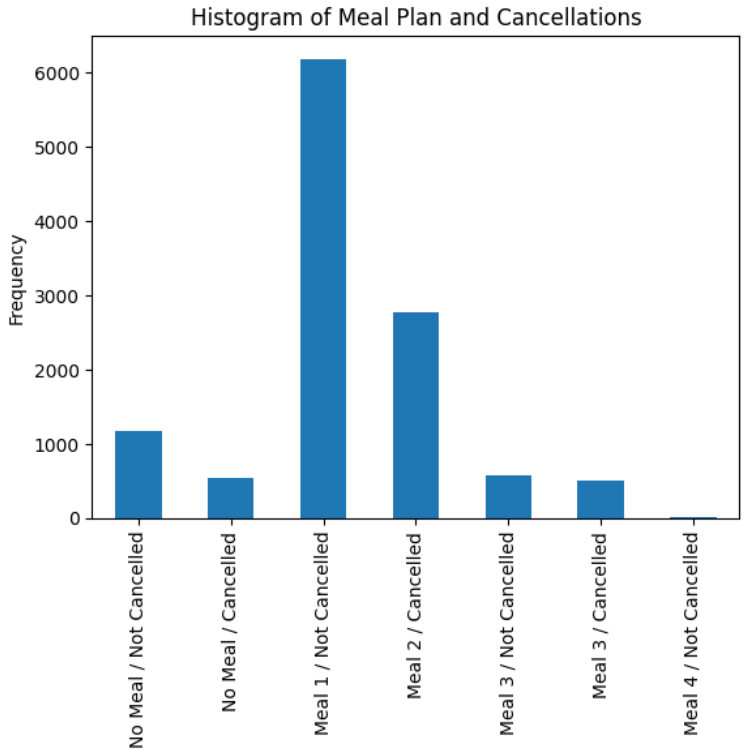
h) Understanding Special Requests



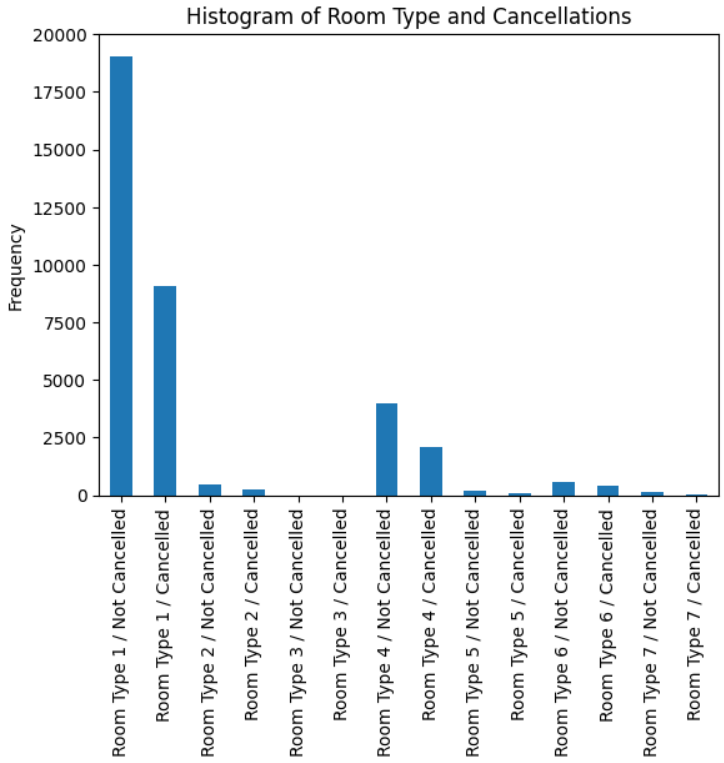
The number of special requests across reservations appears to drop off somewhat linearly with each additional request from 0 to 3. Generally, we can examine in most cases there aren't special request (or there are only one), so these may not be as large of a predictive factor in cancellation rates (aside from 0 vs. 1 requests, if any)

i) Categorical Data Exploration

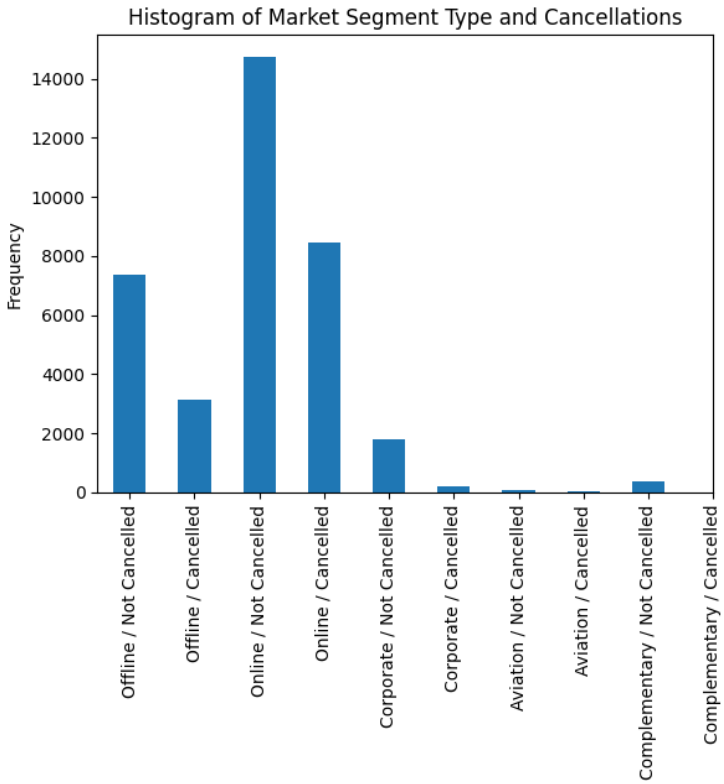
Finally, we seek to examine less-obvious variables and their distribution, when trying to understand potential implications on booking status and what these distributions could mean.



Histogram: Frequency of # of Cancellations vs. Meal Plan (0 - 4)



Histogram: Frequency of # of Cancellations vs. Room Type (1-7)



Histogram: Frequency of # of Cancellations vs. Market Segment Type (Offline, Online, Corporate, Aviation, Complementary)

In general, our data exploration has allowed us to make certain inferences about our variables and the effects they may have on both one another as well as on Booking Status being Canceled or Non-Canceled. We will take these inferences with us as we seek to answer our preliminary research questions and understand how various statistical techniques will allow us to better understand our data. This will ultimately allow us to offer the implications of a standard customer to the hotel for further analysis before really understanding how likely one is to cancel.

6. What are the factors that affect the average price per room in the hotel?

A. Introduction

One of the biggest points that drives how a hotel is conceived is its price. Rich and poor alike, everyone wants to find the best place they can stay at, but at the cheapest rate possible. It is often quite fascinating to learn what the different factors are that drive the price of products, and such analyses could prove to be helpful to both sides, the customers and the hotel management. The customers can make proper calls for their bookings and budget accordingly if they know how each component affects the price compared to either not choosing it or choosing another option. The hotel management too can look at such data, and while they are the ones who set the price, often there might be hidden patterns in such analyses which would help them make changes to their pricing. In this part of our analysis, we will look to answer the following questions:

1. **How do the different factors affect the average price per room?**
2. **Can we create a predictive model to predict the average price per room?**

Our dataset does not require any data cleaning, as mentioned in the above section. However, we need to make some changes for this section. One of the main things we observe is that the arrival year, months and dates are of integer type in the dataset. However, going forward, we will be looking at each year, month and date as different blocks and so, we convert them into factor variables. We also convert the column of repeated_guest to a factor variable as it is a binary(Yes/No) column. Also, going forward, we will be referring to the average price per room just as the price in certain parts of this section, for convenience.

B. Analysis

1. Full regression model

We start off by fitting a regression model of the average price per room(the independent variable) on all other variables in our dataset, excluding the Booking ID and Booking status, as we can logically deduce that these variables will not explicitly have an effect on the price. We look at a snippet of the output here and also elaborate on some interesting observations in this model:

	Estimate <dbl>	P_value <dbl>
(Intercept)	-4.143214e-13	9.735065e-49
no_of_adults	-1.217882e-13	2.289510e-301
no_of_children	-1.421742e-13	2.373410e-172
no_of_weekend_nights	2.962221e-14	4.556102e-67
no_of_week_nights	6.389787e-15	2.243316e-09
type_of_meal_planMeal Plan 2	-2.912233e-13	0.000000e+00
type_of_meal_planMeal Plan 3	1.786208e-13	1.454077e-01
type_of_meal_planNot Selected	1.585518e-13	1.347434e-251
required_car_parking_space	-1.065648e-13	1.022137e-36
room_type_reservedRoom_Type 2	1.202477e-13	3.946128e-28

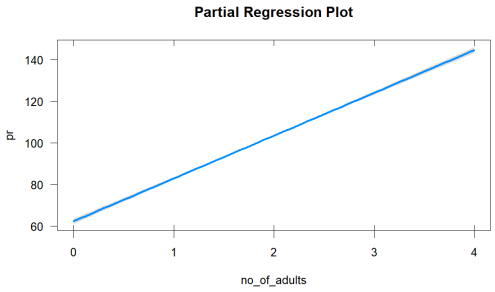
- Our model has 67 covariates in it. It is a large number due to the many levels of date and month in the data. Out of these 67, we find that 45 covariates are significant at 0.05 level of significance, when we do not account for multiple testing.
- To account for multiple testing, we use the Bonferroni correction as well as the FDR method at 5%. On using Bonferroni correction, we see that the new p-value cutoff is 0.0007462687. At this level, we find that 35 out of the 67 variables are significant. However, on using the FDR method at 5%, we see that the new cutoff is 0.02879152, at which level, we find 42 covariates significant. This shows that even after correcting for multiple testing, our number of significant variables do not change much, which is also evident from the regression summary, as most variables have an extremely low p-value.
- Another point to note is that our R^2 and adjusted R^2 values are 1. This happens because we have a large number of covariates, a huge proportion of which are significant, and so, these factors are able to completely explain all variability in the independent variable.
- We also notice that all the covariates have coefficients that are extremely close to 0. We think this must be due to the fact that all these variables have an effect on the price, and the high number of covariates, while decreasing bias, introduces a high amount of variance in the coefficients.

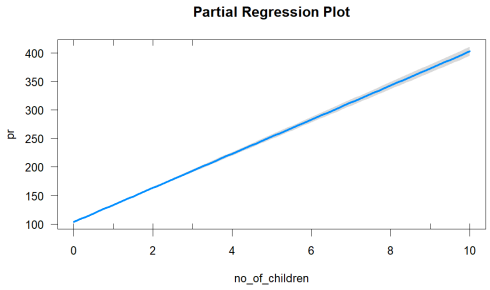
2. Marginal and Section regression models

For interpretability purposes and to better understand the effect of each variable on the price, we model the average price on each variable or groups of similar variables separately. We use the same variables as the ones above, except for the date, as we feel that the prices for each date can be heavily dependent on the month too and may not have a proper explanation. Also, it helps us reduce the number of variables used in the model drastically. After running regression on all these marginal and sectional variables, we find some intriguing insights, and some outputs as we would expect.

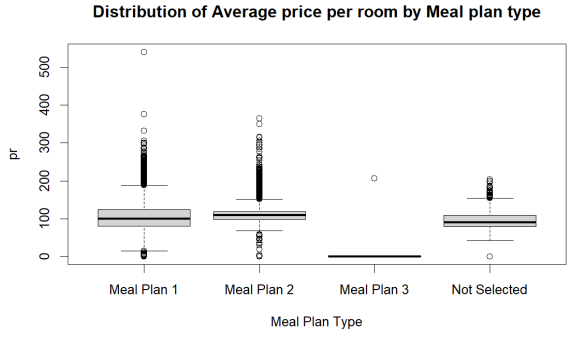
- One of the biggest observations is that the R^2 value for all our models is extremely low. Some are even close to 0 and the top 2 R^2 values are around 0.25. This is because obviously, one or 2 variables by themselves cannot explain the variability in the independent variable. However, since this part only focuses on the effect of each covariate, our regression models had almost all variables as significant, and so it is only the coefficients of significant covariates that matter here.

- On running our regression on the number of adults and the number of children, we see that as expected, we see that both covariates are quite significant and have a positive value. We notice that the coefficient for the number of children is actually higher than the number of adults. This could be attributed to the fact that when kids are present, they will almost always be accompanied by their parents, while adults don't necessarily need to be accompanied by their kids. Due to this, while running a regression, rows where kids are present would tend to have a higher average cost compared to just adults. Another point we must keep in mind is that in this case, our variables do not have very high values. As the number of people increases drastically, it will lead to more rooms being taken and while the total cost would increase, the average price per room will not increase by as large a margin as our current coefficient of around 20 euros.





- On regressing with meal plan types as our independent variable, we notice that keeping all else constant, bookings made with meal plan 2 are the most expensive meal type as per our model and ones with meal plan 3 are the cheapest. However, meal plan 3 is actually cheaper by an extremely large margin which pushes us to find out why it might be so.



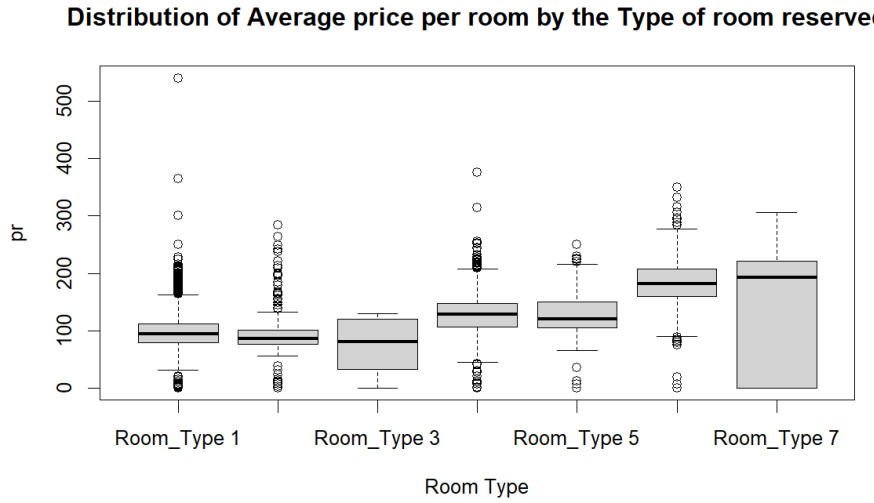
Type of Meal Plan <fctr>	Count <dbl>
Meal Plan 1	27835
Meal Plan 2	3305
Meal Plan 3	5
Not Selected	5130
Sum	36275

As we can see in the plots, meal plan 3 has average price per room always close to 0, except for an extremely high outlier. This could imply that Meal plan 3 could be a part of the complimentary service by the hotel. Also, it is worth noting that in our dataset, only 5 bookings have chosen Meal Plan 3, so we do not have a large enough sample size to conclude accurately.

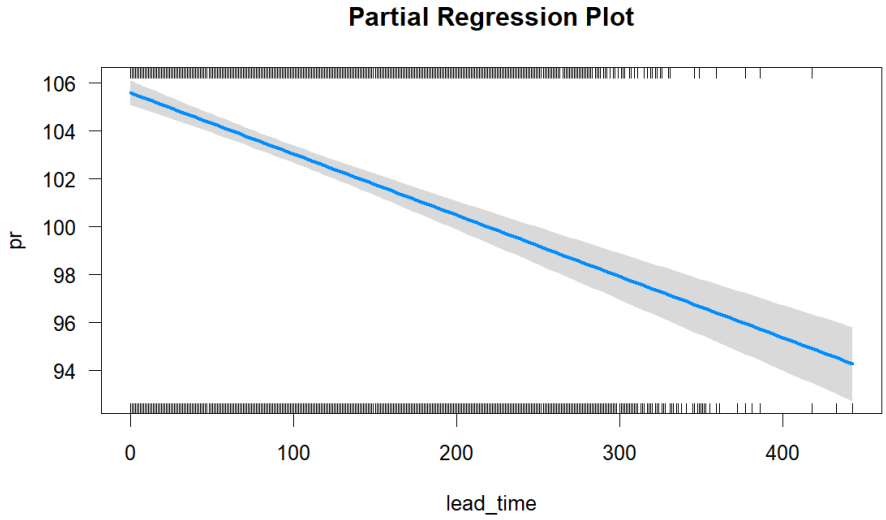
- We see that customers who had required a car parking space had to pay about 12.41 euros more than other customers, keeping all else constant. This would mean that the average

cost of a parking space at the hotel would be expected to be 12.41 euros.

- On regressing with Type of room, we see that all the covariates are actually significant. For room type 3, even though it is greater than 0.05, it is just slightly greater so we will consider it to be significant too. Since this is a categorical covariate, we can see that room type 1 is the baseline category and on further analysis, we see that on average, room type 3 has the most negative value, which indicates it is the cheapest type of room. Also, room 6 has the most positive value, which indicates that it is the most expensive one.



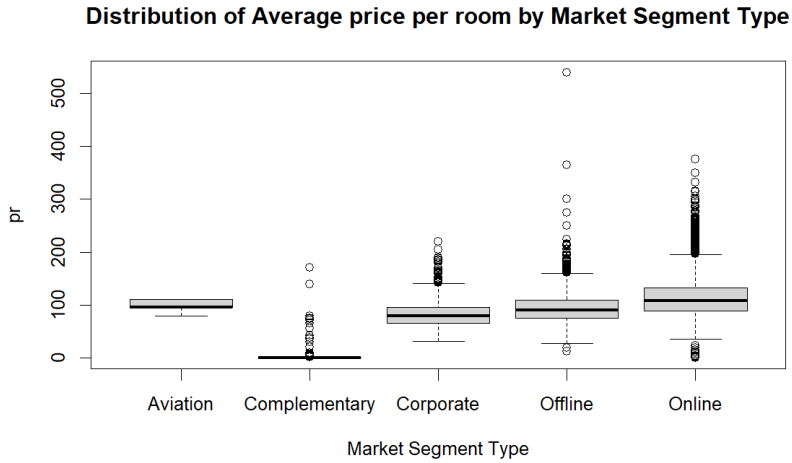
- As is the common trend, we see that the lead time is significant and has a negative value. This makes sense from our real-life experiences where booking accommodations well in advance helps us get it at cheaper rates. In this case, keeping all else constant, the average price goes down by 0.02556 euros for each day in advance that the booking is made.



- On running our regression with the arrival month and arrival year, we observe that the

year 2018 has an average price that is higher than 2017 by 22.5659. This is not entirely surprising and might be due to inflation. Also, if we look at the months, we notice that May to September/October tend to have the higher prices. This makes sense as the summer is a time when people prefer to go for vacations. Due to children having a break from school, the weather being nice, and other factors, summer time is peak vacation time, which means more tourists are attracted to the hotel, and the prices are driven up to improve sales.

- The market segment regression also gives us an output that we would have expected. One important thing to notice is that offline has a lower value on average than online. This happens mainly because, while booking online, most people use third-party sites that actually charge service fees, which drives the cost up. On average, keeping all else constant, we can see that the cost of booking online is actually about 20.6 euros more expensive than booking it offline.



- We club the repeated guest, and number of previous bookings that were canceled and not canceled together for our model. Here, again we see trends that we would normally expect. Repeat guests actually tend to have cheaper rooms compared to first-timers, with a discount of about 36 euros on average. The number of previous cancellations drives the price upwards due to unreliability from their past experiences, and not canceled bookings help reduce the cost since these customers seem reliable and have had a possible good experience and relation with the hotel if they are coming back again.

These were our significant findings from the marginal and sectional regression analysis. Having done this, we now try to create a predictive model for the average price per room.

3. Predictive regression using linear models

We firstly prepare our data by splitting it into training and testing sets, to cross-validate our model, and find a predictive model that would be a good fit for our data.

We split the data randomly to avoid selective inference issues, and split it in a 70-30 split for the training to testing set. We see that it yield the following model:

	Estimate	P_value
(Intercept)	33.61843802	1.847204e-33
no_of_adults	9.45938299	4.074149e-197
no_of_children	11.30528048	3.757604e-118
no_of_weekend_nights	-2.38814352	8.392570e-47
no_of_week_nights	-0.43017667	3.261324e-05
type_of_meal_planMeal Plan 2	23.55837040	0.000000e+00
type_of_meal_planMeal Plan 3	-26.01426795	9.944659e-02
type_of_meal_planNot Selected	-12.99059950	1.418269e-182
required_car_parking_space	9.11548556	1.000994e-28
room_type_reservedRoom_Type 2	-8.28616323	3.196652e-14
room_type_reservedRoom_Type 3	7.01515738	5.277042e-01
room_type_reservedRoom_Type 4	15.52090143	1.653874e-283
room_type_reservedRoom_Type 5	26.00441653	1.104627e-57
room_type_reservedRoom_Type 6	53.67061666	0.000000e+00
room_type_reservedRoom_Type 7	61.42526475	4.065377e-185

	Estimate	P_value
lead_time	-0.08691214	0.000000e+00
arrival_year2018	21.66912302	0.000000e+00
arrival_month2	2.45105789	2.118513e-02
arrival_month3	10.55507262	1.200242e-25
arrival_month4	19.73498553	8.611875e-87
arrival_month5	35.12643870	1.144112e-263
arrival_month6	36.78717586	5.479155e-300
arrival_month7	33.15332322	9.548819e-234
arrival_month8	34.84803825	2.797587e-272
arrival_month9	46.00368122	0.000000e+00
arrival_month10	36.93623953	6.931741e-320
arrival_month11	21.61094414	2.138942e-104
arrival_month12	16.91486847	2.981505e-64
arrival_date2	2.04171680	5.843628e-02
arrival_date3	1.02539090	3.561788e-01

	Estimate	P_value
arrival_date4	-0.05124615	9.616587e-01
arrival_date5	-0.05706903	9.589889e-01
arrival_date6	3.47482648	1.385073e-03
arrival_date7	6.08134035	4.988477e-08
arrival_date8	-0.80210542	4.673053e-01
arrival_date9	0.09043709	9.354803e-01
arrival_date10	-1.24631302	2.642441e-01
arrival_date11	-0.95282928	3.958460e-01
arrival_date12	-1.53655906	1.606293e-01
arrival_date13	-1.78843879	9.273680e-02
arrival_date14	0.58791907	5.901291e-01
arrival_date15	5.97420262	3.194498e-08
arrival_date16	-3.15982965	3.504076e-03
arrival_date17	-0.16468411	8.773934e-01
arrival_date18	-3.28969984	2.543741e-03

	Estimate	P_value
arrival_date19	6.25052948	5.094276e-09
arrival_date20	1.61319496	1.394953e-01
arrival_date21	-1.97417780	7.152932e-02
arrival_date22	0.74760594	5.137241e-01
arrival_date23	2.97337366	1.035559e-02
arrival_date24	2.50466160	2.517718e-02
arrival_date25	3.18890300	3.978634e-03
arrival_date26	-2.42736473	2.832381e-02
arrival_date27	-1.20940793	2.851768e-01
arrival_date28	-0.28026465	8.020091e-01
arrival_date29	0.86400054	4.388601e-01
arrival_date30	2.03873402	6.137864e-02
arrival_date31	2.99681199	2.782836e-02
market_segment_typeComplementary	-93.11721471	1.610230e-235
market_segment_typeCorporate	-0.03403823	9.892439e-01

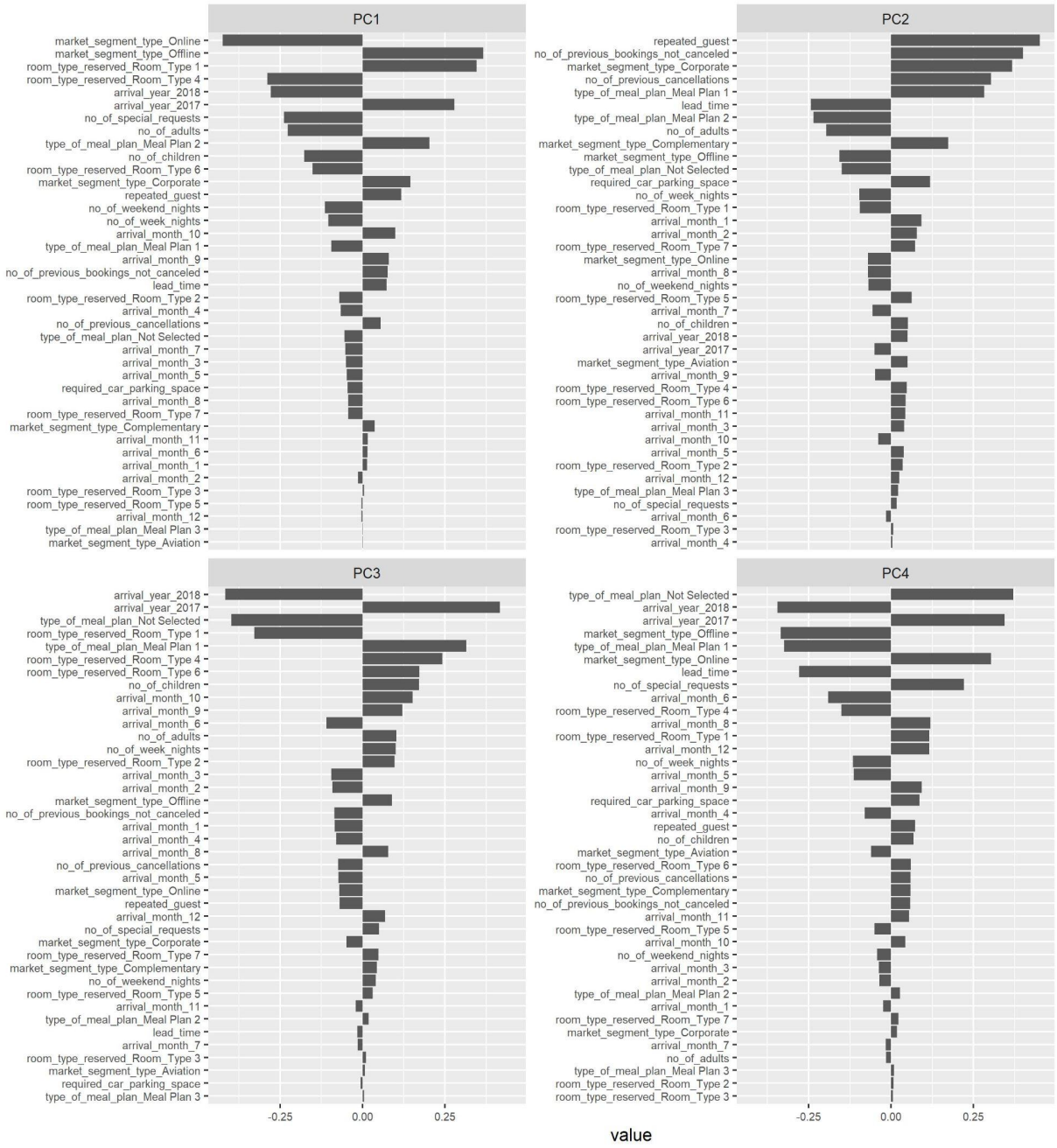
	Estimate	P_value
market_segment_typeOffline	1.5289496	5.383179e-01
market_segment_typeOnline	15.2224962	7.649619e-10
repeated_guest1	-10.4820082	1.506452e-19
no_of_previous_cancellations	0.9530696	2.978428e-02
no_of_previous_bookings_not_canceled	-0.5864740	2.675643e-09
no_of_special_requests	0.7638622	1.188133e-04

The above model is our predictive model. As we can see almost all covariates are significant, except for a select few, most of which are actually a part of the date variable. We also notice that this model has an adjusted R^2 value of 0.6042. On cross-validating this with our testing data, we see that the R^2 value of our predictions is 0.6036. This means that this predictive model is a good one and can be used to predict the average price per room, if we have the entire data.

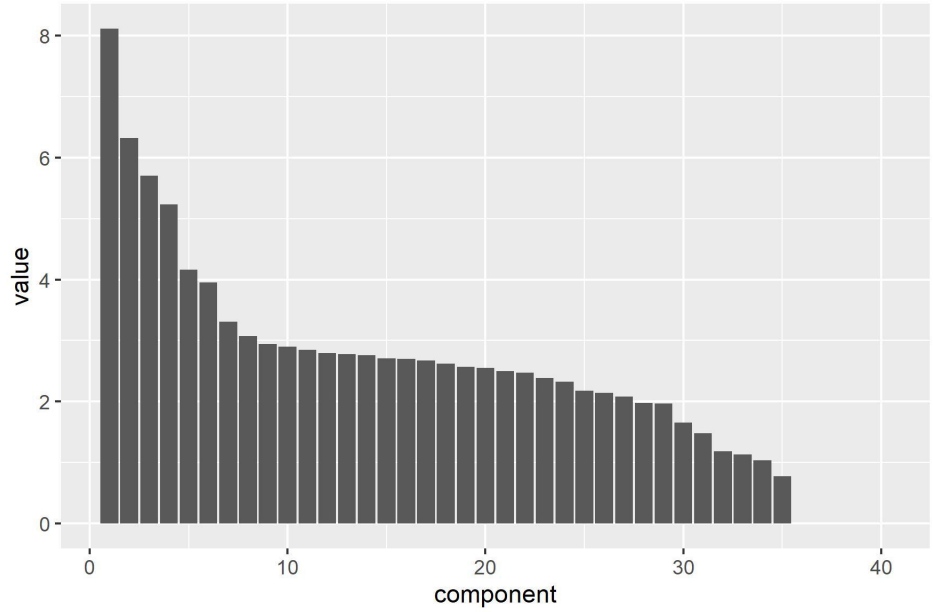
4. Regression using PCA

To deal with high dimensionality in our model, and also correct for other factors like multicollinearity, we use PCA in this section. For ease of fitting the model, we exclude the arrival date variable from our model.

After preparing our data and applying the PCA tools, we find the rotations. We see that the first 4 rotations can be illustrated as shown by the chart below:



Since we have so many variables in our case, we will not try to interpret them, but in turn just try to find the appropriate model to fit. Next, we compare the percentage variances explained by the components and they are illustrated in the chart below:



As we can see in the above graph, there seems to be a decent fall up till component 4 and then 5 to 6 is a very small drop. So we choose the first 4 components for our model. The model summary for this one is as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.4235     0.1593 649.418  < 2e-16 ***
PC1          -8.7238     0.0884 -98.682  < 2e-16 ***
PC2          -3.0601     0.1001 -30.561  < 2e-16 ***
PC3           4.2068     0.1055  39.879  < 2e-16 ***
PC4          -0.3988     0.1101  -3.621 0.000294 ***
```

C. Conclusion

From our analysis in this section, we have seen how each variable contributes to the average price per room. It is worth noticing that almost all the variables we have in our data contribute as a significant covariate in our regression models. Through this section, we have also observed the differences in FDR and Bonferroni correction and how they correct for multiple testing in different ways and provide different outcomes. We also saw how a marginal or sectional regression can be a very useful tool while trying to observe the effect of each individual or small group of variables, as compared to a full model. We were also able to create a predictive model with a good level of accuracy that can be used to predict future prices, provided the economic and other factors remain constant.

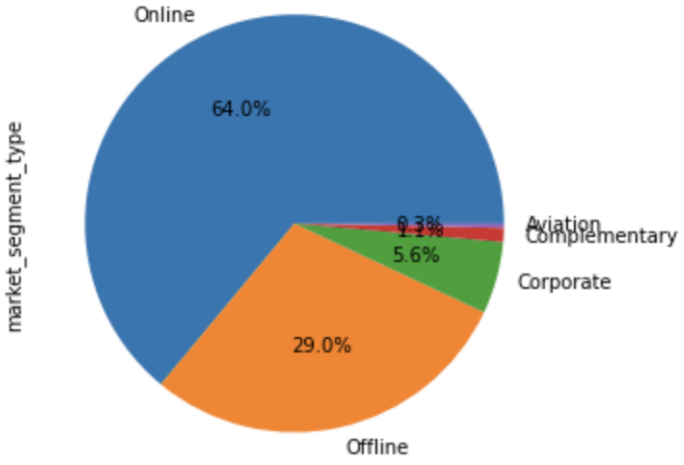
7. How do booking preferences differ among distinct market segments?

A. Introduction

Examining the varying booking preferences across distinct market segments presents a fascinating question, one that delves into the intricacies of customer behavior and market dynamics. From the length of the stay to the type of room booked, these preferences may reveal unique patterns among the different market segments. To ensure a comprehensive analysis, the research question has been deconstructed into three significant questions:

1. **Are there differences in the length of stay among various market segments?**
2. **How do cancellation patterns and repeat bookings differ between market segments?**
3. **What are the common booking preferences for each market segment?**

First, to gain a foundational understanding of the dataset's composition, we will construct a pie chart representing the different market segments - the independent variables in our research question. A clear visual understanding of the proportions of these segments will facilitate a more nuanced interpretation of subsequent analyses.



The distribution of bookings across the different market segments is quite uneven.

- A large majority of bookings are made through the Online segment, indicating that customers prefer to book online, likely due to the convenience, efficiency, and possibly wider range of options available online.

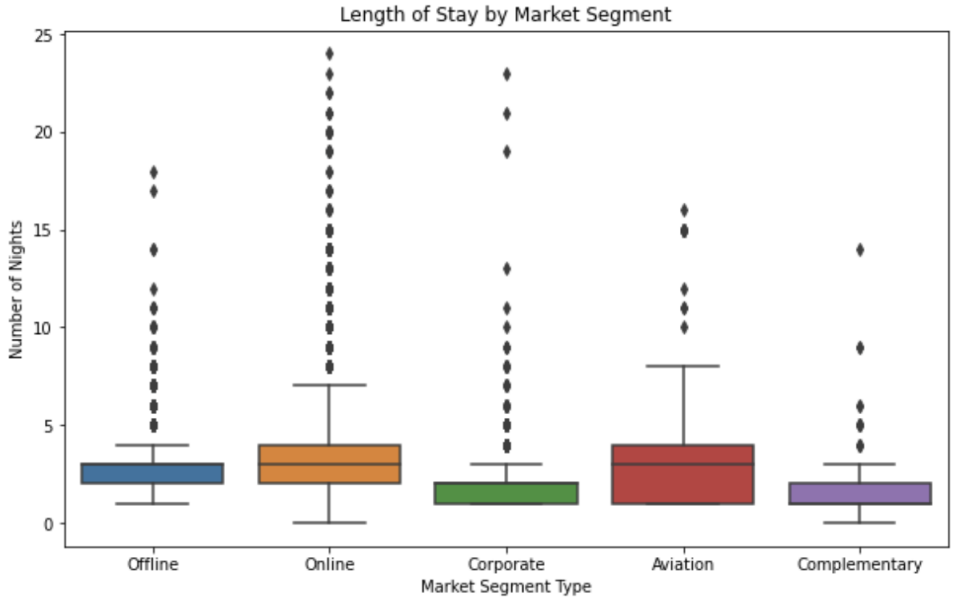
- The Offline and Corporate segments represent the second and third largest portions of bookings, respectively. This suggests that while online bookings are preferred, a significant number of customers still make bookings offline, possibly through a travel agency, direct hotel contact, or other traditional booking methods. The corporate segment might represent business travelers, indicating a significant proportion of the customers are traveling for work or business purposes.

- The Aviation and Complementary segments constitute a small fraction of the total bookings. Bookings from the Aviation segment might indicate partnerships with airlines or bookings made by airline staff. The Complementary segment could represent bookings made using loyalty points, gift cards, or other complimentary offers.

The choice of dependent variables will vary with each sub-question. By utilizing a diverse set of dependent variables, we aim to thoroughly examine the intricate relationships between these variables and the market segments. Ultimately, this comprehensive approach will provide a deeper understanding of how booking preferences differ among distinct market segments.

B. Analysis

1. Are there differences in the length of stay among various market segments?



	count	mean	std	min	25%	50%	75%	max	IQR
Offline	10,528	2.91	1.36	1	2	3	3	18	1.00
Online	23,214	3.18	1.92	0	2	3	4	24	2.00
Corporate	2,017	1.92	1.39	1	1	2	2	23	1.00
Aviation	125	4.02	4.18	1	1	3	4	16	3.00
Complementary	391	1.57	1.22	0	1	1	2	14	1.00

Based on the boxplot with the summary table above, the length of stay varies across different market segments, with the Aviation segment having the longest average stay and the Complementary segment having the shortest. There's considerable variability within each segment, especially the Aviation segment. This information could be useful for the hotel to better anticipate the needs of guests from different market segments.

2. How do cancellation patterns and repeat bookings differ between market segments?

Booking cancellations can impact a hotel's revenue and room management, causing disruptions in scheduling and potential losses due to unoccupied rooms. On the other hand, repeat bookings are a positive indicator of customer loyalty and satisfaction, representing a reliable source of revenue. Given the implications of both phenomena, identifying the patterns and differences in booking cancellations and repeated bookings across various market segments is a valuable research endeavor.

It is necessary to check if the cancellation rate and repeat booking rate have any relationship with each other. It could be possible that segments with higher cancellation rates also have higher repeat booking rates, as customers might be more likely to cancel and rebook their reservations in these segments.

To investigate if there's a relationship between cancellation rate and repeat booking rate, we can calculate the correlation between these two rates.

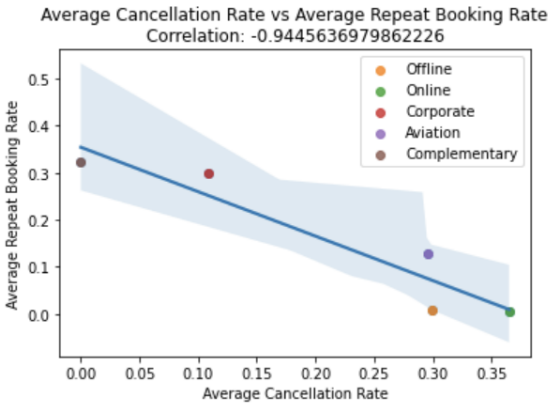
market_segment_type	Aviation	Complementary	Corporate	Offline	Online
Cancellation Rate	0.296000	0.000000	0.109073	0.299487	0.365081
Repeat Booking Rate	0.128000	0.322251	0.298463	0.008549	0.004135

From the summary table above, it appears that the Complementary and Corporate segments have lower cancellation rates and higher repeat booking rates, indicating high customer loyalty and commitment to bookings. The 0% cancellation rate in this segment might be attributed to the nature of complimentary bookings. Complementary customers who receive complimentary services or perks are less likely to cancel since they are not financially invested

in the booking. The high repeat booking rate could be driven by customer satisfaction with the complimentary services, leading them to return for future bookings. The lower cancellation rate in the corporate segment might be due to corporate policies or agreements that discourage cancellations. Companies often have travel policies in place that require employees to follow through with their bookings unless there are exceptional circumstances. The higher repeat booking rate suggests that corporate travelers have consistent travel needs and tend to choose the same hotel for their business trips, contributing to higher loyalty.

On the other hand, the Offline and Online segments have higher cancellation rates and lower repeat booking rates, suggesting lower customer loyalty and commitment. The higher cancellation rate in the offline segment could be due to the more traditional booking process involved, such as making reservations through travel agents or contacting hotels directly. Customers in this segment might have more flexibility to cancel or change their plans, resulting in a higher cancellation rate. The lower repeat booking rate might be attributed to a lack of personalized marketing or loyalty programs targeting offline bookers, leading to reduced customer retention. The higher cancellation rate in the online segment could be influenced by various factors. Online bookings often offer flexibility and the ability to easily compare options, leading customers to make multiple bookings and cancel those that are less desirable. Additionally, the convenience of online bookings might make it easier for customers to change or cancel their plans. The lower repeat booking rate could be due to the vast number of options available online, making it challenging for hotels to establish customer loyalty or incentivize repeat bookings.

The Aviation segment stands in the middle with a relatively high cancellation rate but also a notable repeat booking rate. This might be associated with the dynamic nature of aviation crew schedules. Flight schedules can change unexpectedly, leading to necessary cancellations. However, the notable repeat booking rate suggests that crew members might have regular layovers or recurrent visits to specific locations, leading to repeat bookings despite the higher cancellation rate.



By adding a regression line on the scatterplot between average cancellation rate and average repeat booking rate, we can see a clear negative correlation between the two variables by the negative slope. By calculation, the correlation between the average cancellation rate and average repeat booking rate is approximately -0.945. This indeed indicates a significant negative correlation, which means that as the average cancellation rate increases, the average repeat booking rate tends to decrease, and vice versa.

This might suggest that market segments with higher average cancellation rates have fewer repeat guests, perhaps because these guests have less commitment to their bookings, while market segments with more average repeat guests have lower average cancellation rates, perhaps because these guests are more loyal and thus less likely to cancel their bookings.

3. What are the common booking preferences for each market segment?

Previous analyses have already provided insights into two important aspects related to market segments: the differences in the length of stay and the variations in cancellation patterns and repeat bookings. These findings have shed light on how different market segments behave in terms of stay duration and booking behavior. Now, it is essential to delve further into the examination of booking preferences, as this knowledge will contribute to a more comprehensive understanding of the distinct characteristics and requirements of each segment.

Understanding and analyzing booking preferences across different market segments is crucial for hotels and hospitality businesses to tailor their offerings and services effectively. By examining the common booking preferences for each market segment, hotels can better meet the specific needs and expectations of their diverse customer base.

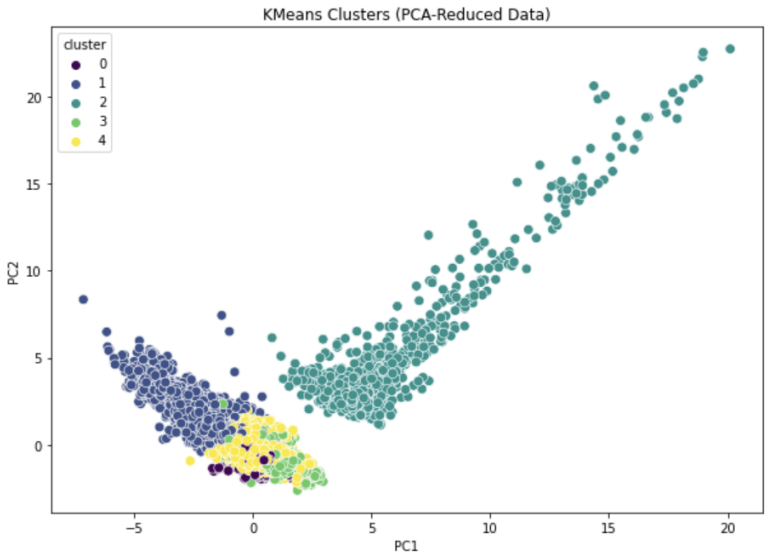
Here, we could use a mixture of descriptive statistics and clustering. We can describe booking preferences by market segment, and then use clustering to identify groups of market segments with similar preferences.

market_segment_type	room_type_reserved	type_of_meal_plan	no_of_special_requests	required_car_parking_space	no_of_adults	no_of_children
Aviation	Room_Type 4	Meal Plan 1	0.000	0.048	1.016	0.000
Complementary	Room_Type 1	Meal Plan 1	0.882	0.079	1.483	0.125
Corporate	Room_Type 1	Meal Plan 1	0.222	0.091	1.230	0.010
Offline	Room_Type 1	Meal Plan 1	0.203	0.003	1.778	0.021
Online	Room_Type 1	Meal Plan 1	0.842	0.037	1.939	0.152

market_segment_type	no_of_weekend_nights	no_of_week_nights	lead_time	arrival_year	arrival_month	repeated_guest	no_of_previous_cancellations
Aviation	1.160	2.856	5.488	2,018.000	7.120	0.128	0.040
Complementary	0.330	1.240	12.036	2,017.645	7.724	0.322	0.210
Corporate	0.428	1.489	21.818	2,017.753	7.104	0.298	0.167
Offline	0.731	2.181	122.873	2,017.722	7.573	0.009	0.011
Online	0.887	2.290	75.334	2,017.873	7.380	0.004	0.013

Interpreting the data in the table, we can identify several distinct characteristics across various market segments. For instance, the vast majority of guests, regardless of segment, prefer Room_Type 1 and Meal Plan 1, with the exception of those in the Aviation segment who show a preference for Room_Type 4. This pattern suggests that Room_Type 1 and Meal Plan 1 cater to a wide range of guests' needs, while the preference for Room_Type 4 in the Aviation segment could be indicative of particular requirements or preferences inherent to these guests.

We chose to do a clustering analysis on 5 clusters, which is the number of unique market segments. We hope to figure out the common booking preferences by seeing a distinct difference among five clusters.

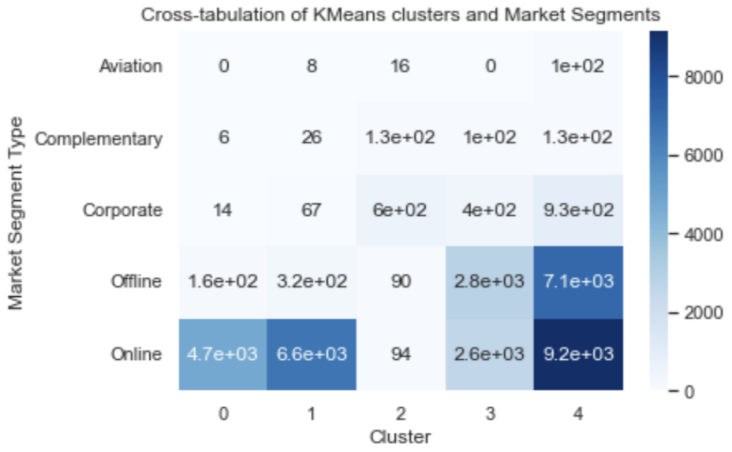


From the scatterplot above, we can clearly see that there is a large group of data points concentrated on the right side, which distinguishes itself with a very clear boundary from the rest. On the other hand, the data points on the left side of the graph are closer to each other, suggesting they might form one or multiple denser clusters. To better understand the structure and differentiate between the clusters, we can proceed with additional statistical analysis.

no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time
1.875	0.018	0.787	1.953	2.999	0.027	0.029	60.592
2.226	0.385	0.956	2.529	0.068	0.053	2.995	68.580
1.228	0.012	0.446	1.330	0.109	0.149	0.470	13.047
1.758	0.025	0.708	2.126	0.287	0.023	0.290	63.245
1.745	0.049	0.813	2.218	0.093	0.020	0.132	110.300

arrival_year	arrival_month	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests
2,017.964	7.323	0.000	0.000	-0.000	95.143	0.715
2,017.974	7.317	0.000	-0.000	0.000	143.523	1.007
2,017.777	7.415	1.000	0.914	6.000	65.219	0.558
2,017.000	9.658	0.000	0.000	-0.000	88.337	0.503
2,018.000	6.733	0.000	-0.000	-0.000	96.723	0.479

The table presents the centroid values for five distinct clusters, representing different guest groups identified using the KMeans algorithm based on their booking and staying behavior in the hotel.



The cross-tabulation table shown above describes the distribution of clusters for each market segment type. This table helps visualize and understand how the different customers' hotel booking preferences are distributed across the various market segments. It shows that hotel reservations in cluster 0, 1, and 4 are mainly from Online customers, hotel reservations in cluster 2 are mainly from Corporate customers, and hotel reservations in cluster 3 are mainly from Offline customers. Even though it might seem unreasonable that the clusters only predict three market segments, the clustering analysis is highly reliable since we have found out that Online, Offline, and Corporate are the main market segments in the dataset that might act a larger role comparing to Complementary and Aviation at the beginning of the data preparation.

C. Conclusion

The comprehensive study of varying booking preferences across distinct market segments reveals compelling patterns and key differences. Our findings demonstrate that distinct market segments have unique characteristics with respect to length of stay, booking cancellations and repeat bookings, and specific preferences such as room type, meal plan, and special requests.

Our cluster analysis further underscored these differences, identifying five distinct guest groups based on their booking and stay behavior. The distribution of these clusters across market segments showed that hotel reservations were mainly from Online customers, Corporate customers, and Offline customers with distinct booking preferences. These findings can be advantageous for hotels to understand and cater to their diverse clientele more effectively.

8. What reservation characteristics lead to hotel cancellation?

A. Introduction

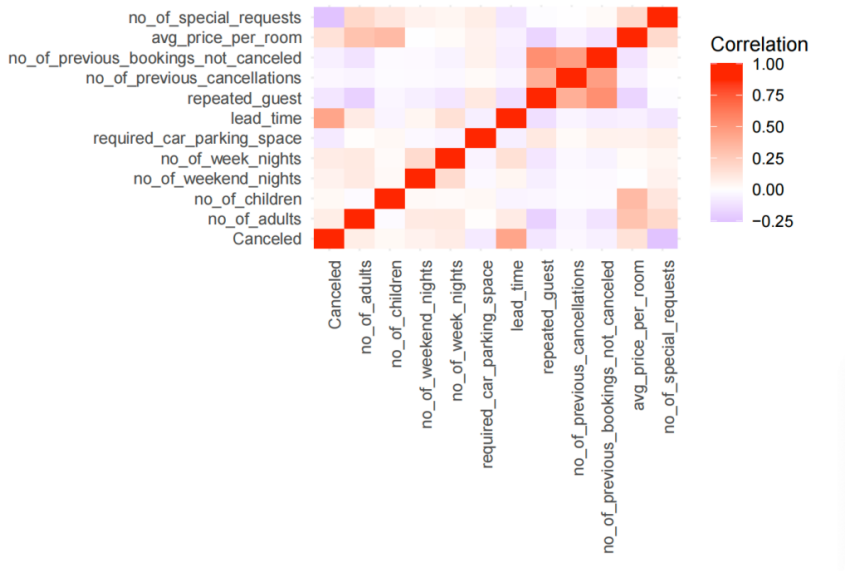
Hotel cancellations can significantly impact a hotel's revenue. Understanding the patterns and factors influencing cancellations is crucial for hotels to minimize cancellations and optimize revenue. This part employs logistic regression analysis to explore the reservation characteristics that lead to hotel cancellations. Correlation analysis is initially conducted to identify the variables correlated with cancellations. Logistic regression is then performed to determine statistically significant predictors of cancellation. Principal component analysis (PCA) is utilized to reduce complexity and address multicollinearity. Based on the PCA results, additional variables with moderate effects on cancellations are added to the logistic regression model. Finally, the accuracy of the logistic regression model is compared with logistic regression model with interactions and lasso, and other machine learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree Classifier, and Random Forest Classifier.

B. Analysis

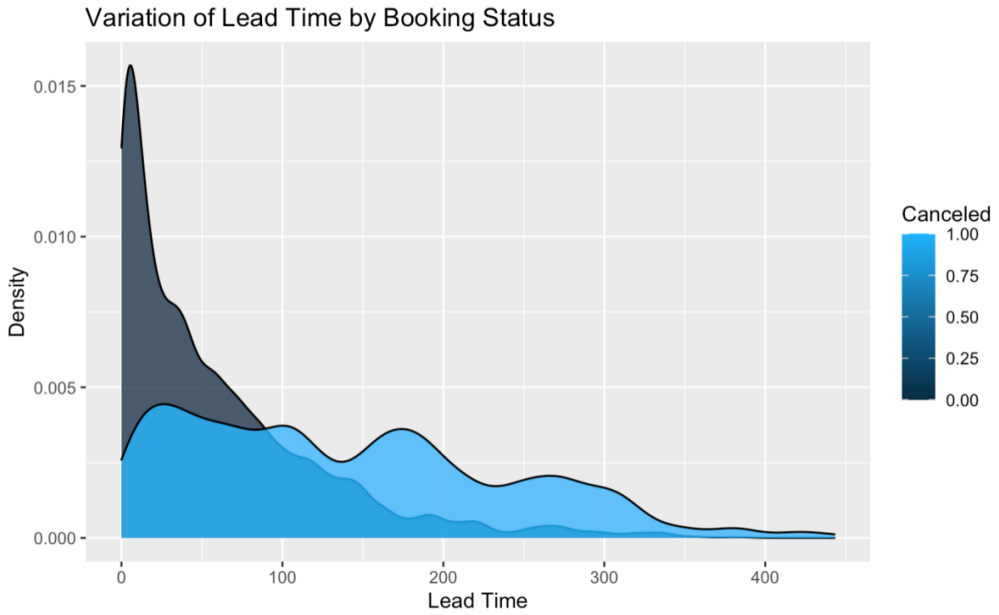
1. Correlation Analysis

Correlation analysis is conducted to explore the relationship between reservation characteristics and cancellations. The analysis reveals several observations: lead time exhibits the strongest positive correlation with cancellations. Additionally, the number of adults, number of weekday nights, and average price per room display positive correlations, while the number of special requests, requirement for car parking space, and repeated guest status show negative relationships with cancellations.

```
booking_status                          1.000000
lead_time                               0.438538
no_of_special_requests                  0.253070
arrival_year                            0.179529
avg_price_per_room                      0.142569
repeated_guest                          0.107287
no_of_week_nights                       0.092996
no_of_adults                            0.086920
required_car_parking_space              0.086185
no_of_weekend_nights                    0.061563
no_of_previous_bookings_not_canceled    0.060179
no_of_previous_cancellations            0.033728
no_of_children                          0.033078
arrival_month                           0.011233
arrival_date                            0.010629
```



Due to the strong magnitude of correlation effect, we decide to dig into the impact of lead time on cancellation.



The plot reveals the disparity in the density of lead time values between different booking statuses. Specifically, it illustrates that the distribution of lead time for not-canceled reservations is mainly concentrated within the range of less than 100 days. On the other hand, for canceled reservations, the lead time distribution spans a broader range of 0 to 300 days. This observation suggests that there may be a relationship between lead time and the likelihood of a reservation being canceled -- a larger number of days between the date of booking and the arrival date may tend to lead to cancellation.

2. Logistic Regression

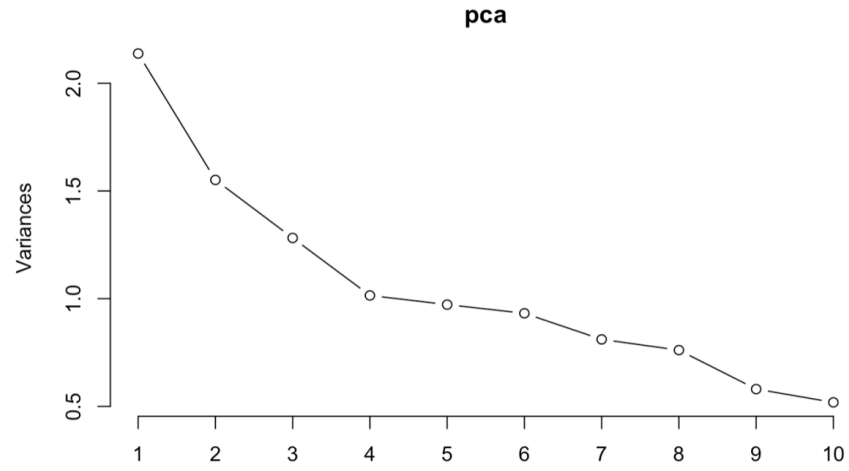
To investigate the predictors of cancellations, logistic regression is performed.

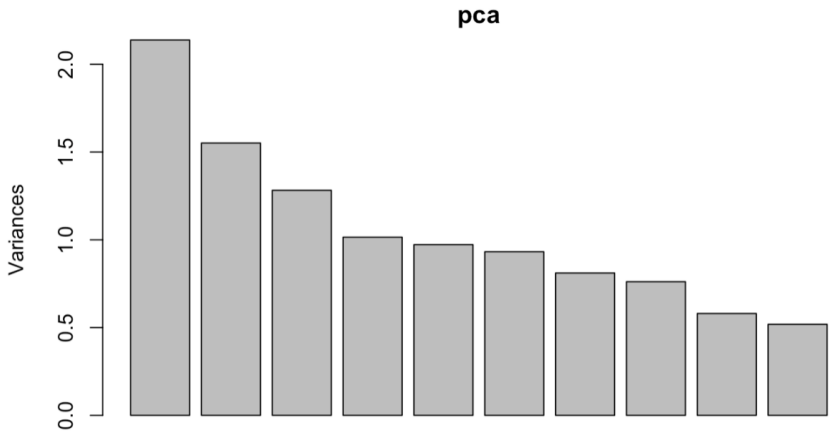
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8857  -0.7284  -0.4233   0.7146   3.0651
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -3.9086480  0.0707695 -55.231  < 2e-16 ***
## no_of_adults                          0.1702330  0.0282467   6.027 1.67e-09 ***
## no_of_children                        0.0061780  0.0359261   0.172 0.863466
## no_of_weekend_nights                  0.1956117  0.0155594  12.572  < 2e-16 ***
## no_of_week_nights                     0.0548741  0.0096129   5.708 1.14e-08 ***
## required_car_parking_space           -1.2517909  0.1110942 -11.268  < 2e-16 ***
## lead_time                             0.0125430  0.0001788  70.165  < 2e-16 ***
## repeated_guest                       -2.3553556  0.3812734  -6.178 6.51e-10 ***
## no_of_previous_cancellations          0.2605623  0.0721902   3.609 0.000307 ***
## no_of_previous_bookings_not_canceled -0.0915952  0.0822655  -1.113 0.265532
## avg_price_per_room                    0.0191016  0.0004878  39.162  < 2e-16 ***
## no_of_special_requests               -1.0854258  0.0218393 -49.701  < 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: 45887  on 36274  degrees of freedom
## Residual deviance: 33837  on 36263  degrees of freedom
## AIC: 33861
##
## Number of Fisher Scoring iterations: 8
```

The results indicate that several variables are statistically significant predictors of cancellations, other than lead time, also including number of adults, number of weekend nights, number of week nights, requirement for car parking space, repeated guest status, number of previous cancellations, average price per room, and number of special requests.

3. Principal Component Analysis (PCA)

To reduce complexity and address multicollinearity, we decide to conduct PCA to do further analysis, which is a dimensionality reduction method that transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components, capturing most of the variation.





The plot showed a clear elbow point after the 1st principal component, where the second, third and fourth principal components have continuous drop-offs afterwards. Our plot above suggests that the first four principal components were the most important as the principal components' variance are close after a certain point.

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.4259 0.8498 0.51283 0.40201
Proportion of Variance 0.6394 0.2271 0.08271 0.05082
Cumulative Proportion  0.6394 0.8665 0.94918 1.00000
                       PC1   PC2   PC3   PC4
no_of_adults          0.05 -0.07 -0.99 -0.05
no_of_children        0.01 -0.01  0.05 -1.00
no_of_weekend_nights 0.17 -0.98  0.08  0.02
no_of_week_nights     0.98  0.18  0.03  0.01
```

The analysis identifies four principal components that capture the majority of the variance in the data. The first principal component (PC1) explains the highest proportion of variance (0.6394), followed by PC2 (0.2271).

PC1 has positive loadings for "no_of_adults" (0.05), "no_of_children" (0.01), and "no_of_week_nights" (0.98), suggesting a positive correlation between these variables and PC1. It also has a negative loading for "no_of_weekend_nights" (0.17), indicating a negative correlation. PC2 has negative loadings for "no_of_weekend_nights" (-0.98) and "no_of_adults" (-0.07), indicating a negative correlation between these variables and PC2.

The PCA results indicate that the variables "no_of_adults", "no_of_children", "no_of_weekend_nights" are the most influential in capturing the variation in the data. PC1 is influenced by variables such as the number of adults, number of children, and number of week nights, while PC2 is mainly influenced by the number of weekend nights.

4. Moderating Effect Exploration

Considering the strong correlation between lead time and cancellations, a further investigation of the moderating effects is conducted. Specifically, the moderating effects of weekends and weekdays, as well as the number of adults and children, on the relationship between lead time and cancellations are explored. The results indicate that weekends and weekdays, along with the number of adults, significantly moderate the impact of lead time on cancellations.

(1) the moderating effect of number of weekends

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.1089431  0.0274939  -76.71   <2e-16 ***
lead_time           0.0139375  0.0002294   60.76   <2e-16 ***
weekends            0.3710120  0.0216449   17.14   <2e-16 ***
lead_time:weekends -0.0026754  0.0001813  -14.76   <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: 45887  on 36274  degrees of freedom
Residual deviance: 38586  on 36271  degrees of freedom
AIC: 38594

Number of Fisher Scoring iterations: 4
```

The estimate for the interaction term "lead_time:weekends" is -0.0026754. It suggests that the effect of "lead_time" on the cancellation differs depending on whether weekends are

included in the stay duration. Specifically, for each unit increase in lead time, the log-odds of the outcome decrease by approximately 0.0026754 when weekends are included in the stay.

The plot provides visual evidence of the moderating effect of weekends on the relationship between lead time and cancellation. It suggests that the impact of lead time on cancellation varies depending on the number of weekends in the stay. When there are fewer weekends, lead time has a stronger positive effect on cancellation, while when there are 7 weekend nights, lead time has a negative effect on cancellation.

This suggests that customers are less likely to cancel their reservations if the stay duration spans across 7 weekend nights. This could be due to the inclusion of weekends making the stay more appealing or providing more flexibility to customers.

(2) the moderating effect of number of weekdays

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.0980024  0.0345658 -60.696   <2e-16 ***
lead_time           0.0138503  0.0003100  44.680   <2e-16 ***
weekdays            0.1378514  0.0130170  10.590   <2e-16 ***
lead_time:weekdays -0.0009587  0.0001110  -8.639   <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: 45887  on 36274  degrees of freedom
Residual deviance: 38771  on 36271  degrees of freedom
AIC: 38779

Number of Fisher Scoring iterations: 4
```

The coefficient for the interaction term between lead time and weekdays is -0.0009587. This suggests that the effect of lead_time on the log-odds of cancellation is moderated by weekdays. Specifically, as the value of weekdays increases, the positive effect of lead_time on cancellation diminishes slightly.

The plot provides visual evidence of the moderating effect of weekdays on the relationship between lead time and cancellation. It suggests that the impact of lead time on cancellation varies depending on the number of weekdays in the stay. When there are fewer

weekends, lead time has a stronger positive effect on cancellation, while when there are 20 weekday nights, lead time has a negative effect on cancellation.

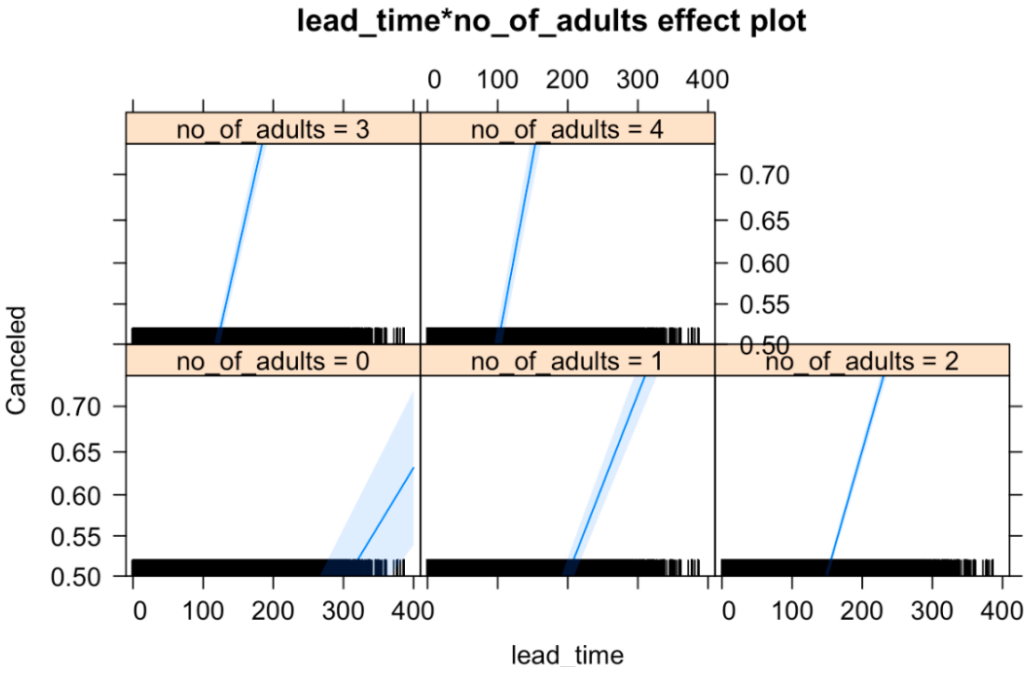
This suggests that customers are less likely to cancel their reservations if the stay duration includes a substantial number of weekday nights. This could be because longer stays with more weekday nights may indicate a purposeful or business-related trip, which could reduce the likelihood of cancellation.

(3) the moderating effect of number of adults

In the principal component analysis, we also find that the number of adults and number of children are influential. Therefore, we generate the hypothesis that the number of adults and number of children may interact with lead time to affect the likelihood of cancellation.

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.7652993  0.0694021 -25.436   <2e-16
lead_time                 0.0057635  0.0006039   9.543   <2e-16
no_of_adults             -0.0285290  0.0370281  -0.770    0.441
lead_time:no_of_adults    0.0032413  0.0003242   9.998   <2e-16

(Intercept)              ***
lead_time                ***
no_of_adults
lead_time:no_of_adults   ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



The interaction between lead time and the number of adults is significant, suggesting that the influence of lead time on cancellation is dependent on the number of adults present. This implies that the relationship between lead time and cancellation may be stronger depending on the number of adults involved in the hotel reservation.

(4) the moderating effect of number of children

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.8467278  0.0202183 -91.339  < 2e-16
lead_time                 0.0117963  0.0001633  72.236  < 2e-16
no_of_children            0.2728157  0.0438444   6.222  4.9e-10
lead_time:no_of_children  0.0007425  0.0004761   1.560    0.119

(Intercept)              ***
lead_time                ***
no_of_children           ***
lead_time:no_of_children
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On the other hand, the interaction between lead time and the number of children is not significant, suggesting that the influence of lead time on cancellation is independent of the number of children present.

5. Model Comparison

To evaluate the effectiveness of the logistic regression model with interaction terms, its predictive accuracy is compared with other machine learning algorithms.

We split data into a training set and test set where the test set is specified to be 30% of the total data. Then we respectively train six models on the training data, predict the target variable for the test data, and then evaluate the model's performance by calculating the accuracy score, generating a confusion matrix, and printing a classification report.

Model	Accuracy Score
Random Forest Classifier	90.08%
Decision Tree Classifier	86.56%
KNN	82.28%
Lasso	81.43%
Logistic Regression with interaction	80.65%
Logistic Regression	79.06%

The comparison of different models reveals the Random Forest Classifier as the most accurate model for predicting cancellations.

While the Random Forest Classifier may have higher predictive accuracy, it is a complex model that may not provide direct interpretability of individual predictor effects. On the other hand, the lasso model's feature selection capability allows for a more focused analysis on the causal effects of specific predictors, even if its overall predictive accuracy is slightly lower (still higher than logistic regression). Therefore, the lasso model is considered the most suitable model for our further causal effect analysis.

C. Conclusion

Cancellation of hotel reservations poses a significant challenge for hotels, impacting their revenue and operations. This analysis aims to identify reservation characteristics that influence hotel cancellations, providing valuable insights for developing strategies to minimize cancellations and optimize revenue.

The analysis reveals that several factors have a significant impact on hotel cancellations. These factors include lead time, the number of adults, the number of weekend nights, the number of week nights, the requirement for car parking space, repeated guest status, the number of previous cancellations, the average price per room, and the number of special requests.

Furthermore, the exploration of moderating effects highlights the influence of weekends, weekdays, and the number of adults on the relationship between lead time and cancellations. This suggests that the impact of lead time on cancellations varies depending on whether the stay includes weekends or weekdays and the number of adults involved in the reservation.

Additionally, the comparison of different predictive models indicates that the Random Forest Classifier performs the best in accurately predicting cancellations. On the other hand, the lasso model is considered most suitable for further causal effect analysis.

These findings offer valuable insights for hotel managers, enabling them to develop effective strategies and policies aimed at reducing cancellations and maximizing revenue generation.

9. Can we identify any causal factors for predicting hotel cancellations?

A. Introduction

Although running a logistic regression of booking status on the dependent variables can provide some information on what variables are correlated to booking status, it does not tell us about what variables may have a causal effect. In this section, we will run a causal double lasso to try to identify any variables that have a causal effect on booking status. We represent categorical variables (market segment type, room type reserved, type of meal plan) as dummy variables.

B. Analysis

1. Causal lasso with hotel cancellation (yes or no) as the dependent variable

Results (significant causal treatment variables highlighted):

Treatment Variable	R^2 (of regressing treatment variable on other predictors)	Coefficient in causal lasso
Number of adults	0.25	0
Number of children	0.5	-0.01
Number of weekend nights	0.06	-0.01
Number of weekday nights	0.09	-0.01
Required car parking space	0.03	0.02
Lead time	0.28	-0.22
Arrival year	0.3	-0.02
Arrival month	0.21	0.01
Arrival date	0.01	0
Repeated guest	0.43	0
Number of previous cancellations	0.26	0
Number of previous bookings	0.38	-0.01

not canceled		
Average price per room	0.51	-0.08
Number of special requests	0.2	0.15
Meal plan 1	0.47	0.01
Meal plan 2	0.53	0
Meal plan 3	0.02	0
Room 2	0.09	0
Room 3	0	0
Room 4	0.27	0
Room 5	0.03	0
Room 6	0.5	0
Room 7	0.1	0
Market Segment Online	0.44	-0.24
Market Segment Corporate	0.35	-0.06
Market Segment Aviation	0.04	0
Market Segment Complementary	0.24	-0.16

We choose 0.1 as a possible threshold for identifying significant variables. From the causal lasso, we identify lead time, number of special requests, and market segment type as significant causal predictors of hotel cancellation. Interpreting the coefficients:

- Lead time is the number of days between the booking date and the arrival date. For every 1 sd increase in lead time (86 days), the odds of keeping the hotel reservation are multiplied by a factor of exp(-0.22) = 0.8. Essentially, the further out the booking date is, the less likely the reservation is to be kept. This makes sense because people are less sure of their plans a year from now than a week from now, and so if they book a year in advance, they are more likely to change their plans.
- For every 1 sd (0.78) increase in the number of special requests, the odds of keeping the hotel reservation are multiplied by a factor of exp(0.15) = 1.16. This makes sense because

a hotel that is able to grant special requests is more likely to be one of the only ones that can meet the demand of the person who booked it; thus, they would be less likely to go to another hotel that cannot meet their special request, and more likely to keep the current one.

- Keeping all else constant, the odds of keeping the hotel reservation for those in the online market segment are exp(-0.24) = 0.79 times the odds of keeping the hotel reservation for those in the offline market segment. Keeping all else constant, the odds of keeping the hotel reservation for those in the complementary market segment are exp(-0.14) = 0.85 times the odds of keeping the hotel reservation for those in the offline market segment. Supposing that online market segment means normal customers who booked online, it makes sense that those who booked online are more likely to cancel than those who booked in person because booking in person requires more of a time investment and opportunity cost. Supposing that the complementary market segment means those who got the hotel for free, it makes sense there are more cancellations because someone who got something for free is less likely to value it than someone who got it for a price.

Based on some basic exploration, the causal variables we just identified make sense. Among reservations with a lead time of more than 100 days, about 60% of them were canceled. Among reservations with a lead time of less than or equal to 100 days, only 20% of them were canceled. Also, among reservations with at least one special request, about 20% of them were canceled. Among reservations with no special requests, about 40% of them were canceled.

From running the causal double lasso, we have found what seems to be a causal effect of several variables on cancellation outcome: lead time and number of special requests. Greater lead times are likely to cause more cancellations. More special requests are likely to cause less cancellations. Holding all else constant, a reservation being made online is more likely to cause a cancellation than a reservation being made offline.

The way the hotel could use this information is that it could find methods to encourage shorter lead times (perhaps offer a discount for shorter lead times), as well as provide niche services that other hotels don't, so that more people with special requests go to that hotel. Lead time and special requests are the two main variables to focus on to improve cancellation rate.

2. Causal lasso on lead time and number of special requests

Next, we can run a causal double lasso on lead time and special requests and see if we can find what variables are causally related to those two. This would be a second-level causal effect (cause of a cause), so this method is not really robust. It is just to try to find what other variables the hotel can look at to try to affect lead time and special requests.

First, we run the causal lasso on lead time as the response, this time only showing variables we deem to be significant.

Treatment Variable	R^2 (of regressing treatment variable on other predictors)	Coefficient in causal lasso
Number of adults	0.24	0.11
Number of week nights	0.07	0.11
Average price per room	0.49	-0.2

Using our threshold of 0.1, it seems that when there are more adults, there is greater lead time (perhaps a larger crowd). Similarly, when there are more weeknights (signaling a longer reservation), there is also greater lead time. Last but not least, when there is a low price, there is greater lead time. Perhaps the hotel can focus on becoming more attractive to small groups that book expensive rooms for a long amount of time. Or, it can try to offer incentives for large groups to book with less lead time to capture more revenue from that group.

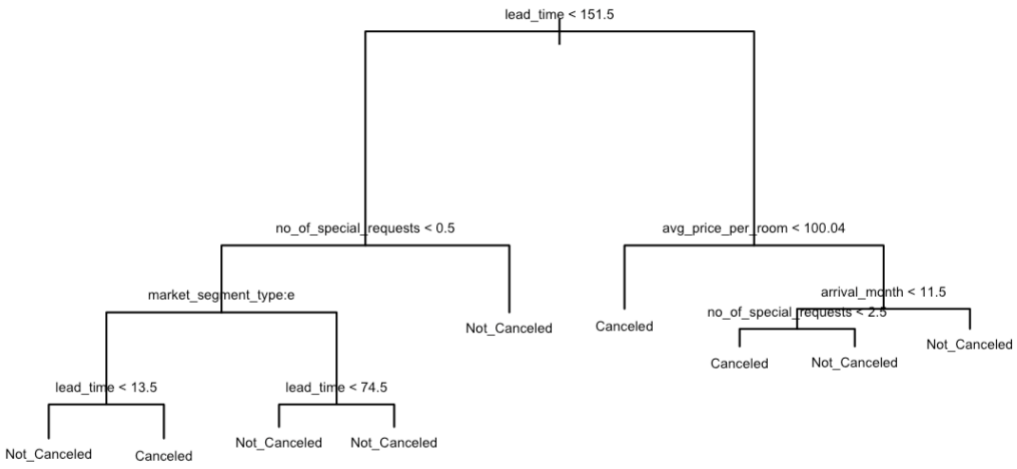
Second, we run the causal lasso on lead time as the response, again only showing variables we deem to be significant.

Treatment Variable	R^2 (of regressing treatment variable on other predictors)	Coefficient in causal lasso
Number of adults	0.24	0.11
Number of children	0.5	0.09

It seems that large groups (more adults and children) result in more special requests, which makes sense since it is more likely for at least one person in a large group to have a special request. This again points to attracting large groups and incentivizing them to book with low lead times.

3. Agreement with Decision Tree model

When we run the decision tree model on the data, we get the following decision tree:



Note that the tree looks at lead time, number of special requests, market segment type online, and average price per room as some of the variables in its classification process. This agrees with our analysis from parts 2 and 3.

C. Conclusion

We ran a causal lasso to try to predict hotel cancellation and found that two variables, lead time and number of special requests, may have a causal effect on whether the reservation is canceled. Thus, the hotel should focus on these two variables to decrease cancellation rate. Then, we ran causal lassos on lead time and number of special requests to see if we could uncover second-level causes. We found that larger groups tend to have larger lead times but also more special requests. This lends itself to one possible strategy of trying to attract larger groups but at the same time incentivizing those groups to book earlier. There are lots of different strategies the hotel can take based on the possible causal factors identified in this section.

10. Conclusion and Improvement

To start off our paper, we cleaned our dataset, re-coded our variable, and performed basic exploratory analysis on our variables. Through this analysis, we were able to locate a number of variables of interest that may place into whether or not a specific customer was likely to cancel on their reservation. Then, we utilized these insights to brainstorm research questions that we explored below:

In section 6, we fit regression models, explored the effect of multiple testing briefly, and majorly, found very insightful results of the effect of each variable on the average price per room using marginal and sectional regression models. We were also able to create a predictive model with a good accuracy level, and also used PCA to show another way of modeling by reducing the dimensionality and multicollinearity effects in our data.

In section 7, we found out the distinct booking preferences across various market segments, including unique characteristics and patterns such as length of stay, cancellation and repeat bookings, and specific preferences. Using clustering analysis, we identified five distinct guest groups, with Online, Corporate, and Offline customers primarily contributing to hotel reservations.

Section 8 highlights the main findings of the logistic regression model on hotel cancellations, identifying significant factors. It also reveals that the impact of lead time on cancellations varies based on weekends/weekdays and the number of adults in the reservation. Moreover, the Random Forest Classifier proves to be the most accurate predictive model for cancellations, while the lasso model is recommended for causal effect analysis.

In section 9, we used the causal lasso statistical technique to find some factors that may be causally related to whether a customer cancels their hotel reservation. We then saw that the causal variables we identified were also the ones that split branches in the decision tree model, giving us confidence in our findings.

In terms of improvement, some methods we could have considered to better analyze the research questions at hand would be to pull in external data relating to the variables we were analyzing. For instance, we could try and find what average hotel room price is across national distributions to understand if the data we are working with is unbiased, and does not systematically differ from real averages for any reason. This would allow us to be able to more confidently present our findings to hotel shareholders. In addition, our data seems to come from Europe, so we could get similar datasets from other regions (e.g., U.S., Asia-Pacific) and verify that the same findings hold for these regions. What's more, additional deep learning models might also be useful to support our findings.

11. Appendix

1. Exploratory Data Analysis

```
df = df.rename(columns={"no_of_adults": "# Adults", "no_of_children": "# Children",
"no_of_week_nights": "# Week Nights", "no_of_weekend_nights": "# Weekend Nights",
"type_of_meal_plan":"Meal Plan", "required_car_parking_space": "Car Space",
"no_of_previous_cancellations": "# Prev Cancels", "no_of_previous_bookings_not_canceled":
"# Prev Non Cancel", "room_type_reserved":"Room Type", "lead_time": "Lead Time",
"arrival_year": "Arrival Year", "arrival_month": "Arrival Month", "arrival_date":"Arrival Data",
"market_segment_type":"Market Segment Type", "repeated_guest": "Repeat Guest"})
df = df.rename(columns={"avg_price_per_room": "Avg Room Price", "no_of_special_requests":
"# Special Requests", "booking_status":"Booking Status"})
df = df.replace({'Room Type': {'Room_Type 1': 1, 'Room_Type 2': 2, 'Room_Type 3': 3,
'Room_Type 4': 4, 'Room_Type 5': 5, 'Room_Type 6': 6, 'Room_Type 7': 7}})
df = df.replace({'Booking Status': {'Canceled': 1, 'Not_Canceled': 0}})
df = df.replace({'Market Segment Type': {'Offline': 0, 'Online': 1, 'Corporate': 2, 'Aviation': 3,
'Complementary': 4}})
df = df.replace({'Meal Plan': {'Meal Plan 1': 1, 'Meal Plan 2': 2, 'Meal Plan 3': 3, 'Not Selected':
0}})

n, bin1, patches = plt.hist(x = df['Booking Status'])
plt.xlabel('Not Cancelled and Cancelled')
plt.ylabel('Frequency')
plt.title('Histogram of Cancellations')
labels = ['Not Cancelled', 'Cancelled']
plt.xticks(np.arange(0.05, .95, .89), labels)
patches[0].set_fc('orange')

ax = df.groupby('Booking Status').size().plot(kind='pie', autopct='%.2f', colors = ['orange',
'royalblue'], labels = ['Not Cancelled', 'Cancelled'])

df.hist(figsize = (12,12))

ax = sns.boxplot(x = 'Avg Room Price', y = 'Booking Status', data = df)

plt.hist(df['Avg Room Price'], bins='auto')
plt.title('Histogram of Average Price Per Room')
plt.xlabel('Price Per Room')
```

```
plt.ylabel('Count')
plt.show()

plt.hist(df['# Adults'], bins='auto')
plt.title('Histogram of # of Adults')
plt.xlabel('# of Adults')
plt.ylabel('Count')
plt.xticks(np.arange(0, 5, 1.00))
plt.show()

plt.close()
plt.hist(df['# Children'], color = 'red')
plt.title('Histogram of # of Children')
plt.xlabel('# of Children')
plt.ylabel('Count')
plt.show()

df['# Children'].value_counts()

plt.hist(df['# Week Nights'])
plt.title('Histogram of # of Week Nights')
plt.xlabel('# of Week Nights')
plt.ylabel('Count')
plt.show()

plt.close()
plt.hist(df['# Weekend Nights'], color = 'red')
plt.title('Histogram of # of Weekend Nights')
plt.xlabel('# of Weekend Nights')
plt.ylabel('Count')
plt.show()

plt.close()
plt.hist(df['# Special Requests'], color = 'green')
plt.title('Histogram of # of Special Requests')
plt.xlabel('# of Special Requests')
plt.ylabel('Count')
plt.show()

plt.hist(df['Lead Time'], bins='auto')
```

```
plt.title('Histogram of Lead Time')
plt.xlabel('Lead Time (Number of Days)')
plt.ylabel('Count')
plt.show()

fig = plt.figure(figsize=(10,6))
sns.kdeplot(df[df['Booking Status'] == 1]['Lead Time'], color="green", shade=True)
sns.kdeplot(df[df['Booking Status'] == 0]['Lead Time'], color="blue", shade=True)
fig.legend(labels=['Cancelled','Not Cancelled'])
plt.title('Variation of the Lead Time (by Booking Status)')
plt.show()

import matplotlib.pyplot as plt
df.corr().style.background_gradient(cmap='coolwarm').set_precision(3)

plt.scatter(x = 'Arrival Month', y = 'Avg Room Price', data = df)

plt.close()
fig = plt.figure(figsize = ( 5 , 3 ))
sns.lmplot(x = 'Arrival Month', y = 'Avg Room Price', data = df, fit_reg=True)
plt.title('2017-2018: Regression of Average Room Price vs. Arrival Month')
plt.xlabel('Month')
plt.ylabel('Average Price')

plt.figure()
df.value_counts(["Meal Plan", "Booking Status"]).sort_index().plot(kind = 'bar')
plt.xlabel('')
plt.ylabel('Frequency')
plt.title('Histogram of Meal Plan and Cancellations')
labels = ['No Meal / Not Cancelled', 'No Meal / Cancelled', 'Meal 1 / Not Cancelled', 'Meal 2 /
Cancelled',
        'Meal 3 / Not Cancelled', 'Meal 3 / Cancelled', 'Meal 4 / Not Cancelled']
plt.xticks(np.arange(0, 7, 1.00), labels)

plt.figure()
df.value_counts(["Room Type", "Booking Status"]).sort_index().plot(kind = 'bar')
plt.xlabel('')
plt.ylabel('Frequency')
plt.title('Histogram of Room Type and Cancellations')
```

```
labels = ['Room Type 1 / Not Cancelled', 'Room Type 1 / Cancelled', 'Room Type 2 / Not
Cancelled', 'Room Type 2 / Cancelled', 'Room Type 3 / Not Cancelled', 'Room Type 3 /
Cancelled',
        'Room Type 4 / Not Cancelled', 'Room Type 4 / Cancelled', 'Room Type 5 / Not Cancelled',
'Room Type 5 / Cancelled', 'Room Type 6 / Not Cancelled', 'Room Type 6 / Cancelled',
        'Room Type 7 / Not Cancelled', 'Room Type 7 / Cancelled']
plt.xticks(np.arange(0, 14, 1.00), labels)

plt.figure()
df.value_counts(["Market Segment Type", "Booking Status"]).sort_index().plot(kind = 'bar')
plt.xlabel('')
plt.ylabel('Frequency')
plt.title('Histogram of Market Segment Type and Cancellations')
labels = ['Offline / Not Cancelled', 'Offline / Cancelled', 'Online / Not Cancelled', 'Online /
Cancelled', 'Corporate / Not Cancelled', 'Corporate / Cancelled',
         'Aviation / Not Cancelled', 'Aviation / Cancelled', 'Complementary / Not Cancelled',
'Complementary / Cancelled']
plt.xticks(np.arange(0, 10, 1.00), labels)
```

2. What are the factors that affect the average price per room in the hotel?

```{r echo=TRUE, include=FALSE}
dat<-read.csv("Hotel Reservations.csv")
head(dat)
```

```{r echo=TRUE, include=FALSE}
#We need to convert some variables with integer values to categorical variables
pr=dat$avg_price_per_room
dat$arrival_year=as.factor(dat$arrival_year)
dat$arrival_month=as.factor(dat$arrival_month)
dat$arrival_date=as.factor(dat$arrival_date)
dat$repeated_guest=as.factor(dat$repeated_guest)

lm1<-lm(pr~.-(Booking_ID+booking_status),data=dat)
summary(lm1)
Estimate<-summary(lm1)$coef[, "Estimate"]
P_value<-summary(lm1)$coef[, "Pr(>|t|)"]
data.frame(Estimate,P_value)
```

```{r echo=TRUE, include=FALSE}
p_values <- summary(lm1)$coef[, "Pr(>|t|)"]

adjusted_p_values <- p.adjust(p_values, method = "fdr")

results <- data.frame(
  p_value = p_values,
  adjusted_p_value = adjusted_p_values
)
 p_bon=0.05/67
 p_bon
print(results)
cutoff <- max(adjusted_p_values[adjusted_p_values <= 0.05])
cutoff
```
```{r echo=TRUE, include=FALSE}
length(lm1$coefficients)
sum(summary(lm1)$coefficients[, "Pr(>|t|)"] < 0.05)
```

```
significant_coeffs <- names(adjusted_p_values[adjusted_p_values <= cutoff])
sum(summary(lm1)$coefficients[, "Pr(>|t|)"] < p_bon)
length(significant_coeffs)
```

```{r echo=TRUE, include=FALSE}
#Include

lm3<-glm(pr~no_of_adults+no_of_children,data=dat)
summary(lm3)
visreg(lm3, scale = "response", main = "Regression Plot")
```

```{r echo=TRUE, include=FALSE}
lm4<-glm(pr~no_of_weekend_nights+no_of_week_nights,data=dat)
summary(lm4)
visreg(lm4, scale = "response", main = "Partial Regression Plot")
```

```{r echo=TRUE, include=FALSE}
lm5<-glm(pr~type_of_meal_plan,data=dat)
summary(lm5)
boxplot(pr ~ type_of_meal_plan, data = dat,
        xlab = "Meal Plan Type", ylab = "pr",
        main = "Distribution of Average price per room by Meal plan type")

meal_plan_freq <- table(dat$type_of_meal_plan)
table_with_margins <- addmargins(meal_plan_freq)
meal_plan_df <- as.data.frame(table_with_margins)
colnames(meal_plan_df) <- c("Type of Meal Plan", "Count")
print(meal_plan_df)
```

```{r echo=TRUE, include=FALSE}
lm6<-glm(pr~required_car_parking_space,data=dat)
summary(lm6)
```

```{r echo=TRUE, include=FALSE}
lm7<-glm(pr~room_type_reserved,data=dat)
summary(lm7)
boxplot(pr ~ room_type_reserved, data = dat,
        xlab = "Room Type", ylab = "pr",
        main = "Distribution of Average price per room by the Type of room reserved")

```

```{r echo=TRUE, include=FALSE}
lm8<-glm(pr~lead_time, data=dat)
summary(lm8)
visreg(lm8, scale = "response", main = "Partial Regression Plot")
```

```{r echo=TRUE, include=FALSE}
lm9<-glm(pr~arrival_year+arrival_month,data =dat)
summary(lm9)
```

```{r echo=TRUE, include=FALSE}
lm10<-glm(pr~market_segment_type,data=dat)
summary(lm10)
# Create a box plot
boxplot(pr ~ market_segment_type, data = dat,
        xlab = "Market Segment Type", ylab = "pr",
        main = "Distribution of Average price per room by Market Segment Type")

```

```{r echo=TRUE, include=FALSE}
lm11<-glm(pr~repeated_guest+no_of_previous_cancellations+no_of_previous_bookings_not_ca
nceled, data=dat)
summary(lm11)
```

```

```{r echo=TRUE, include=FALSE}
lm12<-glm(pr~no_of_special_requests,data=dat)
summary(lm12)
```

```{r warning=FALSE,echo=TRUE, include=FALSE}

set.seed(123)

d <- subset(dat, select = -c(Booking_ID,booking_status))

index <- sample(nrow(d), nrow(d) * 0.7) #choosing random numbers which are 70% of the total datapoints

trainData <- d[index, ]
testData <- d[-index, ]

model <- lm(avg_price_per_room ~ . , data = trainData, family = "gaussian")
summary(model)
Estimate<-summary(model)$coef[, "Estimate"]
P_value<-summary(model)$coef[, "Pr(>|t|)"]
data.frame(Estimate,P_value)


```

```{r echo=TRUE, include=FALSE}
predictions <- predict(model, newdata = testData)

rsquared <- cor(predictions, testData$avg_price_per_room)^2
print(rsquared)
```
```{r echo = FALSE}
require(tidymodels)
pca_dat = dat %>% mutate(
  pr = avg_price_per_room,

```
  across(c(arrival_month, arrival_year, type_of_meal_plan, room_type_reserved,
market_segment_type), as.character)
) %>% pivot_longer(
  cols = c(arrival_month, arrival_year, type_of_meal_plan, room_type_reserved,
market_segment_type),
  values_transform = as.character
) %>% mutate(
  val = paste0(name, "_", value),
  val_fil = 1
) %>% select(
  -one_of(c("name", "value"))
) %>% pivot_wider(
  names_from  = val,
  values_from = val_fil,
  values_fill = 0
) %>% select(
  -one_of(c("avg_price_per_room", "arrival_date", "booking_status"))
)
```

```{r}
pca_recipe = pca_dat %>% recipe(pr~.) %>%
  update_role(Booking_ID, new_role = "id") %>%
  step_normalize(all_predictors()) %>%
  step_pca(all_predictors())
pca_prep = prep(pca_recipe)
```

```{r}
components = tidy(pca_prep, 2) # PCA components

scores = juice(pca_prep) # The scores of each sample with respect to these components

variances = tidy(pca_prep, 2, type = "variance")
```

```{r}
# how much variance is explained by each dimension of the PCA
```

```
# the more rapid the dropoff in variance explained, the better the low-dimensional approximation
b = ggplot(variances[81:120,]) +
  geom_col(aes(component, value))

```

```{r echo = FALSE}
require(tidytext)
components_ = components %>%
  filter(component %in% paste0("PC", 1:4)) %>%
  mutate(terms = reorder_within(terms, abs(value), component))
```

```{r fig.width= 15, fig.height=15}
a = ggplot(components_, aes(value, terms)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ component, scales = "free_y") +
  scale_y_reordered() +
  labs(y = NULL) +
  theme(axis.text = element_text(size = 7))

#ggsave(plot = a, file = "Components.jpg",
#       width = 24, height = 25,
#       units = "cm", dpi = 300)
#ggsave(plot = b, file = "VarianceExplainedPerc.jpg",
#       width = 15, height = 10,
#       units = "cm", dpi = 300)
```

```{r}
summary(yglm <- glm(pr ~ ., data=scores[,2:6]))
```

3. How do booking preferences differ among distinct market segments?

```
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from matplotlib import pyplot as plt
import matplotlib.colors as colors
import seaborn as sns
import pandas as pd
import numpy as np

data = pd.read_csv("/Users/stancy/Desktop/Spring 2023/Big Data/Final Project/Hotel
Reservations.csv")
data.head()

counts = data['market_segment_type'].value_counts()
counts.plot.pie(figsize=(5, 5), autopct='%1.1f%%')

# Add no_of_nights
data['no_of_nights'] = data['no_of_weekend_nights'] + data['no_of_week_nights']

# Get a list of the market segments
segments = data['market_segment_type'].unique()

plt.figure(figsize=(10, 6))
sns.boxplot(x='market_segment_type', y='no_of_nights', data=data)
plt.xlabel('Market Segment Type')
plt.ylabel('Number of Nights')
plt.title('Length of Stay by Market Segment')
plt.show()

summary_df = pd.DataFrame()
for segment in segments:
  segment_data = data[data['market_segment_type'] == segment]['no_of_nights']
  summary = segment_data.describe()
  summary['IQR'] = summary['75%'] - summary['25%']
  summary_df[segment] = summary
```

```
summary_df = summary_df.transpose()

summary_df.style.format({
  'count': '{:,.0f}'.format,
  'mean': '{:,.2f}'.format,
  'std': '{:,.2f}'.format,
  'min': '{:,.0f}'.format,
  '25%': '{:,.0f}'.format,
  '50%': '{:,.0f}'.format,
  '75%': '{:,.0f}'.format,
  'max': '{:,.0f}'.format,
  'IQR': '{:,.2f}'.format
}).background_gradient(cmap='Blues')

cancellation_rates = data.groupby('market_segment_type')['booking_status'].apply(lambda x: (x
== 'Canceled').mean())
repeat_booking_rates = data.groupby('market_segment_type')['repeated_guest'].mean()

rates_df = pd.concat([cancellation_rates, repeat_booking_rates], axis=1)
rates_df.columns = ['Cancellation Rate', 'Repeat Booking Rate']
rates_df = rates_df.transpose()

rates_df.style.format({
  'Cancellation Rate': '{:,.3f}'.format,
  'Repeat Booking Rate': '{:,.3f}'.format,
}).background_gradient(cmap='Blues')

rates_df = pd.DataFrame({
  'cancellation_rate': cancellation_rates,
  'repeat_booking_rate': repeat_booking_rates
})

correlation = rates_df['cancellation_rate'].corr(rates_df['repeat_booking_rate'])
sns.regplot(x='cancellation_rate', y='repeat_booking_rate', data=rates_df)
for segment in segments:
  sns.regplot(x='cancellation_rate',
              y='repeat_booking_rate',
              data=rates_df[rates_df.index == segment],
              label=segment)
```

```
plt.title('Average Cancellation Rate vs Average Repeat Booking Rate\n' + 'Correlation: ' +
str(correlation))
plt.xlabel('Average Cancellation Rate')
plt.ylabel('Average Repeat Booking Rate')
plt.legend()
plt.show()

preferences_market = data.groupby('market_segment_type').agg({
  'room_type_reserved': pd.Series.mode,
  'type_of_meal_plan': pd.Series.mode,
  'no_of_special_requests': 'mean',
  'required_car_parking_space': 'mean',
  'no_of_adults': 'mean',
  'no_of_children': 'mean',
  'no_of_weekend_nights': 'mean',
  'no_of_week_nights': 'mean',
  'lead_time': 'mean',
  'arrival_year': 'mean',
  'arrival_month': 'mean',
  'repeated_guest': 'mean',
  'no_of_previous_cancellations': 'mean',
  'no_of_previous_bookings_not_canceled': 'mean',
  'avg_price_per_room': 'mean'
})

cmap = plt.get_cmap('Blues')

room_type_color = colors.rgb2hex(cmap(0.2))
meal_plan_color = colors.rgb2hex(cmap(0.4))

room_type_colors = {'Room_Type 1': room_type_color, 'Room_Type 4': meal_plan_color}
meal_plan_colors = {'Meal Plan 1': meal_plan_color}

def highlight_cells(x):
  if x.name in ['room_type_reserved', 'type_of_meal_plan']:
    if x.name == 'room_type_reserved':
      return ['background-color: {}'.format(room_type_colors.get(value, '')) for value in x]
    elif x.name == 'type_of_meal_plan':
      return ['background-color: {}'.format(meal_plan_colors.get(value, '')) for value in x]
  else:
```

```
    return [''] * len(x)

preferences_market.style.apply(highlight_cells).format({
  'no_of_special_requests': '{:,.3f}',
  'required_car_parking_space': '{:,.3f}',
  'no_of_adults': '{:,.3f}',
  'no_of_children': '{:,.3f}',
  'no_of_weekend_nights': '{:,.3f}',
  'no_of_week_nights': '{:,.3f}',
  'lead_time': '{:,.3f}',
  'arrival_year': '{:,.3f}',
  'arrival_month': '{:,.3f}',
  'repeated_guest': '{:,.3f}',
  'no_of_previous_cancellations': '{:,.3f}',
  'no_of_previous_bookings_not_canceled': '{:,.3f}',
  'avg_price_per_room': '{:,.3f}',
}).background_gradient(subset=['no_of_special_requests', 'required_car_parking_space',
'no_of_adults','no_of_children','no_of_weekend_nights','no_of_week_nights','lead_time','arrival_
year','arrival_month','repeated_guest','no_of_previous_cancellations','no_of_previous_bookings_
not_canceled','avg_price_per_room'], cmap='Blues')

le = LabelEncoder()
data['type_of_meal_plan'] = le.fit_transform(data['type_of_meal_plan'])
data['room_type_reserved'] = le.fit_transform(data['room_type_reserved'])
variables = ['no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights',
        'type_of_meal_plan', 'required_car_parking_space', 'room_type_reserved', 'lead_time',
        'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations',
        'no_of_previous_bookings_not_canceled', 'avg_price_per_room',
'no_of_special_requests']

scaler = StandardScaler()
scaled_data = scaler.fit_transform(data[variables])
kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(scaled_data)
data['cluster'] = clusters

pca = PCA(n_components=2)
pca_data = pca.fit_transform(scaled_data)
pca_df = pd.DataFrame(pca_data, columns=['PC1', 'PC2'])
pca_df['cluster'] = clusters
```

```
plt.figure(figsize=(10, 7))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='cluster', palette='viridis', s=60)
plt.title('KMeans Clusters (PCA-Reduced Data)')
plt.show()

centroids = scaler.inverse_transform(kmeans.cluster_centers_)
sns.set_theme()
centroids_df = pd.DataFrame(centroids, columns=variables)
formatted_centroids = centroids_df.style.format({
  'no_of_adults': '{:,.3f}',
  'no_of_children': '{:,.3f}',
  'no_of_weekend_nights': '{:,.3f}',
  'no_of_week_nights': '{:,.3f}',
  'type_of_meal_plan': '{:,.3f}',
  'required_car_parking_space': '{:,.3f}',
  'room_type_reserved': '{:,.3f}',
  'lead_time': '{:,.3f}',
  'arrival_year': '{:,.3f}',
  'arrival_month': '{:,.3f}',
  'repeated_guest': '{:,.3f}',
  'no_of_previous_cancellations': '{:,.3f}',
  'no_of_previous_bookings_not_canceled': '{:,.3f}',
  'avg_price_per_room': '{:,.3f}',
  'no_of_special_requests': '{:,.3f}'
}).background_gradient(subset=[
  'no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights',
  'type_of_meal_plan', 'required_car_parking_space', 'room_type_reserved', 'lead_time',
  'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations',
  'no_of_previous_bookings_not_canceled', 'avg_price_per_room', 'no_of_special_requests'],
  cmap='Blues'
)

display(formatted_centroids)

sns.heatmap(cross_tab, annot=True, cmap='Blues')
plt.title('Cross-tabulation of KMeans clusters and Market Segments')
plt.xlabel('Cluster')
plt.ylabel('Market Segment Type')
plt.show()
```

4. What reservation characteristics lead to hotel cancellation?

```{r}
hotel = read.csv('/Users/tongtianyi/Desktop/spring2023/bigdata/finalproject/Hotel_Reservations.csv')
library(tidyverse)
library(corrplot)
library(ggplot2)
library(scales)
library(dplyr)
library(reshape2)
library(effects)
```

```{r}
ggplot(hotel, aes(x = booking_status)) +
  geom_bar() +
  xlab('Booking Status') +
  ylab('Count') +
  ggtitle('Booking Status Distribution')
```

Correlation between cancellation and numerical features

```{r}
hotel$Canceled <- as.numeric(recode(hotel$booking_status, "Not_Canceled" = 0, "Canceled" = 1))

selected_numerical_features <- c('Canceled', 'no_of_adults', 'no_of_children', 'no_of_weekend_nights',
                                 'no_of_week_nights', 'required_car_parking_space', 'lead_time',
                                 'repeated_guest', 'no_of_previous_cancellations',
                                 'no_of_previous_bookings_not_canceled', 'avg_price_per_room',
                                 'no_of_special_requests')

correlation_matrix <- cor(hotel[selected_numerical_features])
melted_cormat <- melt(correlation_matrix)

# plot heatmap of correlation matrix
ggplot(data = reshape2::melt(melted_cormat)) +
  geom_tile(aes(x = Var1, y = Var2, fill = value)) +
```

```
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "", y = "", fill = "Correlation")
```

Distribution of Lead Time by Cancellation Status

```{r}
ggplot(hotel, aes(x = lead_time, fill = Canceled, group = Canceled)) +
  geom_density(alpha = 0.8) +
  labs(x = "Lead Time", y = "Density", fill = "Canceled") +
  ggtitle("Variation of Lead Time by Booking Status")
```

logistic regression

```{r}
hotel_predictors <- hotel %>%
  select(-booking_status, -Booking_ID, -arrival_year, -arrival_month, -arrival_date)
selected_numerical_data <- hotel[selected_numerical_features]
logit_model <- glm(Canceled ~ ., data = selected_numerical_data, family = binomial(link = "logit"))
result <- summary(logit_model)
print(result)
```

principal component analysis

```{r}
# Fit principal components
selected_numerical_features <- c( 'no_of_adults', 'no_of_children', 'no_of_weekend_nights',
                                  'no_of_week_nights', 'required_car_parking_space', 'lead_time',
                                  'repeated_guest', 'no_of_previous_cancellations',
                                  'no_of_previous_bookings_not_canceled', 'avg_price_per_room',
                                  'no_of_special_requests')
selected_numerical_data <- hotel[selected_numerical_features]
pca <- prcomp(selected_numerical_data, scale = TRUE)
zpca <- predict(pca)
# Plot principal components
plot(pca, type = "l")
screeplot(pca)
```

```{r}
```

```
# Perform principal component analysis on the dataset
pca <- prcomp(selected_numerical_data[, 1:4])
# Plot the principal components
plot(pca, type = "l")
# Interpret the principal components
summary(pca)
# Extract the rotation matrix
rotation <- pca$rotation
# Interpret the rotation matrix
round(rotation, 2)
```

Moderating effect exploration

```{r}
# Fit a logistic regression model with an interaction term
hotel <- hotel %>% rename(weekends = no_of_weekend_nights)

moderate_weekends <- glm(Canceled ~ lead_time * weekends, data = hotel, family = "binomial")
# Check the interaction effect
summary(moderate_weekends)
# Create an effects object
effect <- effect("lead_time * weekends", moderate_weekends)
# Plot the interaction effect
plot(effect, x.var = "lead_time",
    xlevels = c(10, 20, 30),
    y.var = "Canceled",
    ylim = c(0, 1),
    lines = TRUE,
    width = 12)
```

```{r}
# Fit a logistic regression model with an interaction term
hotel <- hotel %>% rename(weekdays = no_of_week_nights)

moderate_weekdays <- glm(Canceled ~ lead_time * weekdays, data = hotel, family = "binomial")
# Check the interaction effect
summary(moderate_weekdays)
# Create an effects object
effect <- effect("lead_time * weekdays", moderate_weekdays)
```

```
# Plot the interaction effect
plot(effect, x.var = "lead_time",
    xlevels = c(10, 20, 30),
    y.var = "Canceled",
    ylim = c(0, 1),
    lines = TRUE,
    width = 12)
```
```{r}
moderate_children <- glm(Canceled ~ lead_time * no_of_children, data = hotel, family = "binomial")
# Check the interaction effect
summary(moderate_children)
# Create an effects object
effect <- effect("lead_time * no_of_children", moderate_children)
# Plot the interaction effect
plot(effect, x.var = "lead_time",
    xlevels = c(10, 20, 30),
    y.var = "Canceled",
    ylim = c(0, 1),
    lines = TRUE,
    width = 12)
```
```{r}
moderate_adults <- glm(Canceled ~ lead_time * no_of_adults, data = hotel, family = "binomial")
# Check the interaction effect
summary(moderate_adults)
# Create an effects object
effect <- effect("lead_time * no_of_adults", moderate_adults)
# Plot the interaction effect
plot(effect, x.var = "lead_time",
    xlevels = c(10, 20, 30),
    y.var = "Canceled",
    ylim = c(0, 1),
    lines = TRUE,
    width = 12)
```

```
# importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier

df['booking_status'] = df['booking_status'].map({'Not_Canceled' : 0, 'Canceled': 1})
correlation = df.corr()['booking_status'].abs().sort_values(ascending = False)
correlation
df = df.drop(columns = 'Booking_ID')

cat_cols = [col for col in df.columns if df[col].dtype == 'O']
cat_cols
for col in cat_df.columns:
    cat_df = pd.get_dummies(cat_df, columns=[col], drop_first=True)

num_df = df.drop(columns = cat_cols, axis = 1)
num_df.drop('booking_status', axis = 1, inplace = True)
num_df
num_df['lead_time'] = np.log(num_df['lead_time'] + 1)
num_df['no_of_adults'] = np.log(num_df['no_of_adults'] + 1)
num_df['no_of_children'] = np.log(num_df['no_of_children'] + 1)
num_df['no_of_weekend_nights'] = np.log(num_df['no_of_weekend_nights'] + 1)
num_df['avg_price_per_room'] = np.log(num_df['avg_price_per_room'] + 1)
num_df['no_of_special_requests'] = np.log(num_df['no_of_special_requests'] + 1)
num_df['required_car_parking_space'] = np.log(num_df['required_car_parking_space'] + 1)
num_df['repeated_guest'] = np.log(num_df['repeated_guest'] + 1)
X = pd.concat([cat_df, num_df], axis = 1)
y = df['booking_status']
X.shape, y.shape
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
```

```
# logistic regression
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
acc_lr = accuracy_score(y_test, y_pred_lr)
conf = confusion_matrix(y_test, y_pred_lr)
clf_report = classification_report(y_test, y_pred_lr)
print(f"Accuracy Score of Logistic Regression is : {acc_lr}")
print(f"Confusion Matrix : \n{conf}")
print(f"Classification Report : \n{clf_report}")
df['interaction1'] = df['lead_time'] * df['no_of_weekend_nights']
df['interaction2'] = df['lead_time'] * df['no_of_week_nights']
df['interaction3'] = df['lead_time'] * df['no_of_adults']
interaction = df[['interaction1', 'interaction2','interaction3']]
Interaction
X2 = pd.concat([cat_df, num_df,interaction], axis = 1)
y = df['booking_status']
X2_train, X2_test, y_train, y_test = train_test_split(X2, y, test_size = 0.30)

# Logistic regression with interactions
moderate_lr = LogisticRegression()
lr.fit(X2_train, y_train)
y_pred_lr = lr.predict(X2_test)
acc_lr = accuracy_score(y_test, y_pred_lr)
conf = confusion_matrix(y_test, y_pred_lr)
clf_report = classification_report(y_test, y_pred_lr)
print(f"Accuracy Score of Logistic Regression with interactions is : {acc_lr}")
print(f"Confusion Matrix : \n{conf}")
print(f"Classification Report : \n{clf_report}")

# Lasso (with interactions)
lasso_lr = LogisticRegression(penalty='l1', solver='liblinear')
lasso_lr.fit(X2_train, y_train)
y_pred_lasso_lr = lasso_lr.predict(X2_test)
acc_lasso_lr = accuracy_score(y_test, y_pred_lasso_lr)
conf_lasso = confusion_matrix(y_test, y_pred_lasso_lr)
clf_report_lasso = classification_report(y_test, y_pred_lasso_lr)
print(f"Accuracy Score of Lasso Logistic Regression with interactions is : {acc_lasso_lr}")
print(f"Confusion Matrix : \n{conf_lasso}")
print(f"Classification Report : \n{clf_report_lasso}")
```

```
# KNN
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
acc_knn = accuracy_score(y_test, y_pred_knn)
conf = confusion_matrix(y_test, y_pred_knn)
clf_report = classification_report(y_test, y_pred_knn)
print(f"Accuracy Score of KNN is : {acc_knn}")
print(f"Confusion Matrix : \n{conf}")
print(f"Classification Report : \n{clf_report}")

# Decision Tree Classifier
dtc = DecisionTreeClassifier()
dtc.fit(X_train, y_train)
y_pred_dtc = dtc.predict(X_test)
acc_dtc = accuracy_score(y_test, y_pred_dtc)
conf = confusion_matrix(y_test, y_pred_dtc)
clf_report = classification_report(y_test, y_pred_dtc)
print(f"Accuracy Score of Decision Tree is : {acc_dtc}")
print(f"Confusion Matrix : \n{conf}")
print(f"Classification Report : \n{clf_report}")

# Random Tree Classifier
rd_clf = RandomForestClassifier()
rd_clf.fit(X_train, y_train)
y_pred_rd_clf = rd_clf.predict(X_test)
acc_rd_clf = accuracy_score(y_test, y_pred_rd_clf)
conf = confusion_matrix(y_test, y_pred_rd_clf)
clf_report = classification_report(y_test, y_pred_rd_clf)
print(f"Accuracy Score of Random Forest is : {acc_rd_clf}")
print(f"Confusion Matrix : \n{conf}")
print(f"Classification Report : \n{clf_report}")
```

5. Can we identify any causal factors for predicting hotel cancellations?

---
title: 'Causal Factors'
output: word_document

```{r setup, include=FALSE}
#knitr::opts_chunk$set(echo = TRUE)
```

```{r}
library(gamlr)

hotel = read.csv('Hotel Reservations.csv')
booking_status = as.factor(hotel[,length(hotel)])
hotel_predictors = hotel[,-c(1, length(hotel))]

#scale all numberical variables
hotel_predictors[,-c(5,7,12)] = scale(hotel_predictors[,-c(5,7,12)])

#hotel_predictors['nomealplan'] = hotel_predictors['type_of_meal_plan'] == 'Not Selected'
hotel_predictors['mealplan1'] = hotel_predictors['type_of_meal_plan'] == 'Meal Plan 1'
hotel_predictors['mealplan2'] = hotel_predictors['type_of_meal_plan'] == 'Meal Plan 2'
hotel_predictors['mealplan3'] = hotel_predictors['type_of_meal_plan'] == 'Meal Plan 3'

#hotel_predictors['room1'] = hotel_predictors['room_type_reserved'] == 'Room_Type 1'
hotel_predictors['room2'] = hotel_predictors['room_type_reserved'] == 'Room_Type 2'
hotel_predictors['room3'] = hotel_predictors['room_type_reserved'] == 'Room_Type 3'
hotel_predictors['room4'] = hotel_predictors['room_type_reserved'] == 'Room_Type 4'
hotel_predictors['room5'] = hotel_predictors['room_type_reserved'] == 'Room_Type 5'
hotel_predictors['room6'] = hotel_predictors['room_type_reserved'] == 'Room_Type 6'
hotel_predictors['room7'] = hotel_predictors['room_type_reserved'] == 'Room_Type 7'

#hotel_predictors['marketsegoffline'] = hotel_predictors['market_segment_type'] == 'Offline'
hotel_predictors['marketsegonline'] = hotel_predictors['market_segment_type'] == 'Online'
hotel_predictors['marketsegcorporate'] = hotel_predictors['market_segment_type'] == 'Corporate'
hotel_predictors['marketsegaviation'] = hotel_predictors['market_segment_type'] == 'Aviation'
hotel_predictors['marketsegcomplementary'] = hotel_predictors['market_segment_type'] ==
'Complementary'
```

```
r2_matrix = rep(0, 27)
coef_matrix = rep(0, 27)
j = 0

#do not scale categorical variables, since they already have a good interpretation (eg: is in meal plan 2)
for (i in (1:30)[-c(5,7,12)]) {
  lasso1 <- gamlr(hotel_predictors[,-c(i,5,7,12)], y=hotel_predictors[,i], standardize = FALSE, lambda.min.ratio=1e-3)
  print(summary(lasso1)[which.min(summary(lasso1)$aicc),]$r2)
  print('r2')
  d_hat = predict(lasso1, hotel_predictors[,-c(i,5,7,12)], type="response")
  lasso2 = gamlr(cbind(hotel_predictors[,i],d_hat,as.matrix(hotel_predictors[,-c(i,5,7,12)])), booking_status == 'Not_Canceled', free=2, standardize = FALSE, type = 'binomial')
  print(coef(lasso2)[2])
  print(colnames(hotel_predictors)[i])

  j = j + 1
  r2_matrix[j] = summary(lasso1)[which.min(summary(lasso1)$aicc),]$r2
  coef_matrix[j] = coef(lasso2)[2]
}
```

```{r}
sum(booking_status[hotel['lead_time'] > 100] == 'Canceled') /
length(booking_status[hotel['lead_time'] > 100])
sum(booking_status[hotel['lead_time'] <= 100] == 'Canceled') /
length(booking_status[hotel['lead_time'] <= 100])
```

```{r}
sum(booking_status[hotel['no_of_special_requests'] > 0] == 'Canceled') /
length(booking_status[hotel['no_of_special_requests'] > 0])
sum(booking_status[hotel['no_of_special_requests'] == 0] == 'Canceled') /
length(booking_status[hotel['no_of_special_requests'] == 0])
```

```{r}
lead_time = hotel_predictors[,8]

for (i in (1:30)[-c(5,7,8,12)]) {
  lasso1 <- gamlr(hotel_predictors[,-c(i,5,7,8,12)], y=hotel_predictors[,i], standardize = FALSE,
lambda.min.ratio=1e-3)
  print(summary(lasso1)[which.min(summary(lasso1)$aicc),]$r2)
  print('r2')
  d_hat = predict(lasso1, hotel_predictors[,-c(i,5,7,8,12)], type="response")
  lasso2 = gamlr(cbind(hotel_predictors[,i],d_hat,as.matrix(hotel_predictors[,-c(i,5,7,8,12)])),
lead_time, free=2, standardize = FALSE)
  print(coef(lasso2)[2])
  print(colnames(hotel_predictors)[i])
}
```

```{r}
no_of_specific_requests = hotel_predictors[,17]

for (i in (1:30)[-c(5,7,12,17)]) {
  lasso1 <- gamlr(hotel_predictors[,-c(i,5,7,12,17)], y=hotel_predictors[,i], standardize = FALSE,
lambda.min.ratio=1e-3)
  print(summary(lasso1)[which.min(summary(lasso1)$aicc),]$r2)
  print('r2')
  d_hat = predict(lasso1, hotel_predictors[,-c(i,5,7,12,17)], type="response")
  lasso2 = gamlr(cbind(hotel_predictors[,i],d_hat,as.matrix(hotel_predictors[,-c(i,5,7,12,17)])),
no_of_specific_requests, free=2, standardize = FALSE)
  print(coef(lasso2)[2])
  print(colnames(hotel_predictors)[i])
}
```