Introduction to Data Science - IST 687

Final Project: Exploration of Factors behind Hotel Cancellations

Emmanuel Victor Kamya **|** Abhijith Anil Vamadev **|** Sinem Aytac

December 9, 2021
Syracuse University

<u>Introduction</u>

In this project we are looking at data from hotels and we are trying to find out what the key drivers for cancellations are. In order to find the key drivers we will use the variables available in the data to carry out analysis and draw a conclusion based on the results. The main goal is to give recommendations on what hotel managers/owners can do, in terms of operation, to significantly reduce the number of cancellations. These recommendations will be data driven. We will explore all variables of hotel bookings data, carry out univariate and bivariate analysis and narrow down to the important variables that we deem important to make our recommendations.

<u>Data Preparatory Work</u>

So this is the list of variables that we will be covering for our analysis. And our dependent variable being the IsCancelled variable which is a binary variable that indicates if the booking was cancelled or not. There are overall 40, 600 rows of data with 20 variables. The variables are as follows: IsCancelled, LeadTime, StaysInWeekendNights, StaysInWeekNights, StayInNights, Adults, Children, Babies, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, RequiredCarParkingSpaces, TotalOfSpecialRequests Meal, Country, MarketSegment, IsRepeatedGuest, ReservedRoomType, AssignedRoomType, DepositType, CustomerType. The dependent variable being IsCancelled, looking at the variable IsCancelled we can see that there were 28,938 cancellations, and 11,122 non-cancellations in the data set.

Before starting the exploratory data analysis (EDA) process, we started by preparing the data for viewing. Initially we started by checking for null values, and then negative values, none of which were found in any of the columns. Then, we created a new variable called StayInNights which is the combination of Weekend and Week Nights. We created this variable to see the overall number of days that they would be booking for. After that, we separated the variables as *numerical* and *categorical* variables. We have 12 numerical variables, and 9 categorical variables, as shown here.
- **Numerical**: LeadTime, StaysInWeekendNights, StaysInWeekNights, StayInNights, Adults, Children, Babies, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, RequiredCarParkingSpaces, TotalOfSpecialRequests
- **Categorical**: IsCanceled, Meal, Country, MarketSegment, IsRepeatedGuest, ReservedRoomType , AssignedRoomType, DepositType, CustomerType

After the separation of the variables, we looked at the numerical variables, and it was seen that most of the variables were heavily skewed, so we ran each of the variables through square root transformation. (Initially, we tried logarithmic transformation, but as

0's were present in many of the columns, we decided to go with square root transformation.)

__Data Observations:__

Looking at the statistics for the *cancelled* bookings:
1. On average, Lead time was around 128.7 number of days (SD = 98.8), the data suggests that Lead time is heavily right skewed.
2. On average, StayinWeekenedNights and StayinWeekNights had an average of 1.344 (SD = 1.14) and 3.44 days (SD = 2.46), respectively. Both the variables were rightly skewed.
3. Looking at the adults, children, and babies, the mean of each of the columns is near 0, with very little data in the upper quartiles.
4. The most common meal was BB(Bed and Breakfast) with 7,853 meals while the least was SC(Undefined) with 3 meals.
5. The most common county with cancelled bookings was PRT(Portugal) with 7,438 cancellations while the least being Germany with 146 cancellations.
6. The most common MarketSegment is Online TA with 6,248 while the least common is Complimentary with 33.
7. Booking names are not often not repeated with 11,011.
8. PreviousBookingNotCancelled and PreviousCancellations both have a mean near 0.
9. Most common reserved room type is A with 6,382, and the most common assigned room type is A with 6,047.
10. The most common deposit type is no-deposit with 9,450, while the most common customer type is Transient with 9,416.
11. StayinNights which was the combination of StayinWeekendNights and StayinWeekNights have a mean ner 4.78 days (SD = 3.35), while the maximum is 56 days. The data is very rightly skewed.
12. Required Car Parking Spaces was 0 overall.

Looking at the statistics for the *non-cancelled* bookings:
1. On average, Lead time was around 78.84 number of days (SD = 93.06). The data suggests that Lead time is heavily right-skewed.
2. On average, StayinWeekenedNights and StayinWeekNights had on average 1.3 (SD = 1.14) and 3.09 days (SD = 2.45), respectively. Both of the variables were rightly skewed.
3. Looking at the adults, children, and babies, the mean of each of the columns is near 0, with very little data in the upper quartiles.

4. The most common meal was BB(Bed and Breakfast) with 22,162 meals while the least was SC(Undefined) with 83 meals.
5. The most common county with non-cancelled bookings is PRT(Portugal) 10,192 while the least being Germany with 1,057 non-cancellations.
6. The most common MarketSegment is Online TA with 11,418 while the least common is Complimentary with 168.
7. Booking names are not often mean about 0.
8. PreviousBookingNotCancelled and PreviousCancellations both have a mean near 0.
9. Most common reserved room type is A with 17,017 while the most common assigned room type is A with 10,999.
10. The most common deposit type is no-deposit with 28,749 while the most common customer type is transient with 20,793.
11. StayinNights, which is the combination of StayinWeekendNights and StayinWeekNights, has a mean of 4.14 days (SD = 3.37), while the maximum is 69. The data is very rightly skewed.

Overall Observations:
● Looking at both the data sets, it is observed that there are more observations for the not_cancelled data set compared to the cancelled data, but the trends across the two different data sets remain the same. Nearly all columns regardless if they were in the cancelled or not_cancelled data set showed very similar patterns.

Univariate Analysis

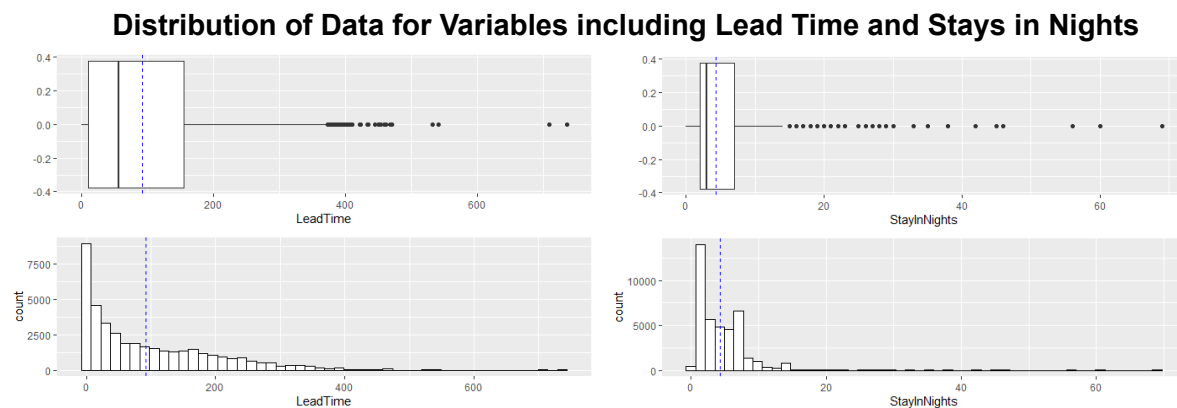**Distribution of Data for Variables including Lead Time and Stays in Nights**

Figure 1: From left to right, LeadTime and StayInNights before transformation

The graphs shown above are the associated Boxplot with the histogram of the associated variables including Lead Time and Stay in Nights before the square root Transformation.

Before Transformation:

- Looking at Lead Time, it looks very rightly skewed with lots of outliers visible after the 3rd quartile.
- StaysInWeekendNights, StaysInWeekNights and StaysInNights all look slightly right skewed, with many data points hitting very few numbers. Most of the frequency of the numbers are right about the mean. Outliers are also visible with all these columns.
- For Adults, Children, Babies, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, RequiredCarParkingSpaces,
- all have means centered around 0. However, all these graphs look to have outliers.
- For Total Number of Special Requests, it looks like the mean is centered between 0 and 1, while most of the special requests seem to be around 0, followed by 1, 2, and 3.
- Overall, looking at the variables, most of the variables have means centered around 0, with very frequency of data. Also, a lot of the variables seem to have outliers. To understand the variables better and to visualize them better, transformation of the data needs to be done. For this we have chosen Square root transformation as most of the variables have data that includes 0. Otherwise, a log transformation would be more appropriate, seeing how skewed the data is.
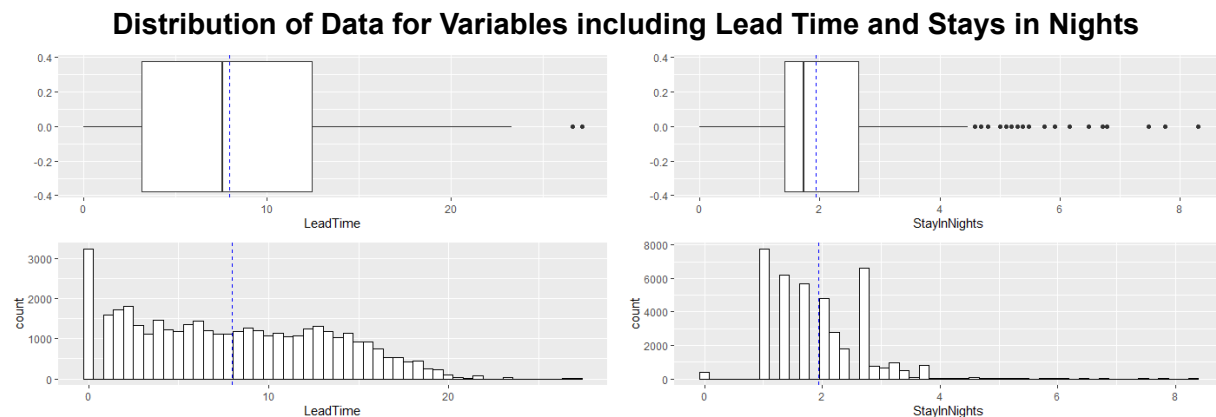
Distribution of Data for Variables including Lead Time and Stays in Nights

Figure 2: From left to right, LeadTime and StayInNights after transformation

After Transformation:

- After the transformations we can see that the Lead Time looks more normally distributed (i.e., less skewed) compared to the original non-transformed data.
- Stays in Weekend Nights, Stays in Week Nights, and Stays in Nights all seem to be more normally distributed relative to their original non-transformed versions.

- For Adults, Children, Babies, Previous Cancellations, Previous Bookings Not-Canceled, Booking Changes, Required Car Parking Spaces, the graphs haven't changed much as most of the data was already centered around 0.
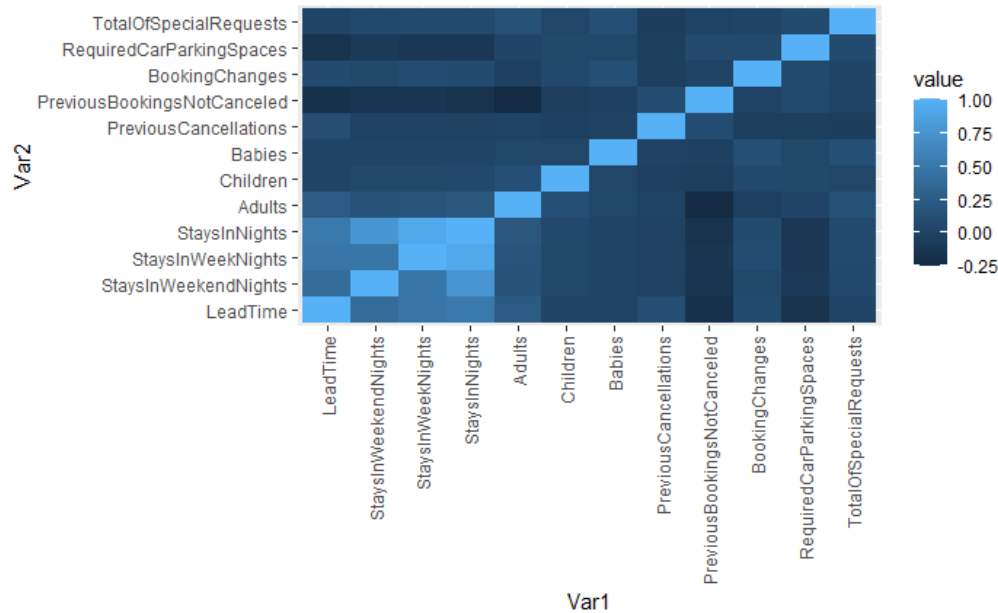
Bivariate Analysis

Figure 3: Heatmap showing the correlation among variables

- We took all the correlation data among each of the numerical variables and plotted them on a heatmap. As the color gets lighter, the correlation gets stronger. Similarly, as the color gets darker, the correlation gets weaker.
- Looking at the heatmap, we can see that Stays in Weekend Nights, Week Nights and Nights are all correlated which makes sense as these variables are dependent on each other. Additionally, Lead Time has a medium correlation with Stays in Nights (r = 0.51).
- Other than that we couldn't find variables that showed significant prominence.

Exploration of the Difference between Cancellations and Non-cancellations:

Next, we explored cancellations and non-cancellations for each variable to understand the patterns predicting cancellations. To do that, we started with all continuous variables in the data set listed in the previous sections. However, we will only report here the variables that might be important for further consideration.

Figure 4 below represents lead time on the y-axis and cancellations on the x-axis while 0 represents non-cancellations and 1 represents cancellations. Specifically, Figure 4 demonstrates the distribution of the number of days elapsed since booking for both cancelled reservations and non-cancellation reservations. If we focus on the mean (i.e., dashed lines), it can be seen that the average lead time is higher for cancellation (M = 128.7, SD = 98.82) than non-cancellations (M = 78.84, SD = 93.06).
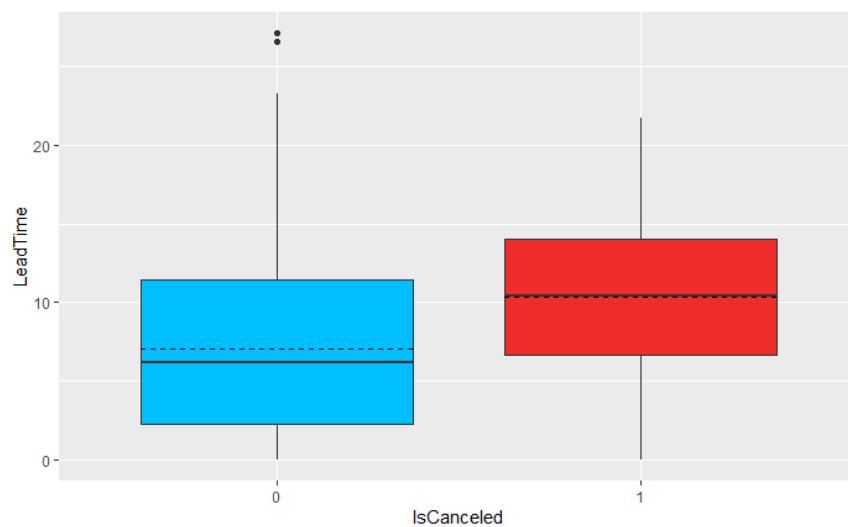

Figure 4: Distribution of Number of Days elapsed since booking (i.e., Lead time) for cancellations and non-cancellations. Red boxplot (i.e., 1) represents cancellations while blue boxplot (i.e., 0) represents non-cancellations. Solid line on each box represents the median whereas the dashed line represents the mean.

Moreover, Figure 5 below demonstrates the distribution of the number of nights booked in a hotel for both cancelled reservations and non-cancellation reservations. Looking at the mean (i.e., dashed lines) for each category, the average nights people booked to stay in a hotel is slightly higher when the reservation is cancelled (M = 4.76, SD = 3.35) versus when the reservation is not cancellations (M = 4.14, SD = 3.37). That is, the visual inspection tells that the number of nights booked to stay in a hotel is slightly more for cancellations than non-cancellations. Additionally, although we do not present the visualizations here for the number of stays in weekend nights and the number of stays in week nights, the data show higher number of stays in nights when book is cancelled than not cancelled for these variables, as well.
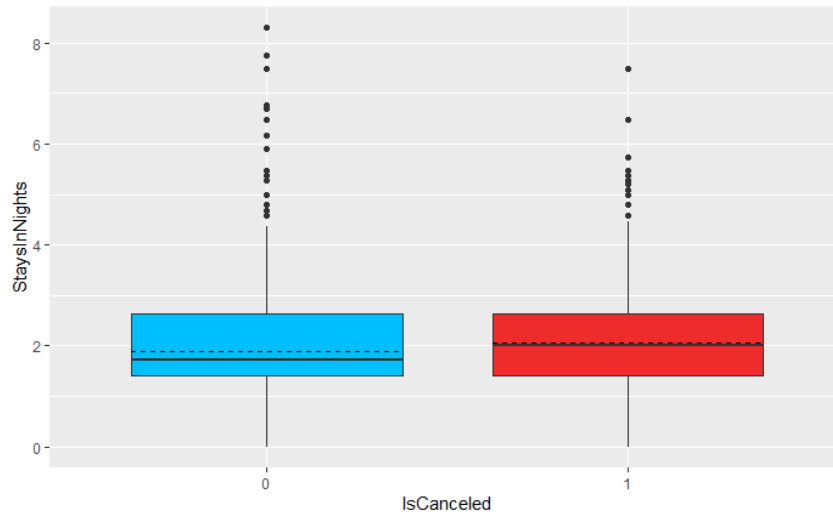
Figure 5: Distribution of Overall Number of Nights people booked to stay in a hotel (i.e., Stays in Nights) for cancellations and non-cancellations. Red boxplot (i.e., 1) represents cancellations while blue boxplot (i.e., 0) represents non-cancellations. Solid line on each box represents the median whereas the dashed line represents the mean.

Furthermore, Figure 6 below illustrates the distribution of the total number of special requests made by customers when the reservation is cancelled versus not cancelled. In contrast to the previous variable presented above, the number of total special requests made by customers is lower for cancellations ($M = 0.49$, $SD = 0.75$) than non-cancellations ($M = 0.67$, $SD = 0.83$). This shows us that cancelled bookings have slightly more special requests than non-cancelled bookings, on average.
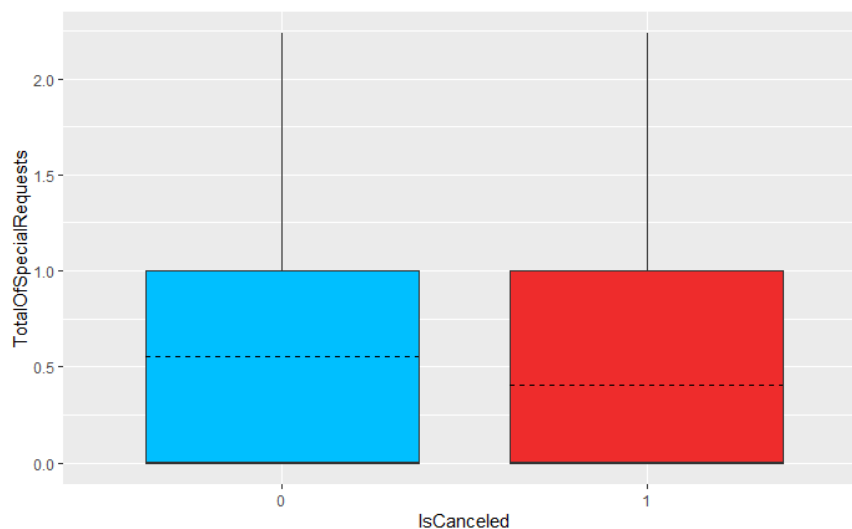

Figure 6: Distribution of Total Number of Special Requests made by customers (i.e., TotalOfSpecialRequests) for cancellations and non-cancellations. Red boxplot (i.e., 1) represents cancellations while blue boxplot (i.e., 0) represents non-cancellations. Solid line on each box represents the median whereas the dashed line represents the mean.

Although we do not take space here to talk about the rest of the continuous variables, we still would like to say a few things before we move on the categorical variables. For the number of adults, children, and babies, previous cancellations, previous bookings not-canceled, booking changes, required car parking spaces, much analysis can't be said between the differences among cancellations and non-cancellations. This is because there are very few data points as well as most of the data points are distributed to very few numbers. However, a lot of outliers are visible. Therefore, in future, it might be a good idea to do some kind of outlier detection. Or it might be better to change these variables into categorical variables, as most of the distribution of the data is allocated to very few points, to explore if they can predict hotel cancellations.

Next, we explored the cancellation and non-cancellation frequencies for the categorical variables in the data set listed in the previous sections. In addition to the raw frequencies, we will also plot the percentages of cancellations and non-cancellations for each level of categorical variables. The reason why we also provide the percentages is to give the reader a better idea upon looking at the visualizations.

Looking at Figure 7A illustrating the frequency of cancellations and non-cancellations for each level of Market Segments includings Complementary, Corporate, Direct, Groups, Offline Travel Agency (TA) and Tour Operators (TO), and Online TA segments. The figure demonstrates that the most common types of market segments for cancelled booking are following: Online TA (N = 6,248), Groups (N = 2,474), Offline TA/TO (N = 1,138), Direct (N = 878), Corporate (N = 351), Complementary (N = 33). However, looking at the total number of cancellations without considering the total number of reservations is useless. Therefore, Figure 7B represents the percentage of cancellations and non-cancellations among the total number of reservations for each market segment. When looking at the figure, it can be more clear that Online TA and Groups have the highest number of cancellation rates compared to the other segments. The cancellation rate for each market segment is following: Online TA with having 35%, Groups with having 42%, Offline TA/TO with having 15%, Direct with having 13%, Corporate with having 15%, and Complementary with having 16% cancellations. Therefore, we can now be more confident that Groups and Online TA have higher cancellation rates compared to other market segments.
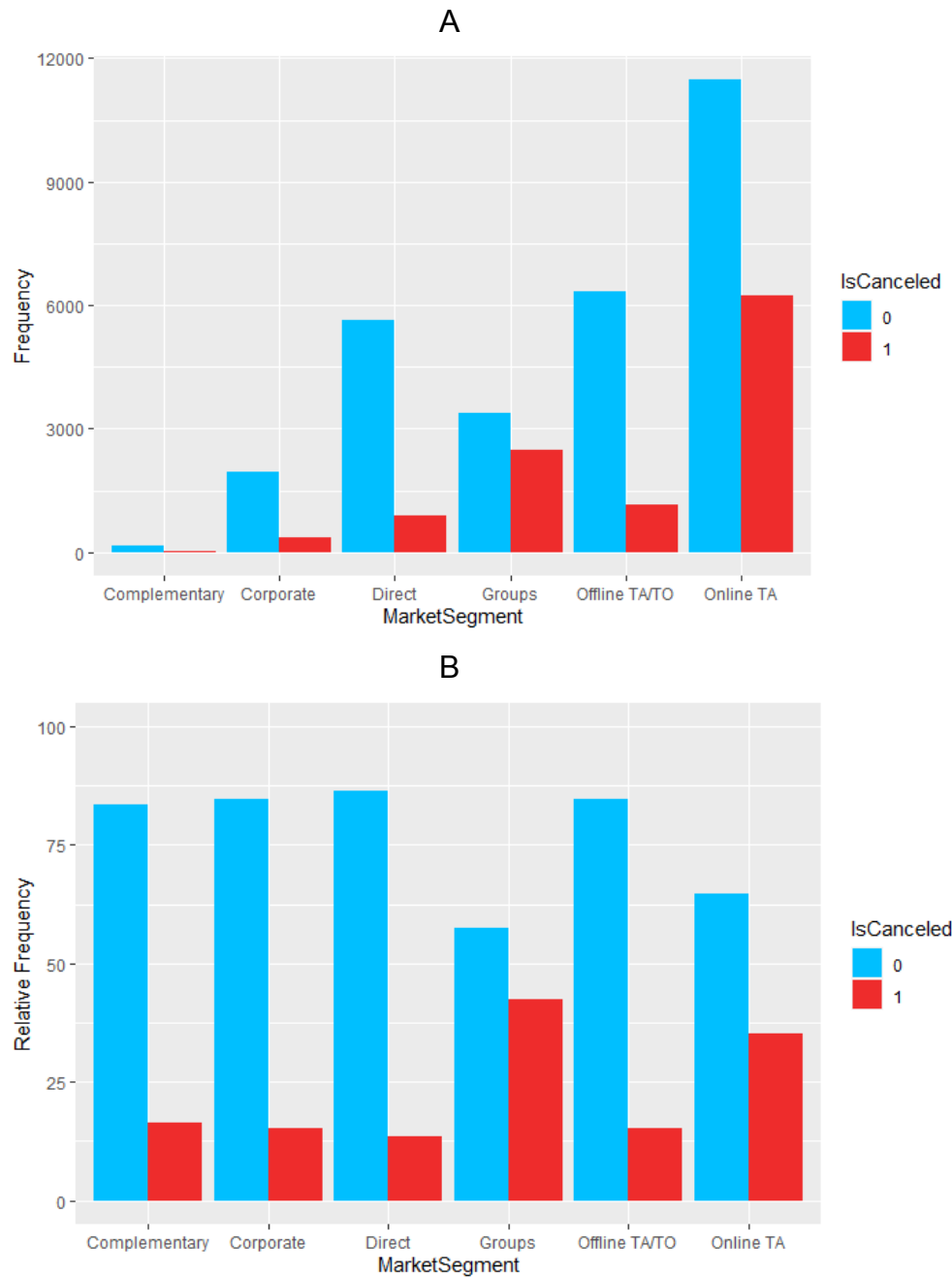
A

B

Figure 7: Both figures A and B have the Market Segment on the x-axis including Complementary, Corporate, Direct, Groups, Offline Travel Agency (TA) and Tour Operators (TO), and Online TA. Both figures represent cancellations in red bars (i.e., 1) and non-cancellations in blue bars (i.e., 0). Specifically, Figure A at the top represents the frequency of whereas Figure B represents the percentages of cancellations and non-cancellations among the total number of reservations in that category.

We will not include any other categorical variable exploration in the report; however, it is still worth mentioning them briefly[1]. For the meals, BB (Bed and Breakfast) is highest among both cancelled and not-cancelled bookings, the least common type of booking is SC (No Meal Package) with very minimal bookings. For countries, as mentioned before, the most common countries where both cancellations and non-cancellations occurred were Portugal, Great Britain, Spain, and Germany, respectively. For both the non-cancelled and cancelled bookings, being a repeated guest was not that common. For the reserved room and assigned Room types, type A was the most common reserved room type irrespective of cancellation terms. No Deposit type was the most-common deposit type among both cancelled and non-cancelled bookings. Finally, transient customer type was the most common among customer type variables, while the least common customer type was contract, irrespective of booking type.
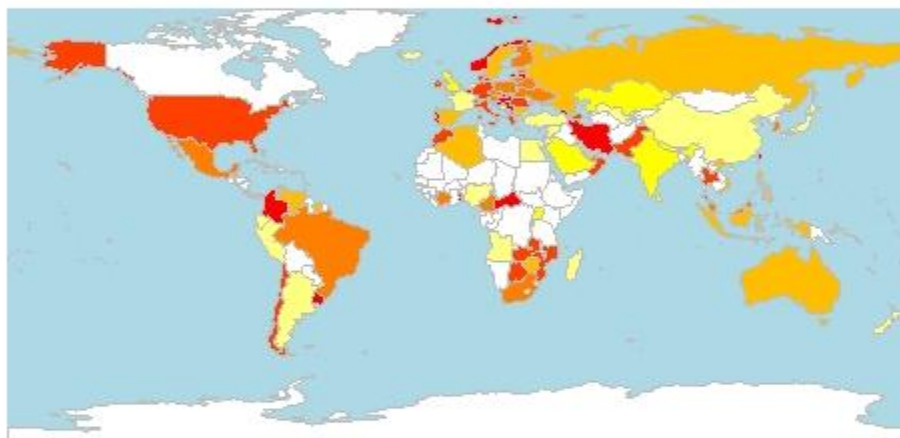
Maps

During the Exploratory data analysis, we focused on the 4 variables of interest namely, Lead time (number of days between booking and cancellation), stay in nights (combination of stay in weekend nights and stay in week nights) market segment, and special requests. Our exploratory data analysis showed a high correlation of lead time and cancellations. Similarly we also observed a high correlation of stay in nights.

---

[1] Interested readers can see our R Markdown code attached along with the report. The code includes the whole data exploration with all variables with the figures inside of the document.

Maps showing clients from countries that cancelled vs clients from countries that didn't cancel

Number of days elapsed: No Cancellations

# Map showing the number of days leading towards cancellation
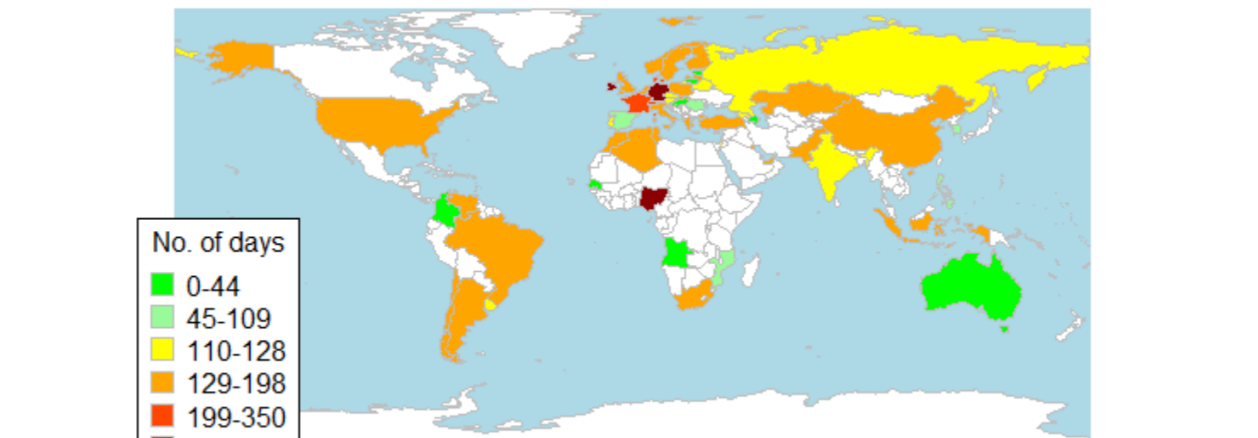
The map above shows the number of days between booking leading to cancellation. The data was filtered to only have data where a client cancelled the booking. The legend shows the color palette and ranges of lead time days used for this map. We can see that the colors ranging from green, pale green and yellow represent those countries that had lead time days between 0-44, 45-109 and 110-128. These countries include but are not limited to; Australia, Angola and Colombia.

Additionally, countries with lead time between 129-198 make up the majority of cancellations. This is evident in the map as many countries have an orange color as shown in the legend. These countries include, but are not limited to; the United States of America, Algeria, South Africa and the United Kingdom.

Lastly, counties with a red and dark red fill represent a lead time above 199 days. We observed that Nigeria (West Africa) , Germany and France (East Europe), are filled with red or dark orange representing the most lead time for cancellations. This implies that We noticed that there is a concentration of such countries in Eastern Europe. To have a closer look, we zoomed in to Europe to observe the clear distribution of lead time days.

High LeadTime cancellations in Europe

1   322
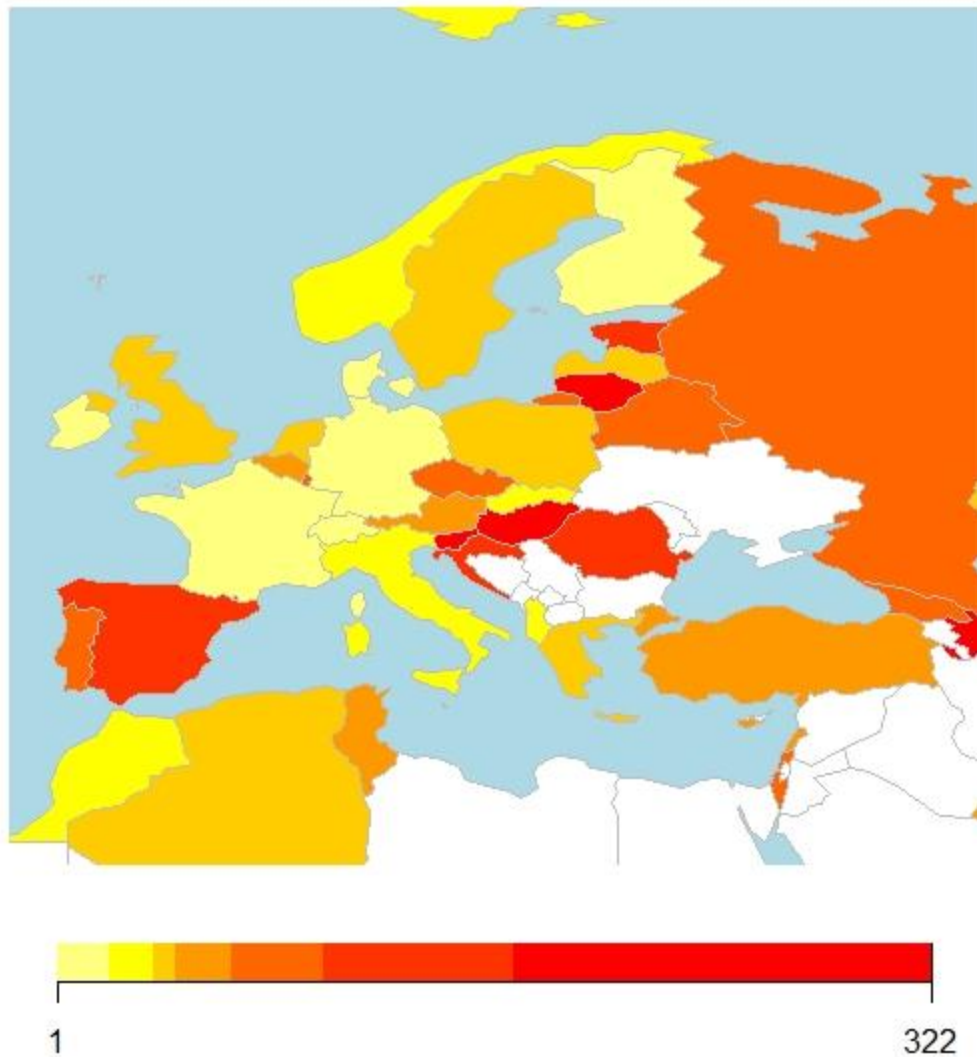
The darker colors ranging from orange, red and dark red represent high lead time days. Since we observed a higher concentration of darker colors in the region of Europe, we decided to zoom in to show the high lead time related to cancellation of bookings.
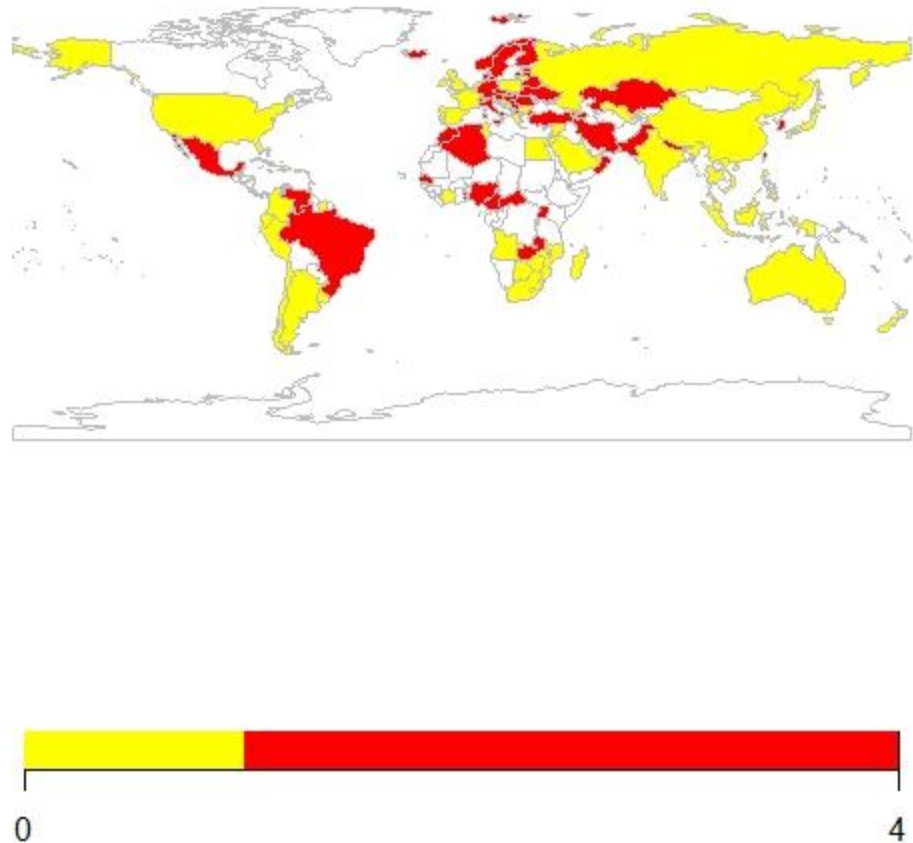
Stay In Week Nights

0   10
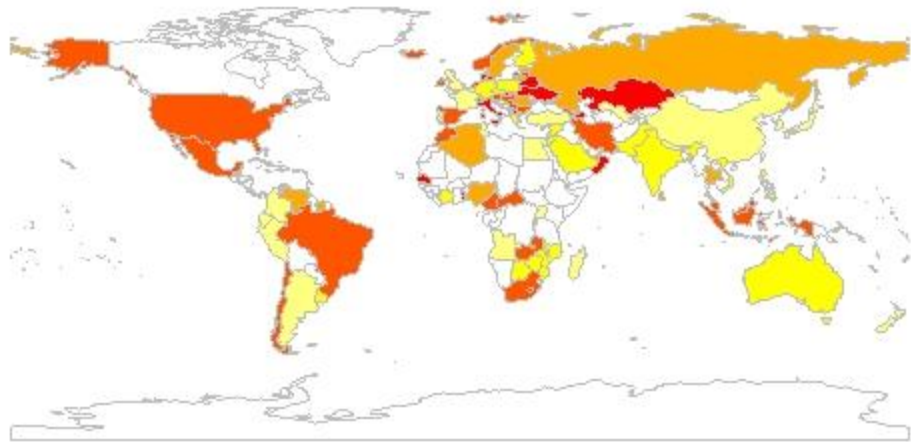
The map above shows a geographic representation of stay in week nights. This data includes both cancellations and none cancellations. The color palette goes from a pale yellow to a dark red.

The map can be interpreted as; countries filled with yellow and light orange having clients that spend less nights during the week at the hotel while countries filled with dark orange and red having clients that spend a higher number of days during the week staying at the hotel. The range for this lead time is from 0-10.
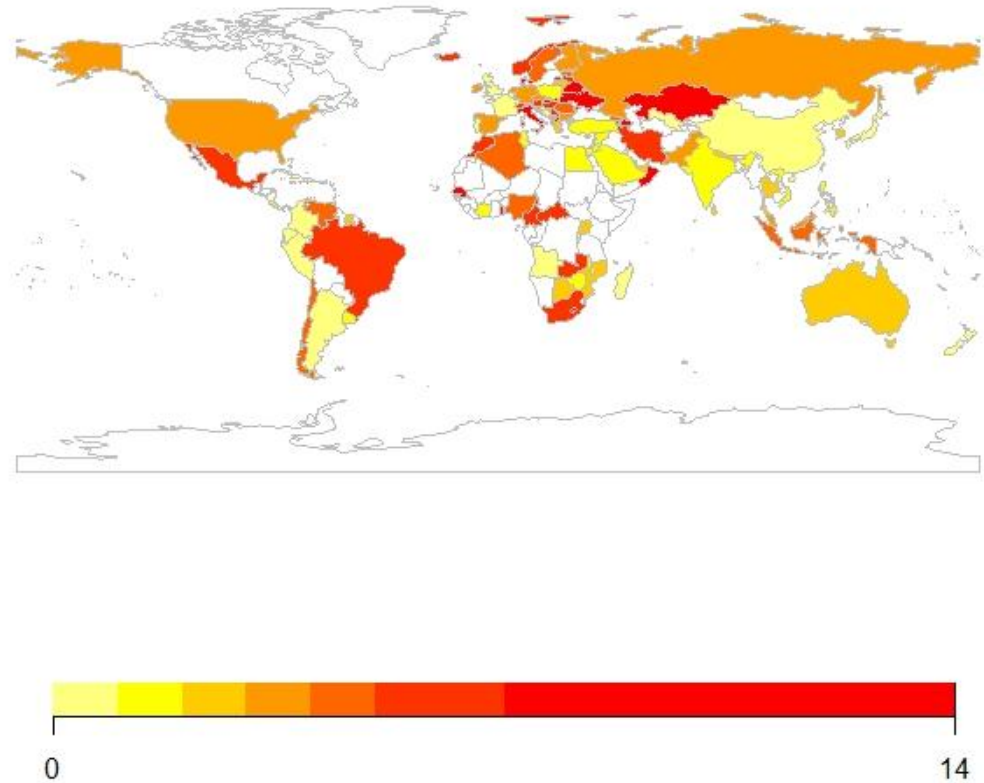
Stay In Weekend Nights

The map above shows a geographic representation of stay in weekend nights. This data includes both cancellations and none cancellations. The color palette goes from a pale yellow to a dark red.

The map can be interpreted as; countries filled with yellow and light orange having clients that spend less nights during the weekend at the hotel. A few examples of these countries include the United States of America, Australia and Russia. Countries filled with dark orange and red having clients that spend a higher number of days during the weekend staying at the hotel. These countries include but are not limited to; Brazil, Algeria and Mexico. The range for this lead time is from 0-4.

Stay In Nights

0   14

We discovered that individually, the variables, StayInWeekNights and StayInWeekendNights did not have a significant effect on cancellations. However we combined these two variables in our exploratory data analysis to have a combination of these variables. This new variable, StayinNights showed to have a high correlation and a significant predictor of cancellations. This was the impetus of prediction using stay in nights.

The map above displays a geographic representation of the origin of clients and the number of days spent at the hotel. The range of stay in nights is 0-14. This range is evident from the combination of the two variables with 14 being the sum of both ranges.

The lighter colors including Yellow, light orange and orange represent less number of days. These are countries like Angola, Australia and the United Kingdom.

Machine Learning: Logistic Regression

Before starting to report the results, we would like to tell the reader which variables we included in our model and why. We included four variables in the logistic regression model as independent variables predicting cancellations. These are lead time, number of nights stayed in a hotel, total number of special requests, and market segment. We decided upon including these variables based off of our data exploration process mentioned in the previous sections. Specifically, we decided upon the factors that could be significant predictors of cancellations if there is a difference between cancellations and non-cancellations by inspecting the boxplots and barplots for each variable. At the end, we had these four variables that seem interesting to explore further with logistic regression.

The analysis conducted through logistic regression has revealed that lead time, number of nights stayed in a hotel, and total number of special requests made by customers significantly predict hotel cancellations ($p < 0.001$). Specifically, when everything else is controlled, every additional day passed after the reservation has been made (i.e., lead time) increases the odds of cancellation by around 13%; every additional night stayed in a hotel increases the odds of cancellation by around 7%; and finally, every additional special request by customers decreases the odds of cancellation by around 58%. These factors significantly predict the cancellations such that lead time and overall stays in nights increases the probability of cancellation; whereas total number of special requests decreases the probability of cancellation.

When focusing on the market segment, the results have shown that market segments including direct ($p = 0.01$), offline TA/TO ($p < 0.001$), and online TA ($p < 0.001$) are significantly different from the reference point, complementary, in terms of hotel cancellations. For instance, the odds for a reservation being canceled is around 45% less if made through direct platforms rather than complementary platforms. Similarly, the odds for a cancellation is around 70% less if made through offline platforms rather than complementary platforms. On the other hand, reservation through online platforms increases the odds for a cancellation by around 95% compared to complementary platforms (when everything else is controlled). However, the model didn't reveal any significant difference between groups and complementary platforms ($p > 0.05$) as well as between corporate and complementary platforms ($p > 0.05$).

Next, we will explore each variable in detail and make predictions regarding cancellations.

1) **Market Segment as a Predictor of Hotel Cancellations**

When all the other variables are controlled, the model suggests the following: Holding lead time, stays in nights, and total number of special requests constant at zero, online bookings have around 25% chance of cancellation, groups have around 13% chance of cancellation, corporates have around 10% chance of cancellation, and complementaries have around 15% chance of cancellation; whereas direct bookings have only around 9% and offline bookings have only around 5% chance of cancellation (see Figure 13).
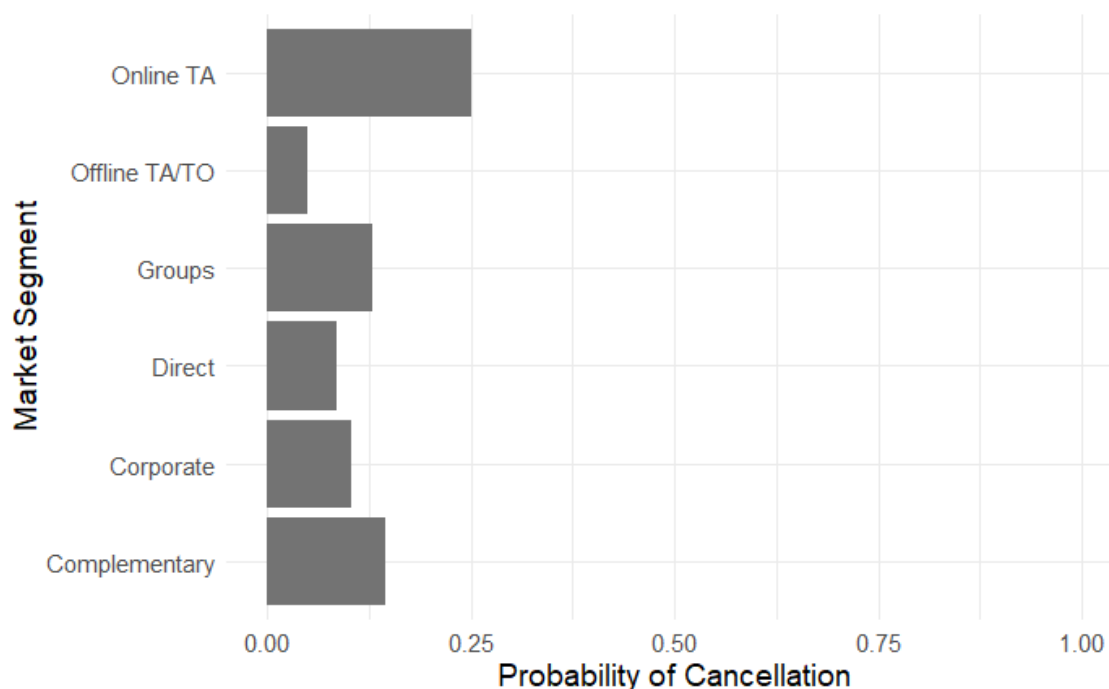

Figure 13: The probability of cancellation for each market segment including online travel agents (TA), offline TA/TO, Groups, Direct, Corporate, and Complementary.

These results have suggested that people will be the most likely to cancel their hotel reservation when made through online platforms rather than the other platforms. Whereas, people will be least likely to cancel their hotel reservation when made through offline platforms rather than the other platforms. Precisely, people will be 6.5 times more likely to cancel their bookings when made through online compared to when made through offline platforms.

2) **Lead Time as a Predictor of Hotel Cancellations**

Further, we explored the probability of cancellation with different lead times including 10 days, one month, six months, and a year when stays in nights and total number of special requests are constant at zero. The results suggest that the probability of cancellation increases with increasing lead time (see Table 1).

Table 1: Probability of cancellation for each market segment after 10 days, one month, six months, and a year has elapsed since booking.

Market Segment/Lead Time	10 Days	1 Month	6 Months	1 Year
Direct	0.12	0.16	0.34	0.51
Corporate	0.15	0.19	0.39	0.56
Online TA	0.33	0.40	0.65	0.79
Offline TA/TO	0.07	0.09	0.22	0.36
Complementary	0.20	0.25	0.48	0.66
Groups	0.18	0.23	0.45	0.62

As also shown in Figure 14 below, people will be more likely to cancel hotel bookings after six months if made through online platforms. However, the model suggests that after a year has passed since booking, cancellations are more likely to be expected for all market segments except offline TA/TO. This is because if a year has elapsed between the entering date of the booking into and the arrival date year, the probability of cancellation expected is over 50% except offline platforms.
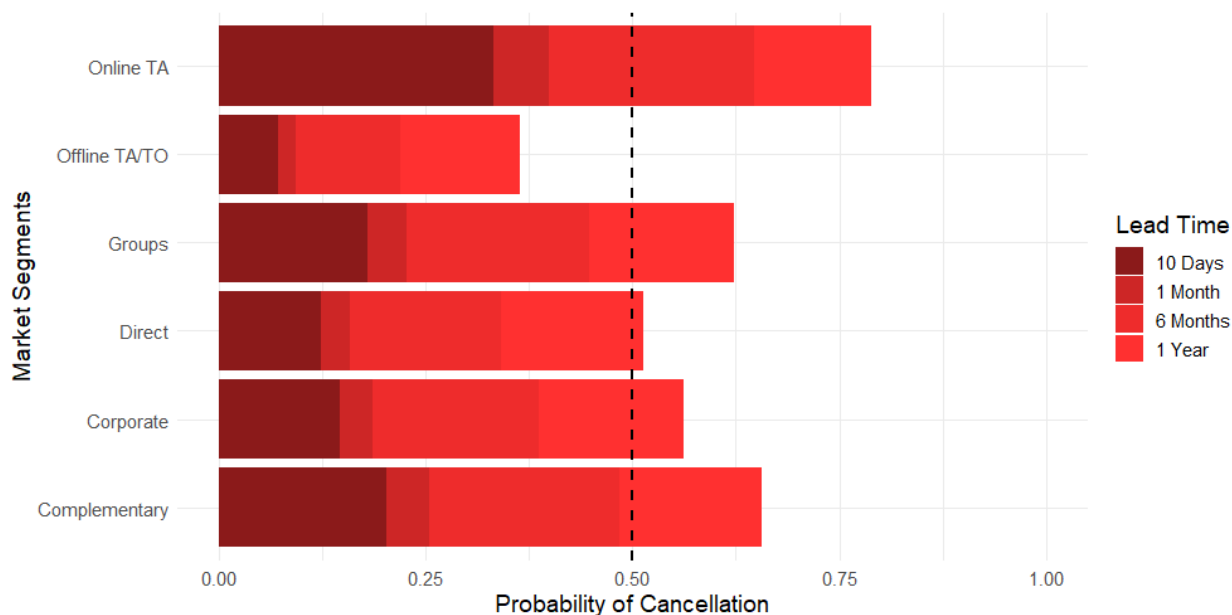
Figure 14: The probability of cancellation for each market segment with different lead times including ten days, one month, six months, and a year. The dashed line at 0.50 represents the chance level.

3) Stays in Nights as a Predictor of Cancellation

Every additional night booked to stay in a hotel means an increase in the odds for a cancellation by around 7%. This suggests that we expect the probability of cancellation to increase as the number of nights booked by customers to stay in a hotel increases.

4) Total Number of Special Requests as a Predictor of Cancellation

Total number of special requests is different than the previous factors because the probability of cancellation decreases as it increases. Thus, we will use it to predict the amount of decrease in probability of cancellations with different numbers of special requests made by customers.

First, we explored the probability of cancellations with a total number of special requests equal to 0, 1, 2, and 3 when lead time is approximately three months and stays at night approximately for four days. The reason that we picked three months and four days as constants is because these values are the average values of these variables. Therefore, we wanted to explore the decrease in probability of cancellation with each additional request when the other variables are constant are their means to have an estimation closer to the real-life situation. Figure 15 demonstrates the results. As seen in the figure, each additional special request strongly decreases the probability of cancellation.
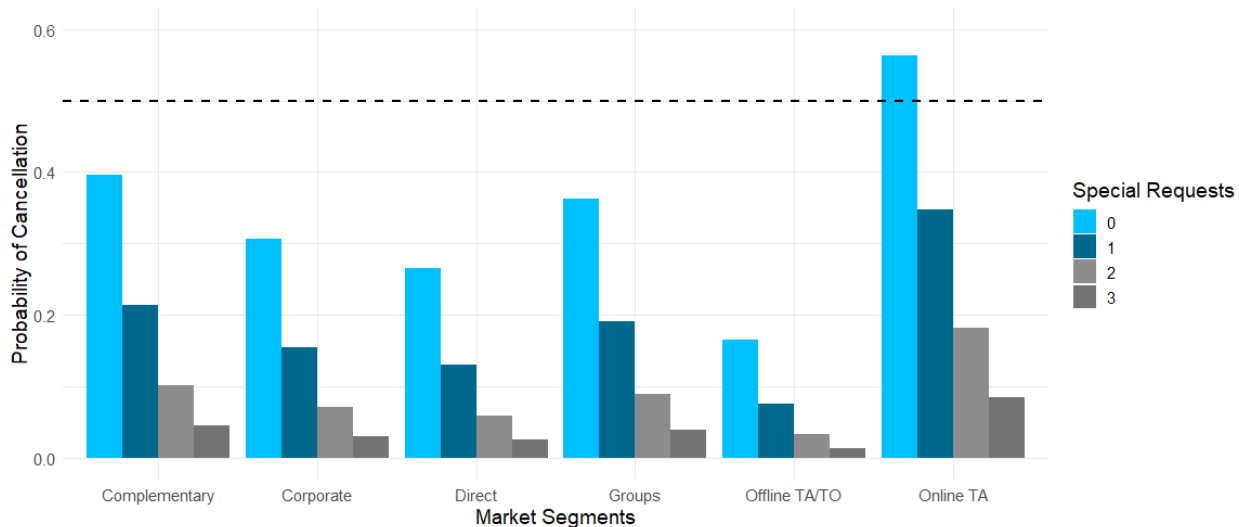
Figure 15: Probability of cancellation for each market segment with a different number of special requests made by customers, when lead time and stays in nights are set to their averages (~3 months and ~4 days, respectively). The dashed line represents the chance level at 0.5. Special requests are zero requests (i.e., 0), one request (i.e., 1), two requests (i.e., 2), and three requests (i.e., 3).

Second, we wanted to explore the probability of cancellation with each additional special request made by customers when lead time is a year and stays at night is four days. Again, the variable representing the number of nights people booked to stay in a hotel is set to its average. On the other hand, this time we wanted to show when a year has passed since booking because we have previously learned that a year lead time predicts more cancellations except for offline platforms. Figure 16 demonstrates the results. As shown in the figure, although we expect people to be more likely to cancel their bookings for all market segments except for offline TA/TO, each additional special requests decreases the probability sharply.
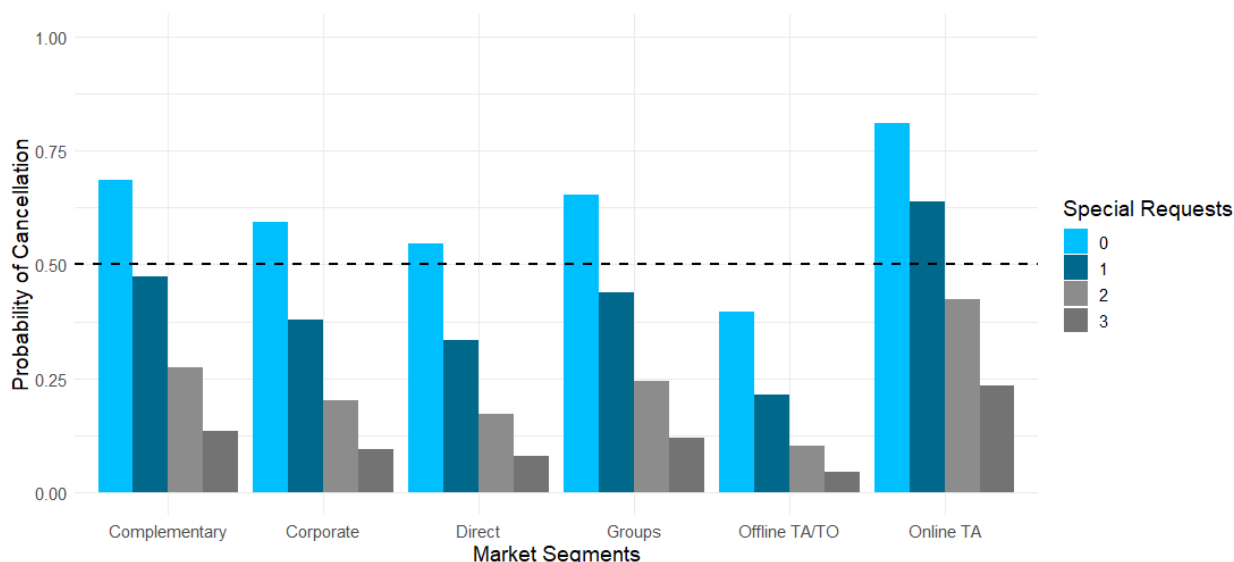
Figure 16: Probability of cancellation for each market segment with a different number of special requests made by customers, when lead time is one year and stays at night for four days. The dashed line represents the chance level at 0.5. Special requests are zero requests (i.e., 0), one request (i.e., 1), two requests (i.e., 2), and three requests (i.e., 3).

CONCLUSION

As the findings have shown, we expect more cancellations of booking when made through online platforms whereas less cancellations when made through offline platforms. Therefore, one short-term solution could be encouraging people to book their hotels though offline travel agencies and tour operators instead of online travel agencies. As a long-term solution, we would like to address the differences between online and offline platforms. Online and offline platforms are naturally different from each other. For instance, offline platforms probably provide more information to their customers because having a representative or an expert physically there for you gives an opportunity to ask any questions that you have before booking. This might decrease the chances of future cancellations. However, this is just an insight and one difference between them. Hence, as a long-term solution, we suggest identifying the differences between online and offline platforms, more importantly finding the reasons behind the least cancellation rate behind offline platforms so that the managers can implement these characteristics to online platforms, as well. For example, providing a 7/24 customer service available for each customer any time they have a question.

Moreover, we expect more cancellations after a year has elapsed between the entering date of the booking into and the arrival date. Therefore, customers can be reached to ask if they still want to keep their reservation if a year has passed since their booking. This way, if there is going to be an opening in the hotel, this could be known before the customer reaches out to the hotel.

Finally, we expect less cancellations with each additional special request made by customers. We suggest encouraging customers to make special requests. This could be simply asked by the travel agencies when booking or incentivizing customers with sales for special requests. Nevertheless, we strongly recommend conducting a controlled experiment to explore the effect of special requests on cancellations by controlling extraneous variables. This is because we cannot know without a controlled experiment if customers made special requests are characterically different from customers who did not make special requests. In other words, it could be that customers are already certain that they will not cancel their bookings when making special requests. Thus, we want to point out the importance of experiments to explore this factor further in the long-term.