IST 687 Final Project
Thursday 3:30-4:50pm
Group 2
May 5th, 2022

Aye Thada Hla
Chandra Shekar Varma Manthena
Erick Leon
Mike DeMaria

Table of Contents

Mission

The executive committee has tasked the data analysis team with determining why our customers cancel their hotel reservations. The team has been provided with an anonymized set of business records containing multiple data points. This analysis will illustrate our findings and recommendations.

The team set out with 4 major objectives. First, to determine what data points correlate with cancellations. Second, how to convert our prospects to customers by reducing the number of cancellations. Third, what actions should be taken to protect the business' revenue streams by determining who are our best, most reliables customers. Fourth, how to grow the business by targeting untapped markets.

Executive Summary

We analyzed 40,060 bookings, 24 different factors and nearly a million data points. We have arrived at ten recommended courses of action. Our statistical analysis has given us a high degree of confidence that these recommendations will lead to improved prospect to client conversion, protect our current revenue streams and assist in growing the business to attract new customers.

Data Provided

We were provided with the following data.

LeadTime - Days the reservation was made in advance

StaysInWeekendNights - Number of weekend nights

StaysInWeekNights - Number of weekday nights

Adults - Number of adults in the party

Children - Number of children in the party

Babies - Number of babies in the party

Meal - Meal package selected by the customer

Country - Customer's country of origin

MarketSegment - Market Segment (such as travel agent, tour operators, etc)

IsRepeatedGuest - First time customer or repeat customer

PreviousCancellations - Number of prior cancellations by the customer

PreviousBookingsNotCanceled - Number of prior bookings by the customer that resulted in a stay

ReservedRoomType - Type of room reserved

AssignedRoomType - Type of room actually assigned

BookingChanges - Number of changes made before check-in

DepositType - Type of deposit and refund policy

CustomerType - Customer segmentation (such as contracted, group, etc)

RequiredCarParkingSpaces - If parking was required

TotalOfSpecialRequests - Special requests made

Data Assumptions and Cleansing

The data set we received was relatively clean. This section will outline any data assumptions we made in this analysis. We recommend consulting with a subject matter expert to confirm if our assumptions are correct. While most of our data was clean on an individual basis, there was some data that did not correlate correctly when taken holistically. In this document, the terms guest and customer should be considered similar. Unless otherwise specified, a guest is any individual, of any age, that stays on our property.

Lead Time: Lead time spans from 0 days to just over two years at 737 days. We are assuming that same day reservations are possible, thus represented by 0 days, and that bookings are permitted two years out. We did not find any missing data in this variable.

Stays In Weekend Nights: Any stay with 0 weekend nights is assumed to be a weekday only booking, such as check-in Monday and check-out Friday. Any number above 2 is assumed to span across multiple weekends, such as a 14 day long stay. We did not find any missing data in this variable.

Stays in Week nights: Any stay with 0 week nights is assumed to be a weekend only booking, such as check-in Saturday and check-out Sunday. Any number above 5 is assumed to span across multiple weeks, such as an 8 day long stay. We did not find any missing data in this variable.

Adults, Children, Babies: These values range from 0 to 55. The customer type for any entry with more than 4 adults is listed as ‘Group’. We are assuming that this represents a group discount, and the entire group is being counted as one entry, as opposed to individual entries, for these large volumes. We should also note that there are many smaller group entries of 1 to 3 adults. We cannot, by the data alone, resolve these discrepancies. Large groups (5 or more guests) represent only 16 of the 5836 group entries, which we consider statistically insignificant, and not worth performing a statistical analysis to derive an estimated value.

TotalPartySize: We derived this column by adding the number of adults, children and babies together. There are 13 entries that have a party size of zero out of 40,060 entries. We do not believe it is possible to have a booking of no guests. Given the small volume of data with this discrepancy, we consider this statistically insignificant, and not worth performing a statistical analysis to derive an estimated value.

Meal: We are assuming a meal plan of Undefined means no meal package was selected. We did not find any missing data in this variable.

Country: The country is represented by a 3 character ISO code. One country is listed as CN with 710 entries. We are unsure if these bookings are meant to represent China, which is elsewhere in the dataset as CHN. There are 464 entries that have a country code of NULL. This value is not a null value, it is the literal string “NULL”. Although this number is very small when compared to the 40,060 observations, the null country represents the 10th largest country when sorted by total number of bookings. Regardless, we still consider these two countries to be fairly small in our analysis. We do not believe we can use statistical analysis to reasonably guess at an associated country.

Market Segment: We are considering Offline TA/TO and Online TA to be two distinct market segments, and not to overlap. We did not find any missing data in this variable.

Is Repeated Guest: We did not find any missing data in this variable.

Previous Cancellations: These values range from 0 to 26. We are considering all values in this range as valid possibilities. We did not find any missing data in this variable.

Previous Bookings Not Cancelled: These values range from 0 to 30. We are considering all values in this range as valid possibilities. We did not find any missing data in this variable.

Reserved Room Type: We did not find any missing data in this variable.

Assigned Room Type: We cannot tell from the data if the assigned room type was changed at the time of check-in, or if the customer knew of the change beforehand. Likewise, we cannot tell if these changes were considered negative situations. For example, the customer may have asked for a room change after booking, or the customer may have been given a free upgrade. We did not find any missing data in this variable.

roomWasChanged: We derived this column by comparing if the reserved room type and assigned room type was changed.

Booking Changes: These values range from 0 to 17. We are considering all values in this range as valid possibilities. We did not find any missing data in this variable.

Deposit Type: We did not find any missing data in this variable.

Customer Type: We are considering Transient and Transient-party to be two distinct customer types, and not to overlap. We did not find any missing data in this variable.

Required Car Parking Spaces: These values range from 0 to 8. We are considering all values in this range as valid possibilities. We did not find any missing data in this variable.

Total of Special Requests: These values range from 0 to 5. We are considering all values in this range as valid possibilities. We did not find any missing data in this variable.

Total Days Stayed: We derived this column by adding the number of week nights and weekend nights together. This data ranges from 0 to 69 days. There are 384 entries where no days were stayed. We do not believe it is possible to have a booking of no days. Given the small volume of data with this discrepancy, we consider this statistically insignificant, and not worth performing a statistical analysis to derive an estimated value.

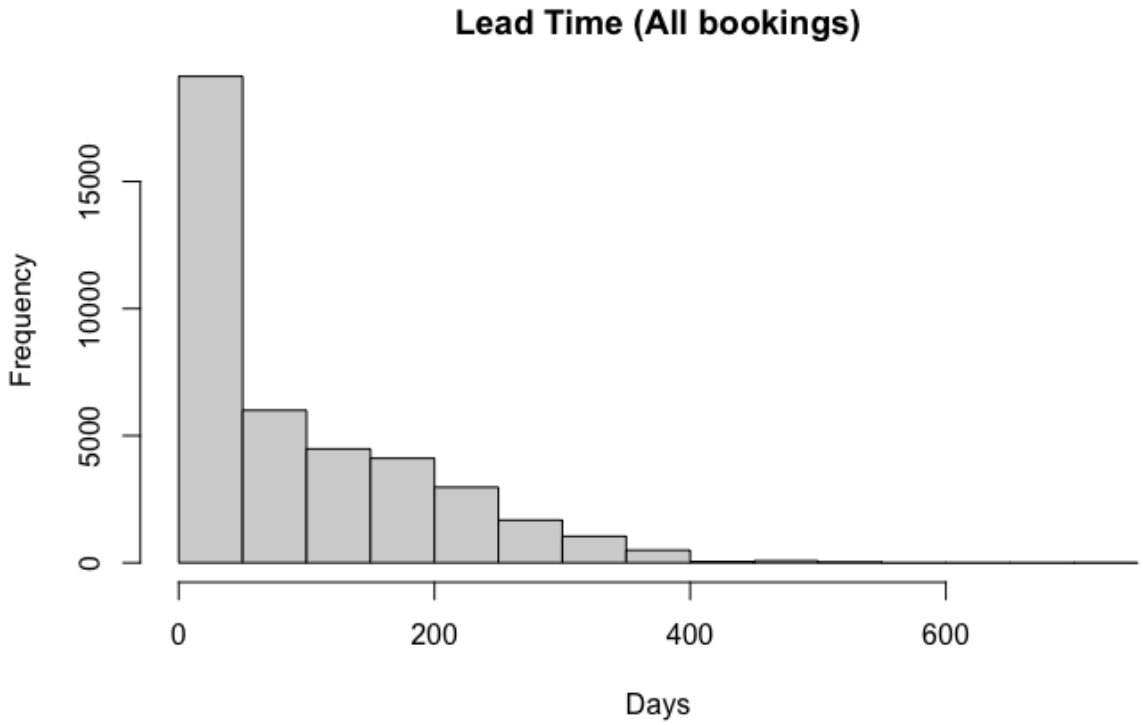
Long Weekend: We derived this column by finding guests who stayed at least one weeknight, at least one weekend and a total of 3 days. This would include Thursday night to Sunday morning bookings, as well as Friday night to Monday morning bookings.

Extended Stays: We derived this column by finding guests who stayed for at least 8 nights.

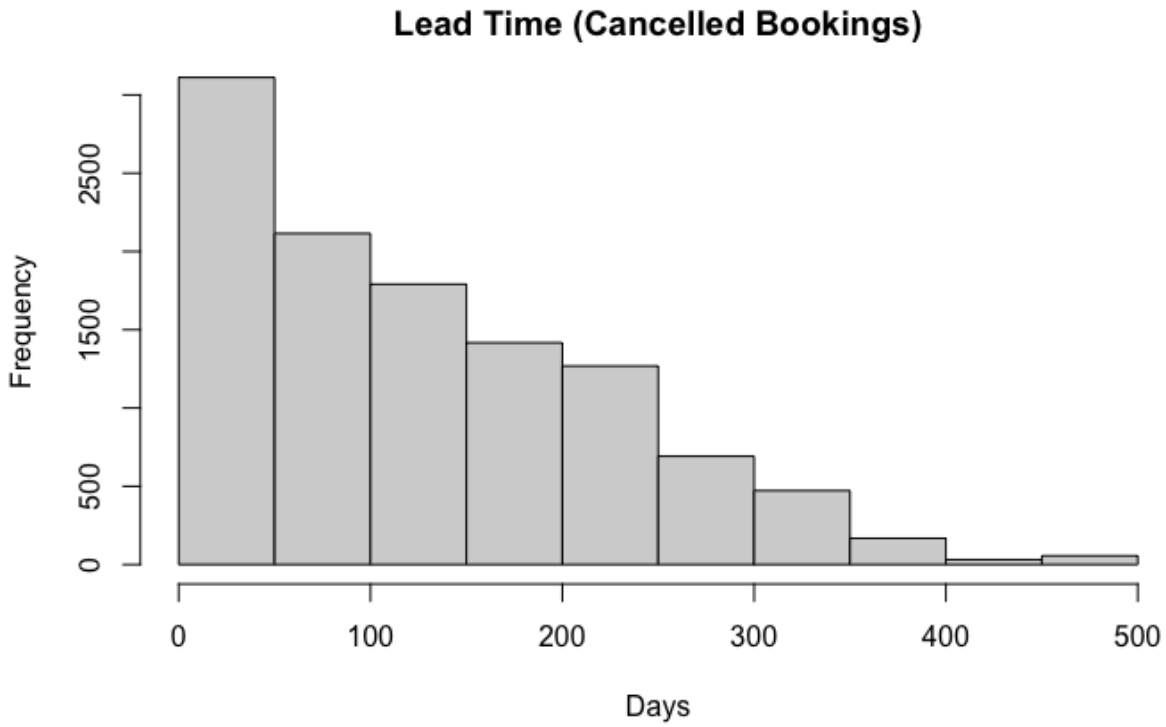
Variable Analysis

The following is our analysis of each individual variable, in isolation. We will look at the variables in combination later in the analysis.

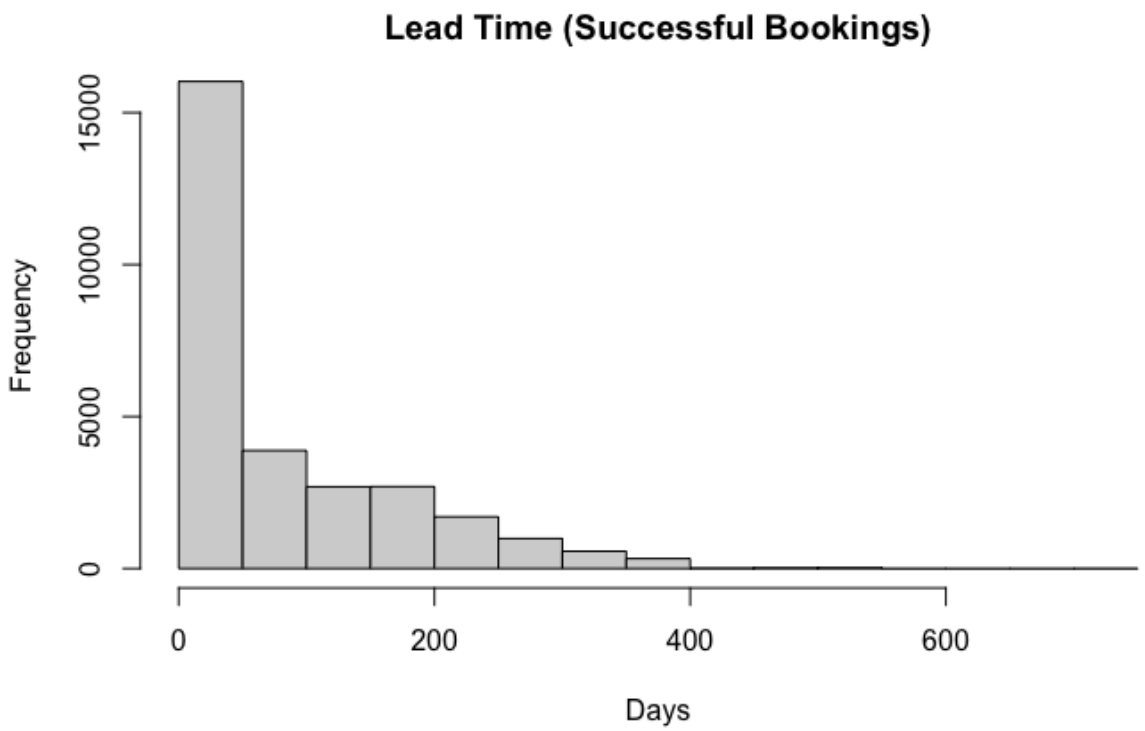
Lead Time:
Lead Time has a mean of 92.68 and a median of 57 . The range is from 0 to 737 days.

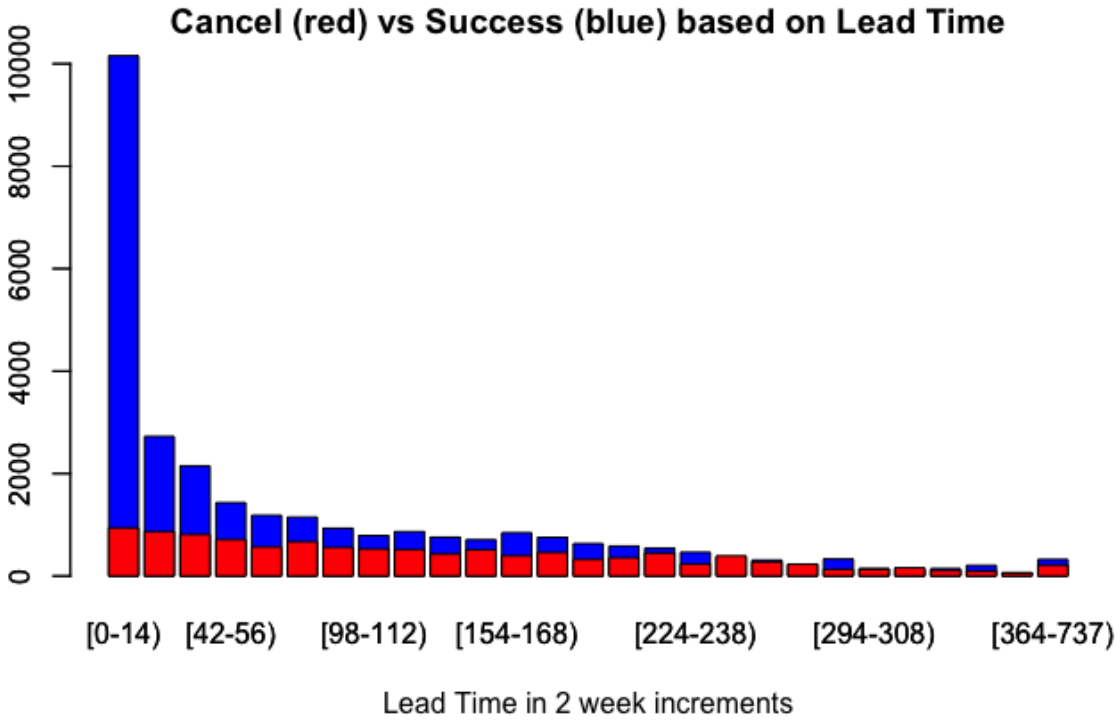


Lead Time (Cancelled Bookings) has a mean of 128.68 and a median of 109 . The range is from 0 to 471.



Lead Time (Successful Bookings) has a mean of 78.84 and a median of 38 . The range is from 0 to 737.

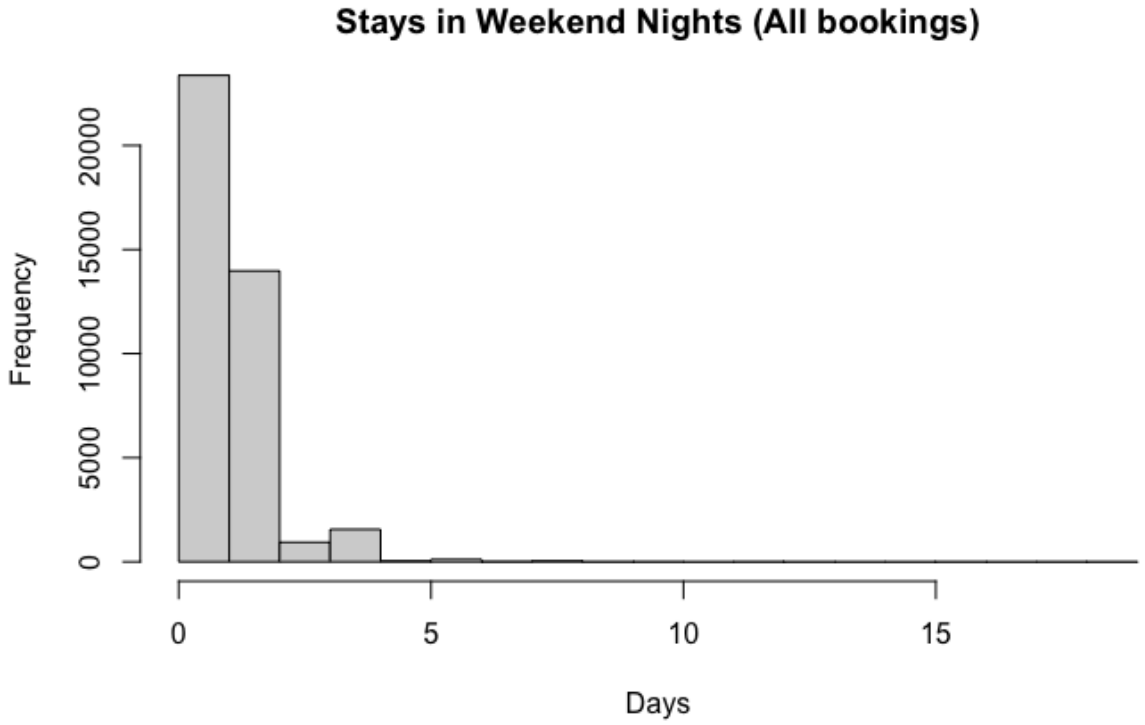




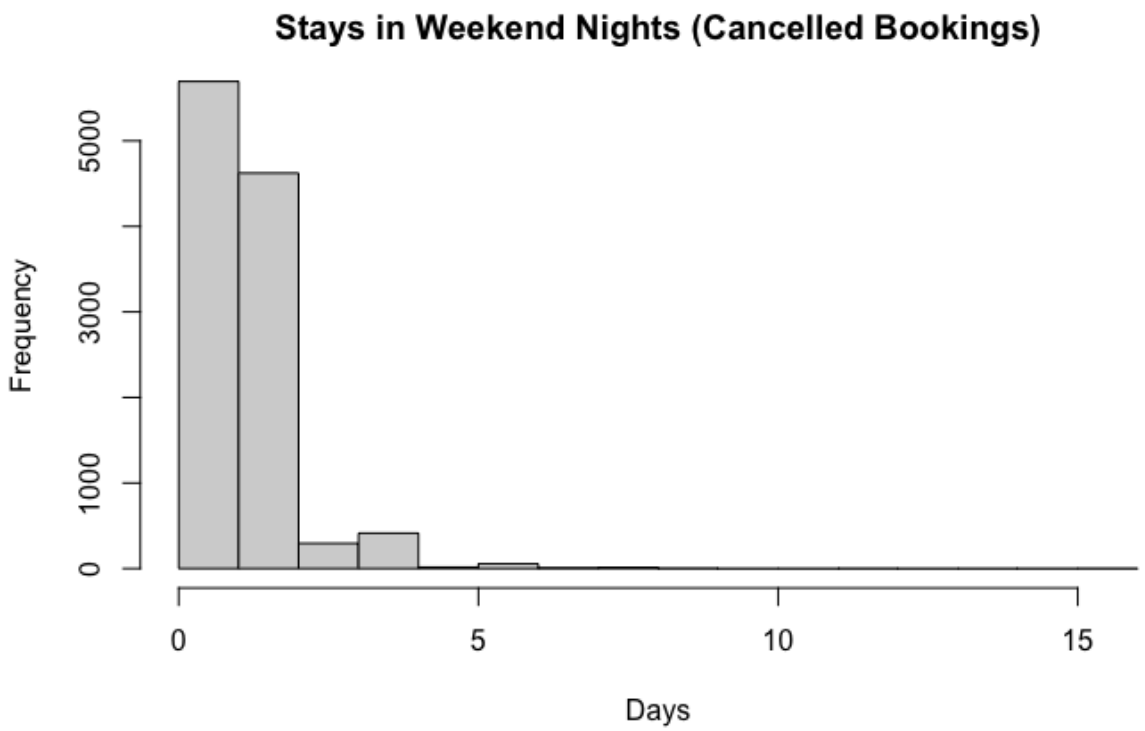
Analysis: When we bucket the cancellations into two week increments, we can see that the number of cancellations, as a percentage of the total bookings, change over time. If a booking occurs within 6 weeks of the stay, there is a high chance that the customer will not cancel.

After 6 weeks, it's about evenly likely that the customer will cancel. After 6 months, it's more likely that the customer will cancel.

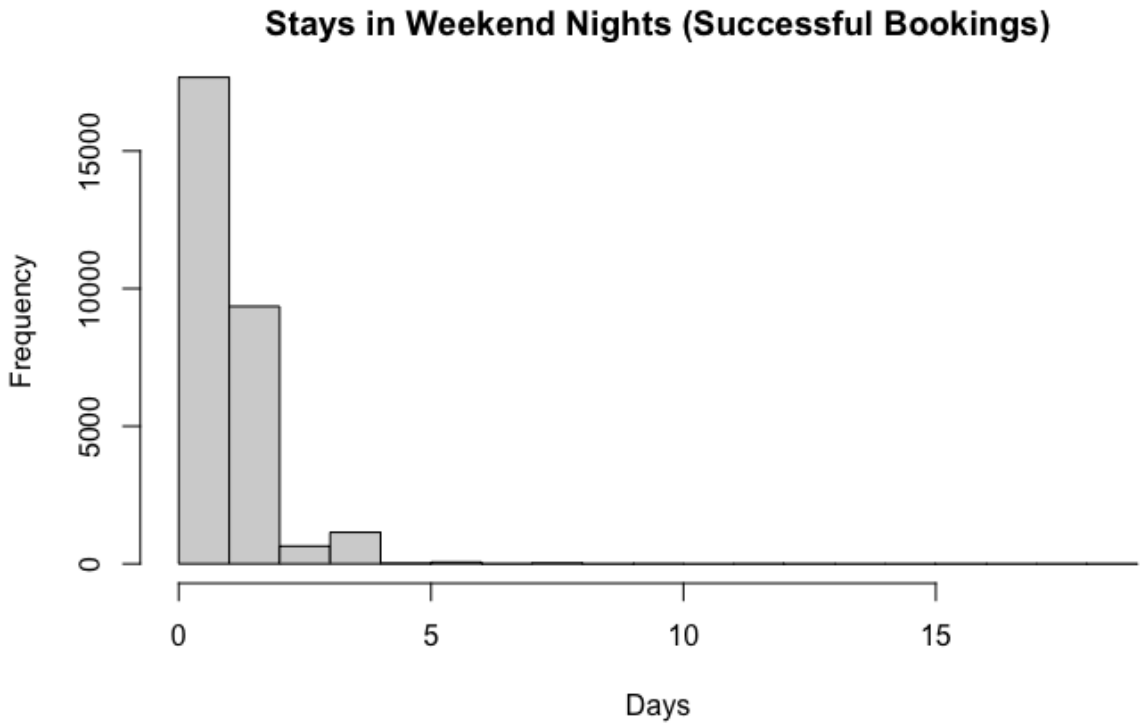
Stays In Weekend Nights:
Stays in Weekend Nights has a mean of 1.19 and a median of 1 . The range is from 0 to 19.

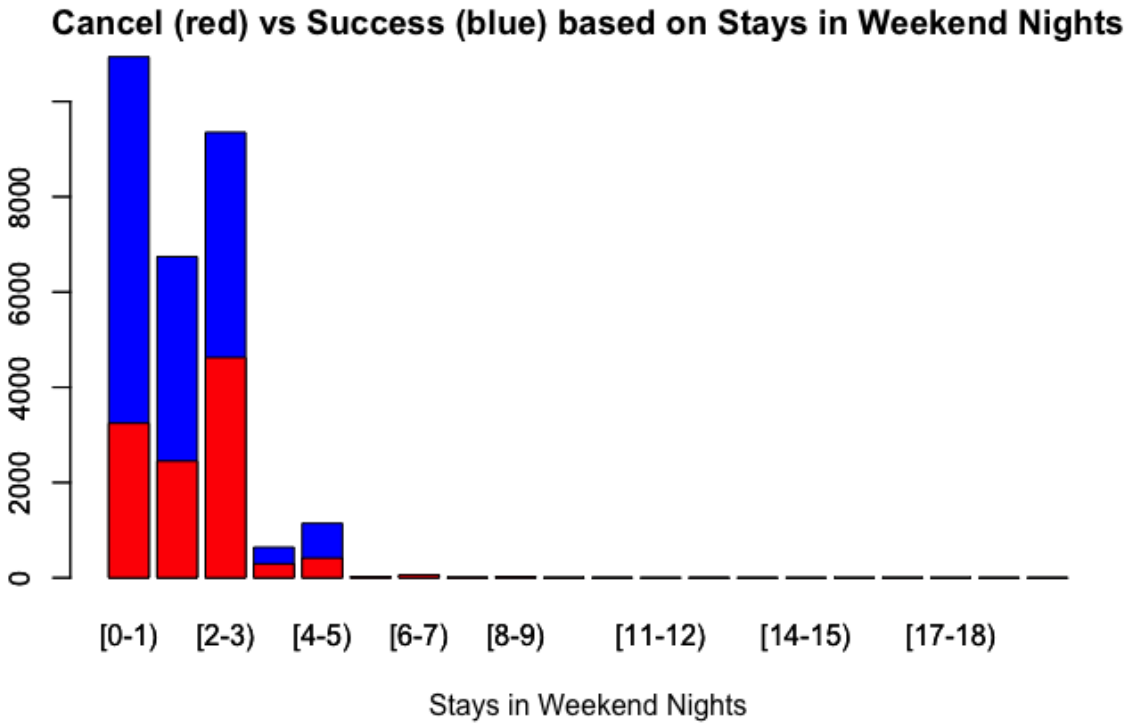


Stays in Weekend Nights (Cancelled Bookings) has a mean of 1.34 and a median of 1 . The range is from 0 to 16.



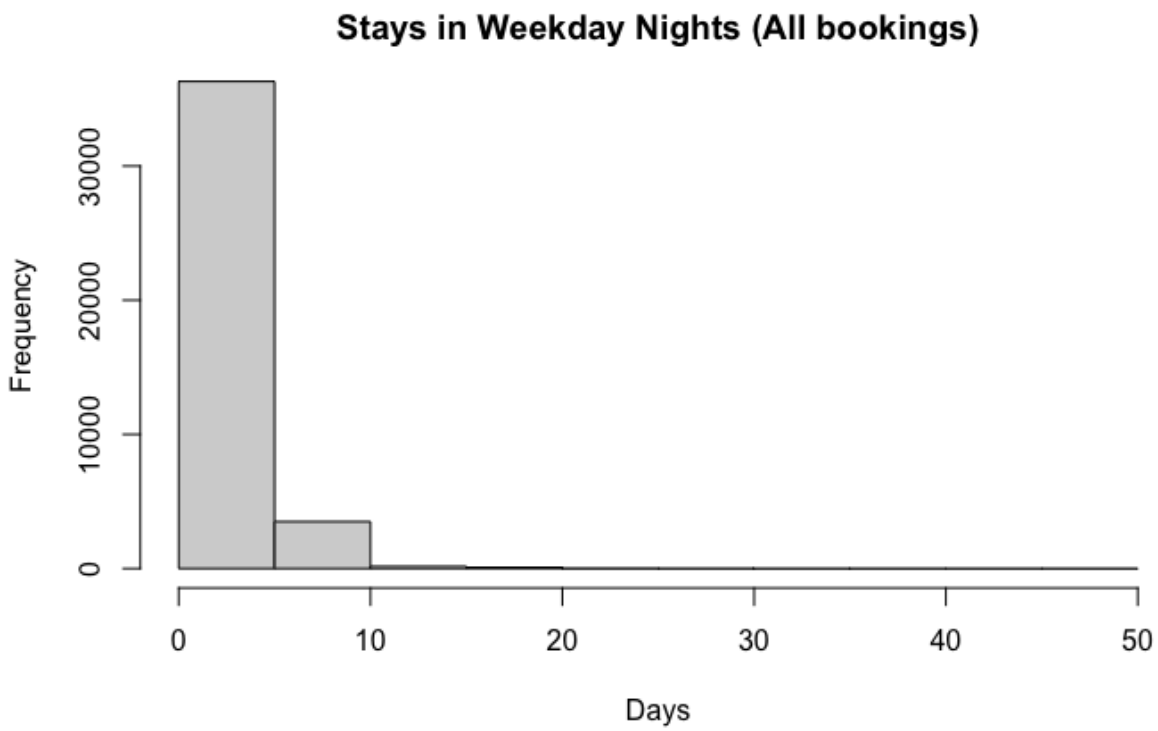
Stays in Weekend Nights (Successful Bookings) has a mean of 1.13 and a median of 1 . The range is from 0 to 19.



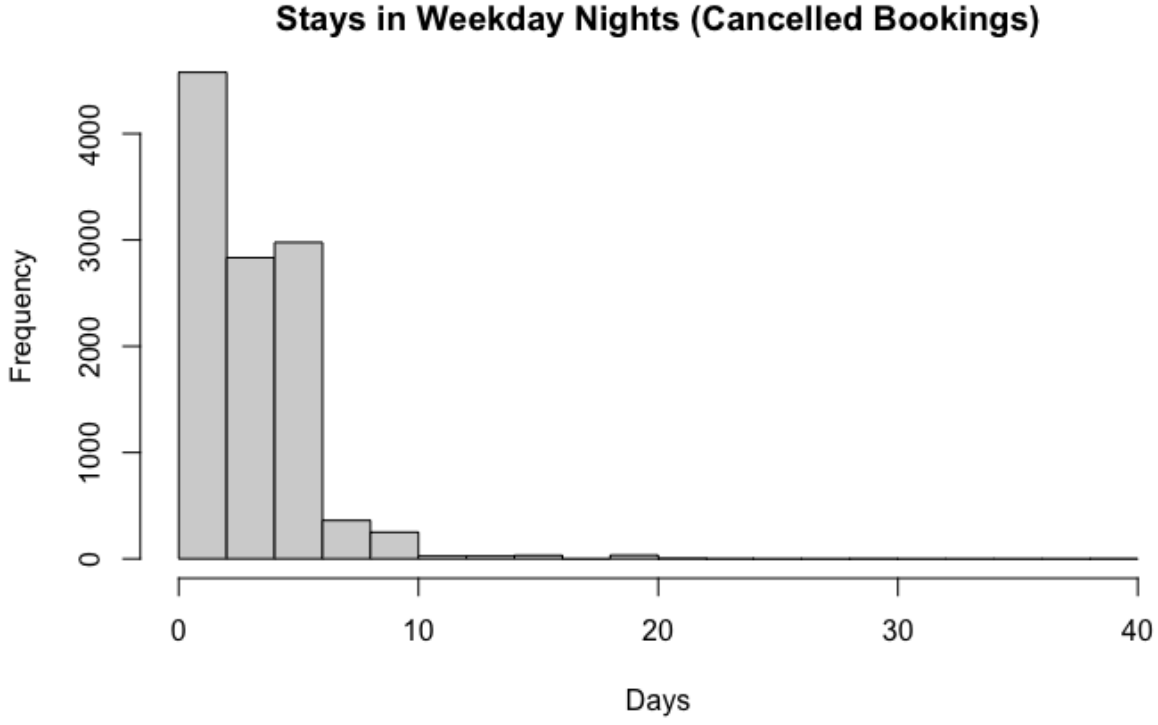


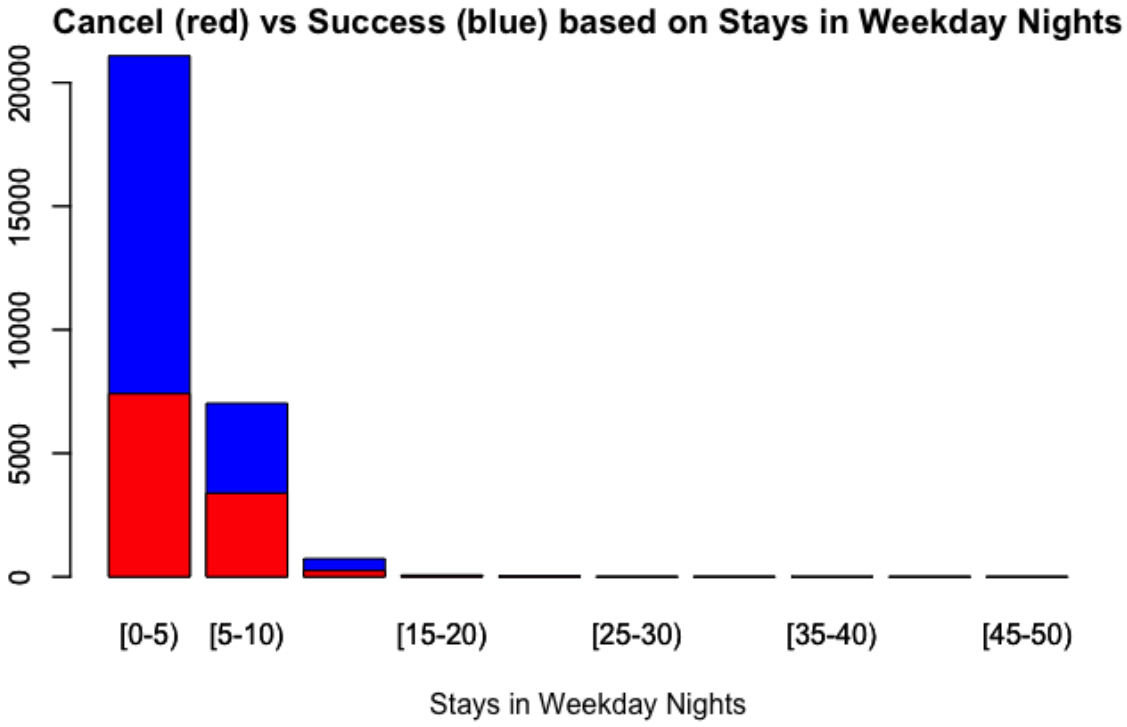
Analysis: We do not see any significant correlation in cancellations and stays in weekend nights.

Stays in Weekday nights:
Stays in Weekday Nights has a mean of 3.13 and a median of 3 . The range is from 0 to 50.



Stays in Weekday Nights (Cancelled Bookings) has a mean of 3.44 and a median of 3 . The range is from 0 to 40.

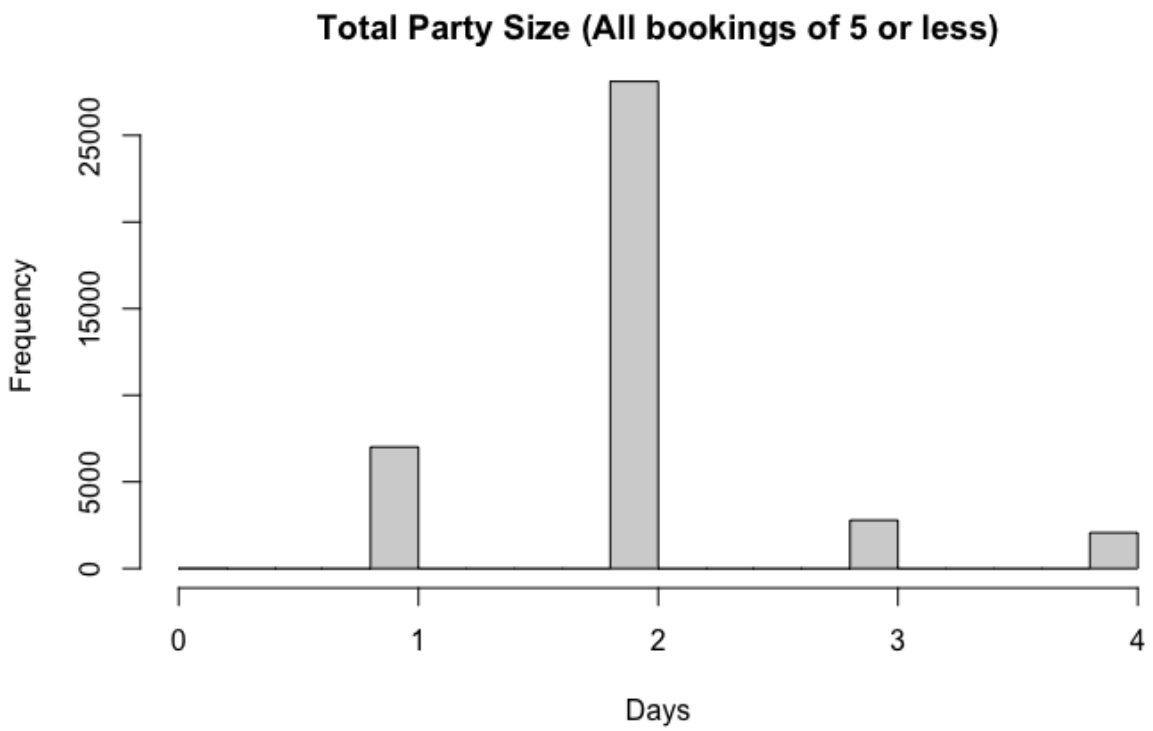




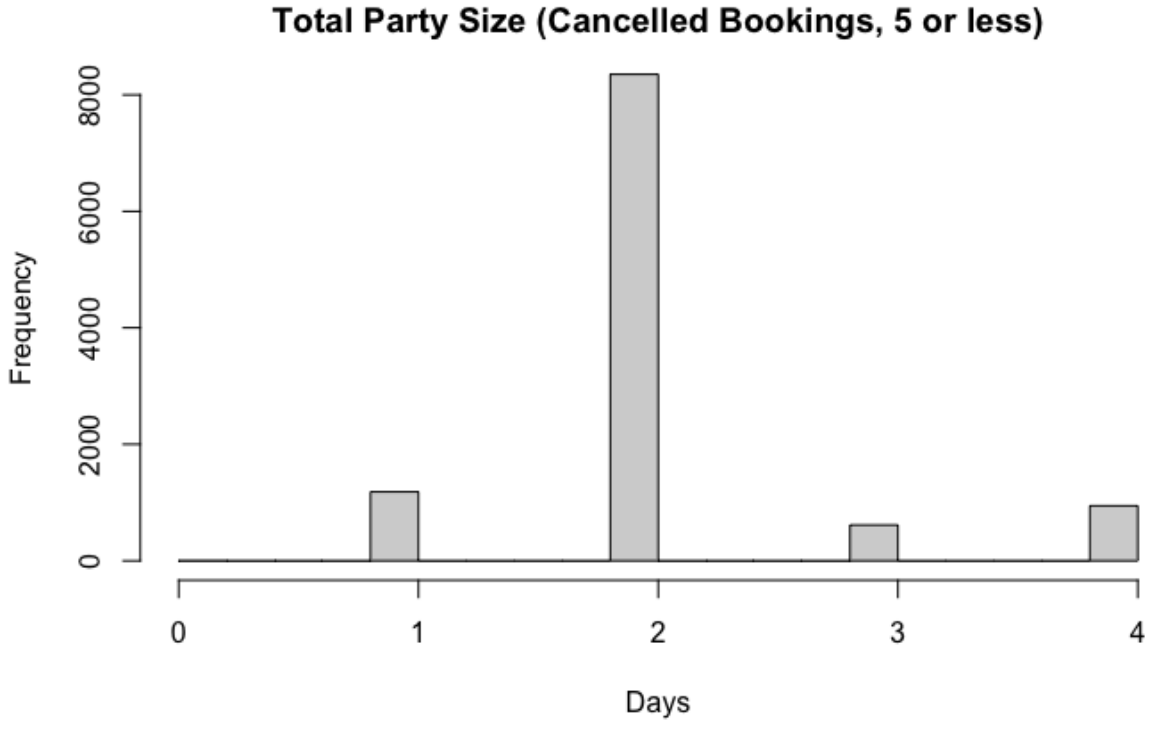
Analysis: We do not see any significant correlation in cancellations and stays in weekend nights.

Adults, Children, Babies:
We have opted to provide statistics based on total party size.

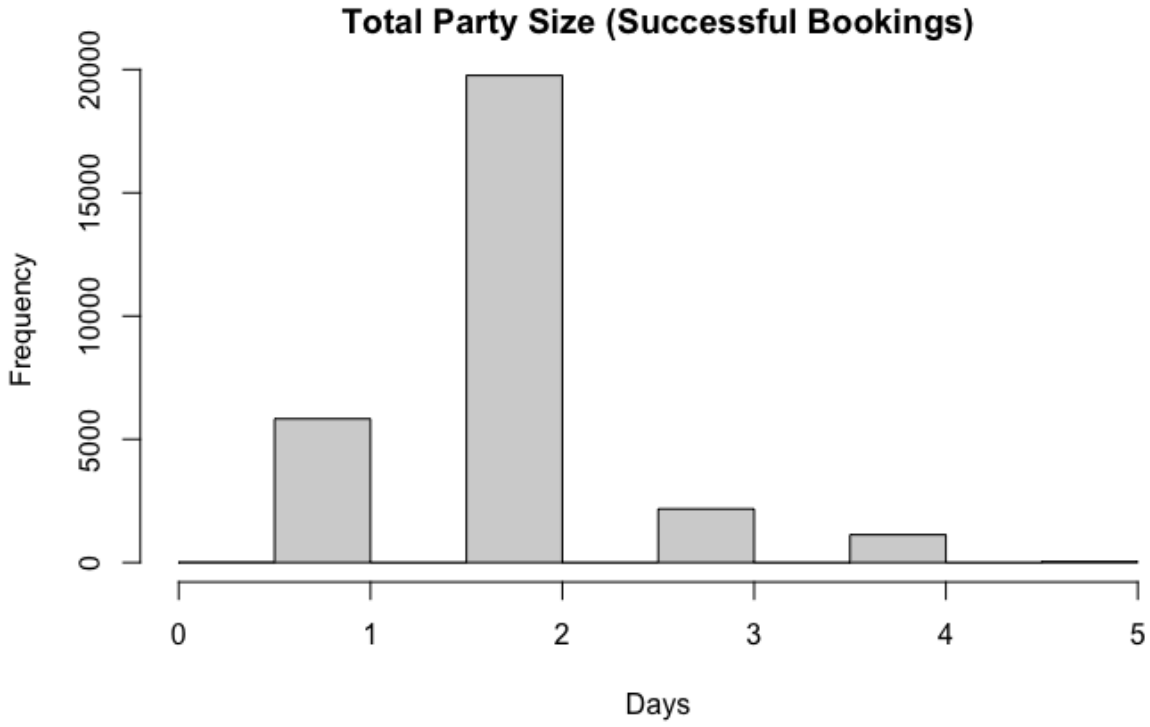
TotalPartySize:
Total Party Size has a mean of 2.01 and a median of 2 . The range is from 0 to 55.

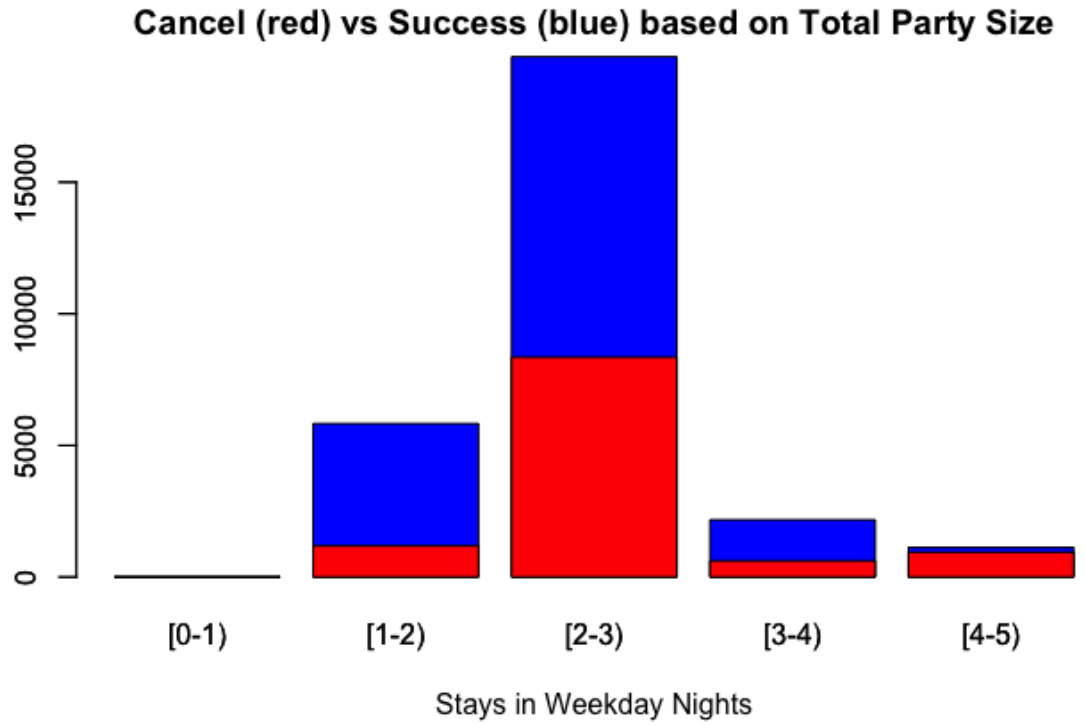


Total Party Size (Cancelled Bookings) has a mean of 2.15 and a median of 2 . The range is from 0 to 55.



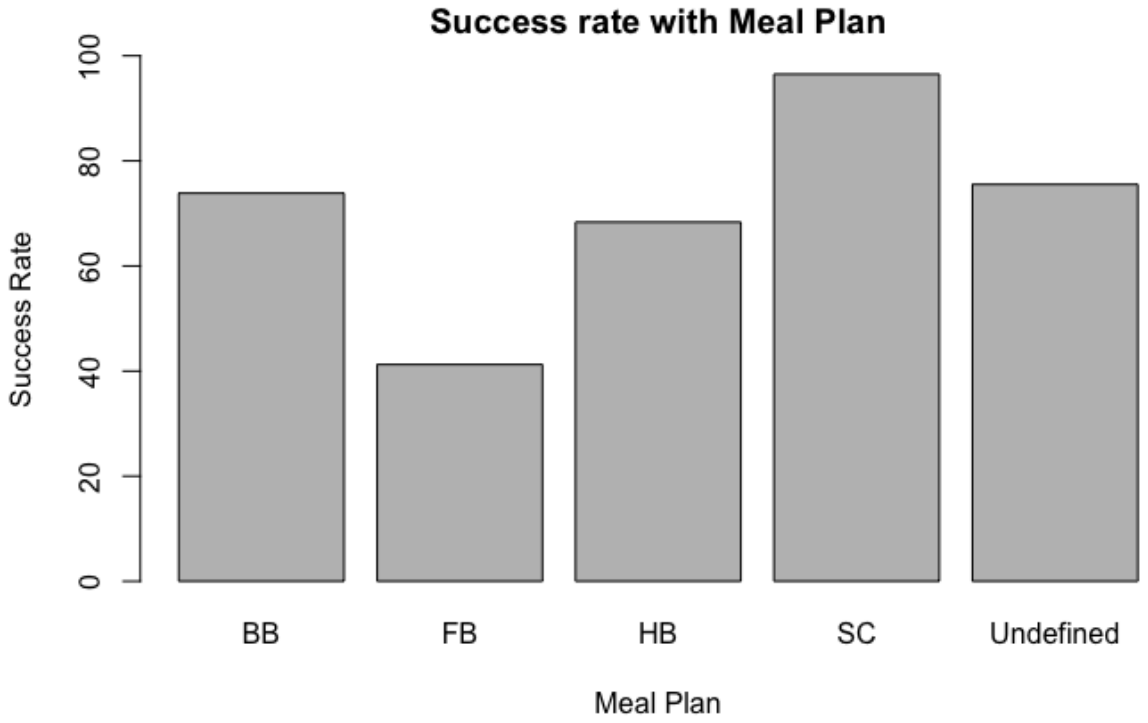
Total Party Size (Successful Bookings) has a mean of 1.95 and a median of 2 . The range is from 0 to 5.





Analysis: Stays of 1-2 days and 3-4 days perform better than 2-3 days.

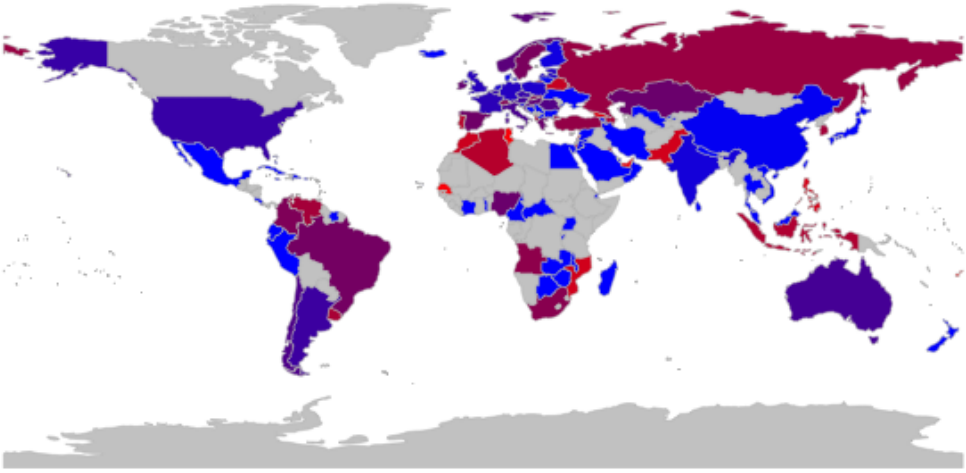
Meal:
Meal Success Rate has a mean success rate of 71.10 and a median of 73.86 . The range is from 41.25 to 96.51.



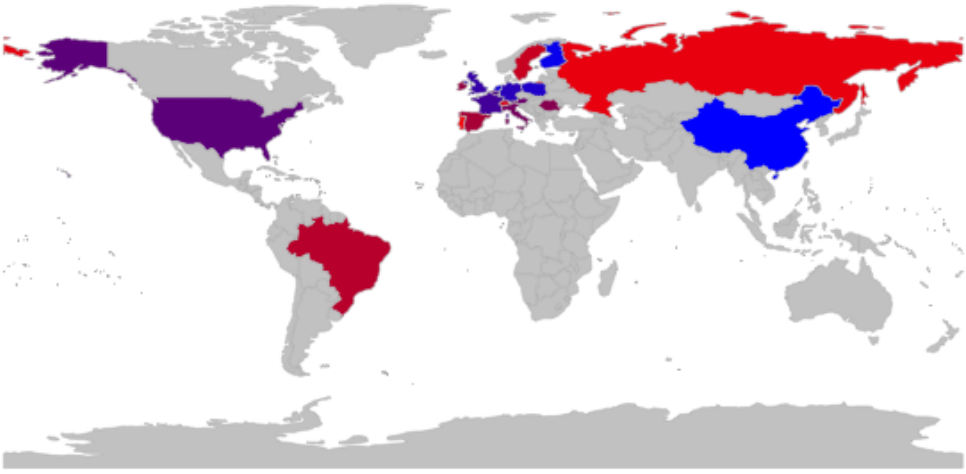
Analysis: The FB meal plan performs significantly worse than average, while the SC plan performs significantly better. Note that the FB represents full board, while SC means no meal plan. Our data analysis does not have the costs associated with these choices. We cannot ascertain if the full meal plan costs more than the no plan option. A cost difference could influence the cancellation rate.

Country:
Guests have arrived from 125 distinct countries. Country Cancellation Rate has a mean of 20% and a median of 11.912%.

Cancellations Rate By Country (Lowest is blue, Highest is red)



Cancellations Rate For Top 20 Countries



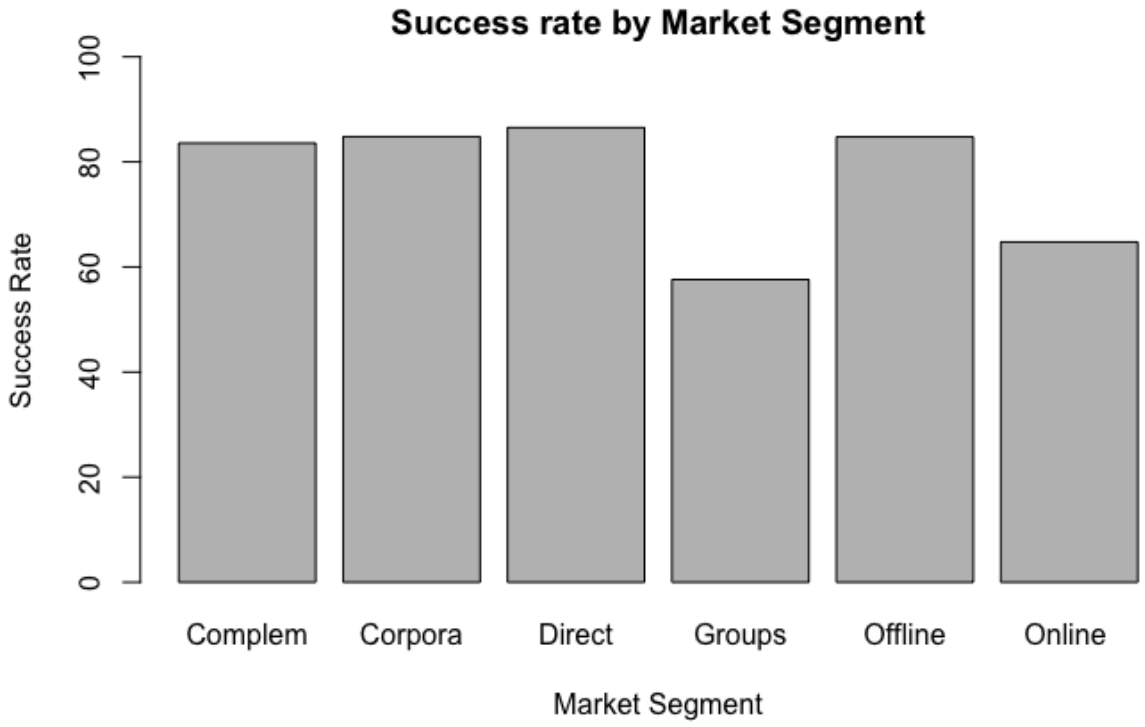
Analysis: We illustrated the cancellation rate for every distinct country, and another map for the top 20 countries. The top 20 countries account for 38,354 bookings, or 95.7% of our business. Each country in the top 20 has at least 131 bookings. The top 10 countries, by quantity of bookings are as follows: Portugal (PRT), Great Britain/UK (GRB), Spain (ESP), Ireland (IRL),

France (FRA), Germany (DEU), Netherlands (NLD), United States (USA), Italy (ITA), Belgium (BEL). When evaluating based on party size, the top 6 countries retain the same ordering. The 7-10th highest visitor count are United States (USA), Netherlands (NLD), Switzerland (CHE), Brazil (BRA).

We also evaluated the likeliness to cancel for the top 20 countries. Of the top 20 countries that book at our hotel, the most to least likely to cancel are: Portugal (PRT), Russia (RUS), Switzerland (CHE), Sweden (SWE), Brazil (BRA). The least likely to cancel are China (CHN), Finland (FIN), Neatherlands (NLD), Poland (POL) and Germany (DEU). We noted that Portugal (PRT) has both the highest number of bookings for any given country and the worst cancellation rate amongst the top 20 countries. Great Britain (GRB), conversely, is very low on the cancellation rate but represents the second highest number of guests.

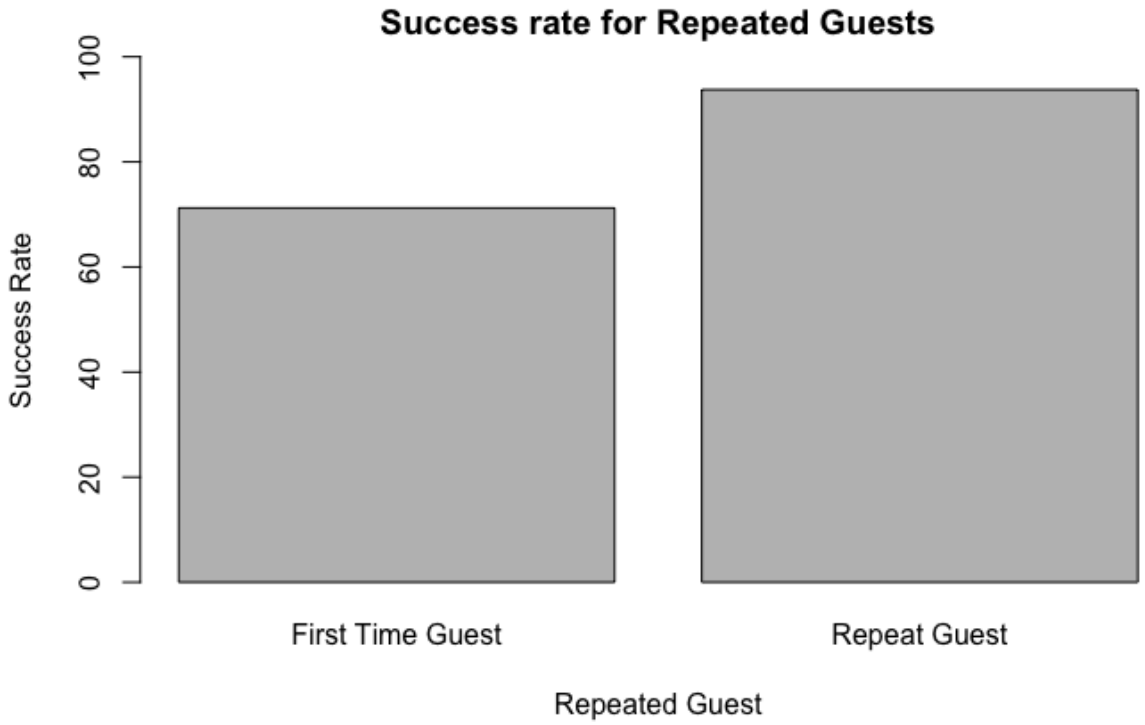
Portugal (PRT) was listed in 17,630 entries, representing 44% of our business, and had a cancellation rate of 42.18%. Great Britain/UK (GRB) was listed in 6,814 entries, representing 17% of our business, and had a cancellation rate of 13.08%. From our analysis, we estimate about 90% of our business comes from western European countries.

Market Segment:
Market Segment Rate has a mean success rate of 77.01 and a median of 84.18 . The range is from 57.61 to 86.52.



Analysis: Groups and Online perform slightly worse than average.

Is Repeated Guest:

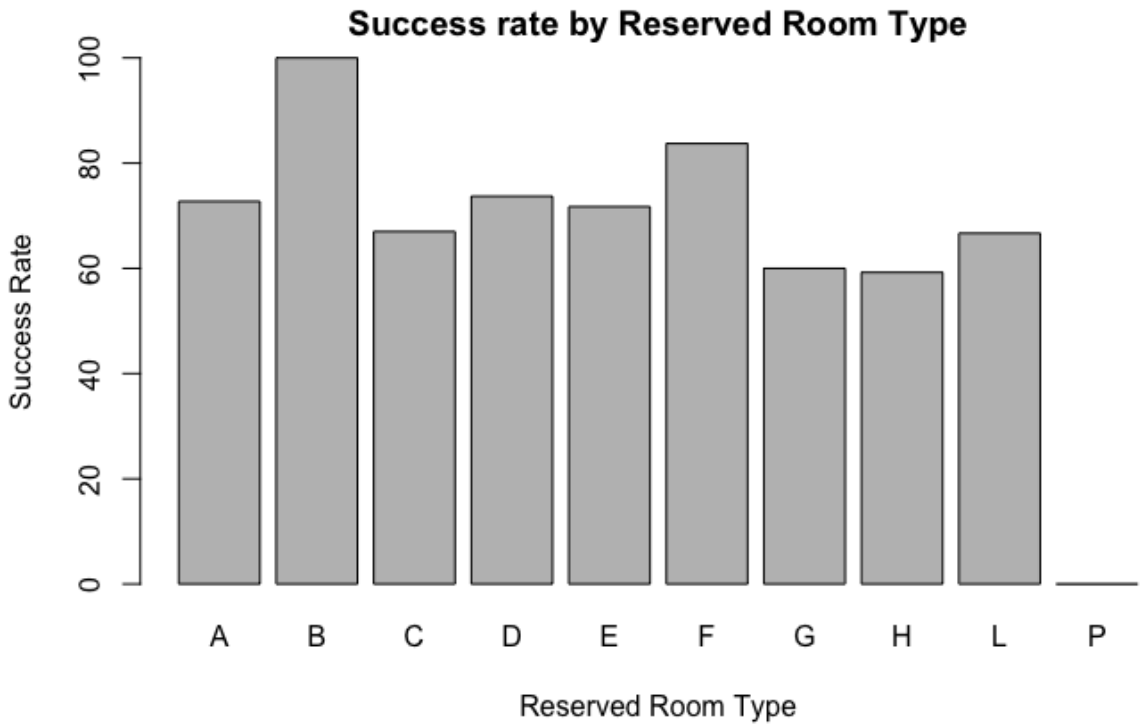


Analysis: The repeat guest has a significantly higher success rate. The success rate of the first time guest is 71.2%, while the repeat guest is 93.8%.

Previous Cancellations:
Previous Cancellations has a mean success rate of 0.10 and a median of 0 . The range is from 0 to 26. Only 1095 bookings have a prior cancellation, representing about 2.5% of the total bookings. Of these with prior cancellations, 924 entries cancelled, and just 171 did not cancel. Due to this small variation, we are opting not to generate a graphical representation of this data.

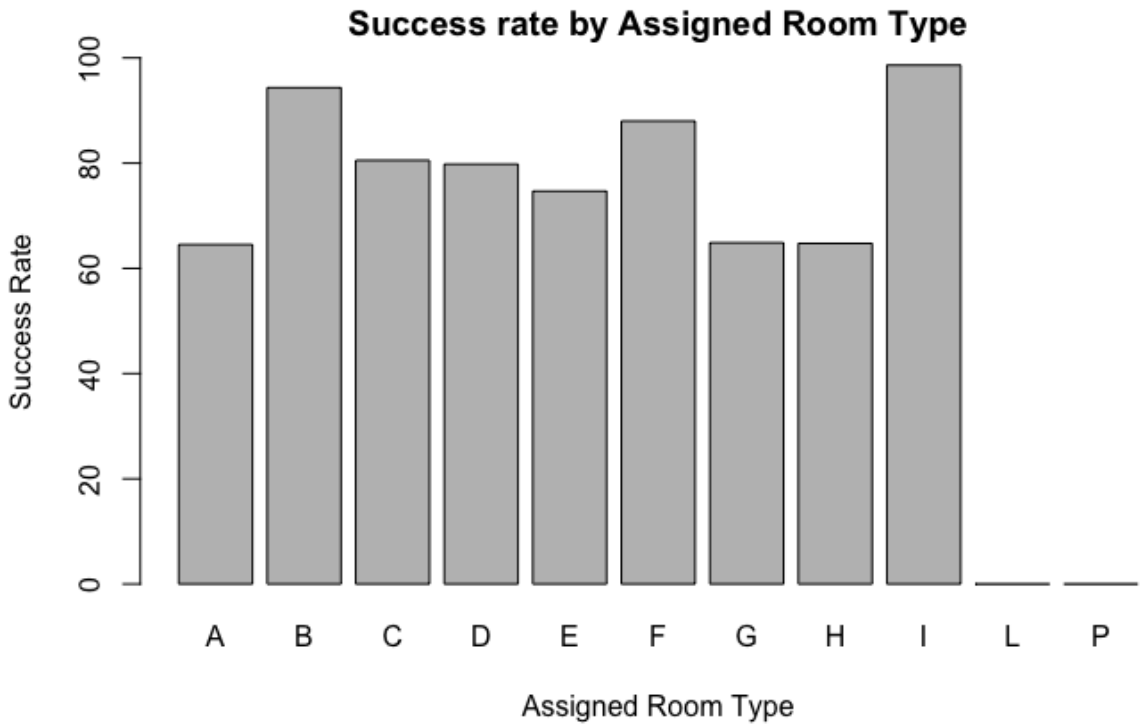
Previous Bookings Not Cancelled:
Previous Bookings Not Cancelled has a mean success rate of 0.15 and a median of 0 . The range is from 0 to 30. Only 2032 bookings had prior cancellations, representing about 5% of the total bookings. Of these with prior cancellations, 1953 entries did not cancel, and just 79 did cancel. Due to this small variation, we are opting not to generate a graphical representation of this data.

Reserved Room Type:
Reserved Room Type Rate has a mean success rate of 65.48 and a median of 69.34 . The range is from 0 to 100.



Analysis: Most reserved room types perform about equally well. Type B performed quite well, at 100%, while type P performed horribly at 0%. However, we should note that B only accounts for 10 bookings and P only accounts for 2 bookings. We consider these to be statistically insignificant.

Assigned Room Type:
Assigned Room Type Rate has a mean success rate of 64.55 and a median of 74.67 . The range is from 0 to 98.62.



Analysis: Assigned room types B and I perform slightly better than average. While L and P performed horribly, these room types only account for 1 and 2 bookings respectively. We consider this to be statistically insignificant.

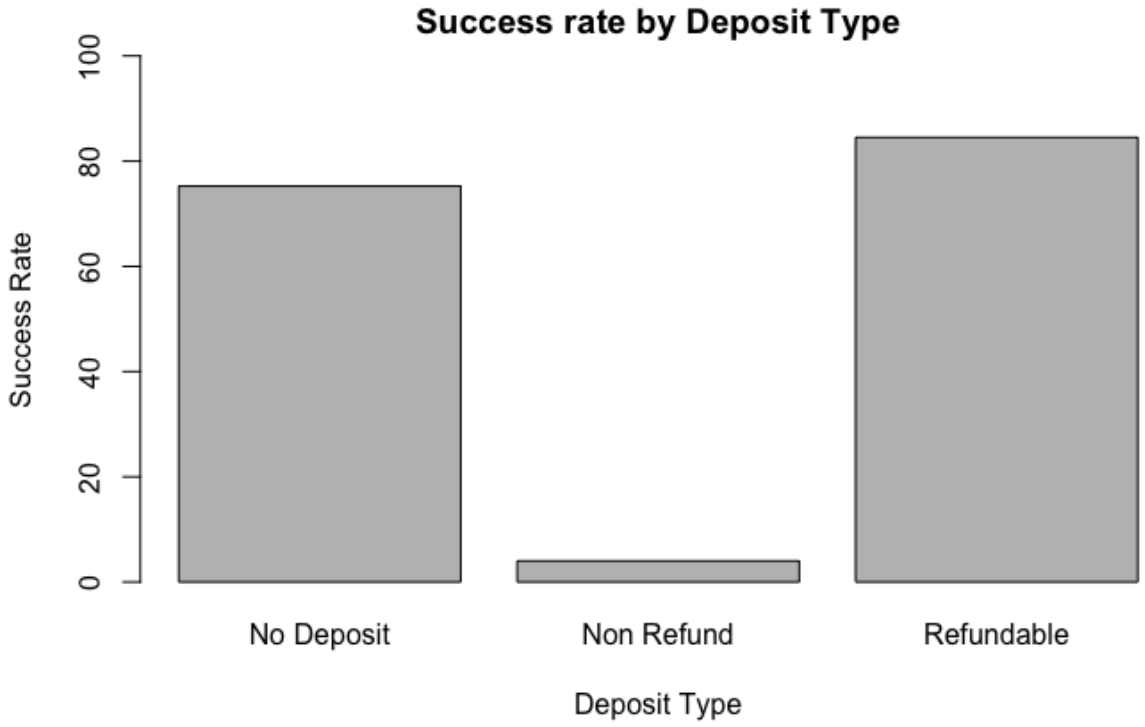
Room Was Changed:
The room was changed for 19.3% of bookings. When the room was not changed, approximately 33% of bookings cancelled. When the room was changed, approximately 5% of booking cancelled.

Booking Changes:
Booking Changes Rate has a mean of 80.06 and a median of 85.13. The range is from 0 to 100.

Analysis: We see no direct correlation between success rate and the number of booking changes. While 16 booking changes were all associated with cancellations, this only occurred in 1 booking. We consider this to be statistically insignificant.

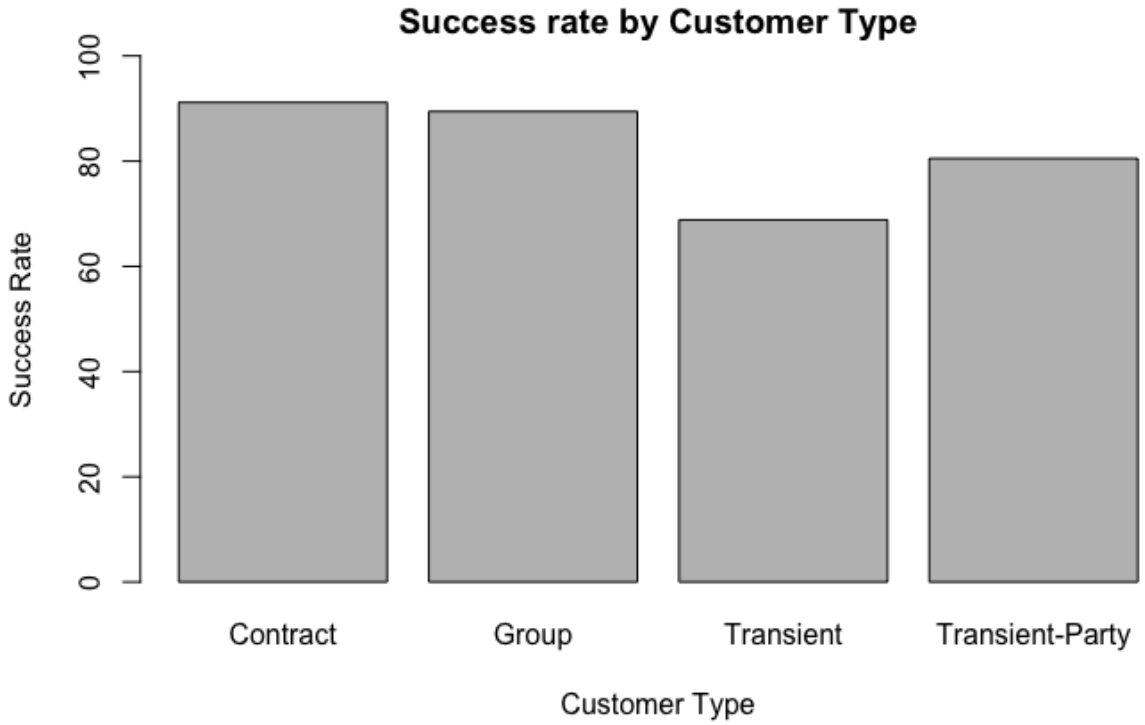
Deposit Type:

Deposit Type Rate has a mean of 54.59 and a median of 75.26. The range is from 4.01 to 84.51.



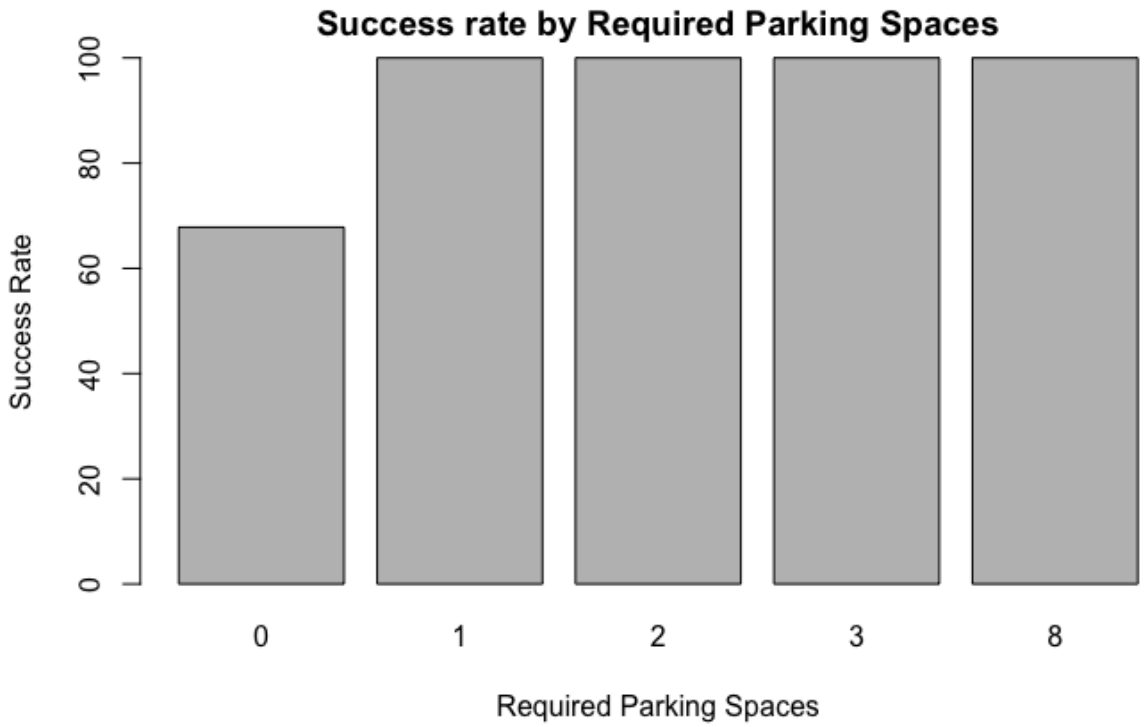
Analysis: We see a large number of cancellations when a non refundable deposit was used. Only about 4% of reservations were not cancelled. The number of bookings made with a non-refundable deposit was fairly small, representing only 4% of our business.

Customer Type:
Customer Type has a mean of 82.48 and a median of 84.97 . The range is from 68.83 to 91.16.



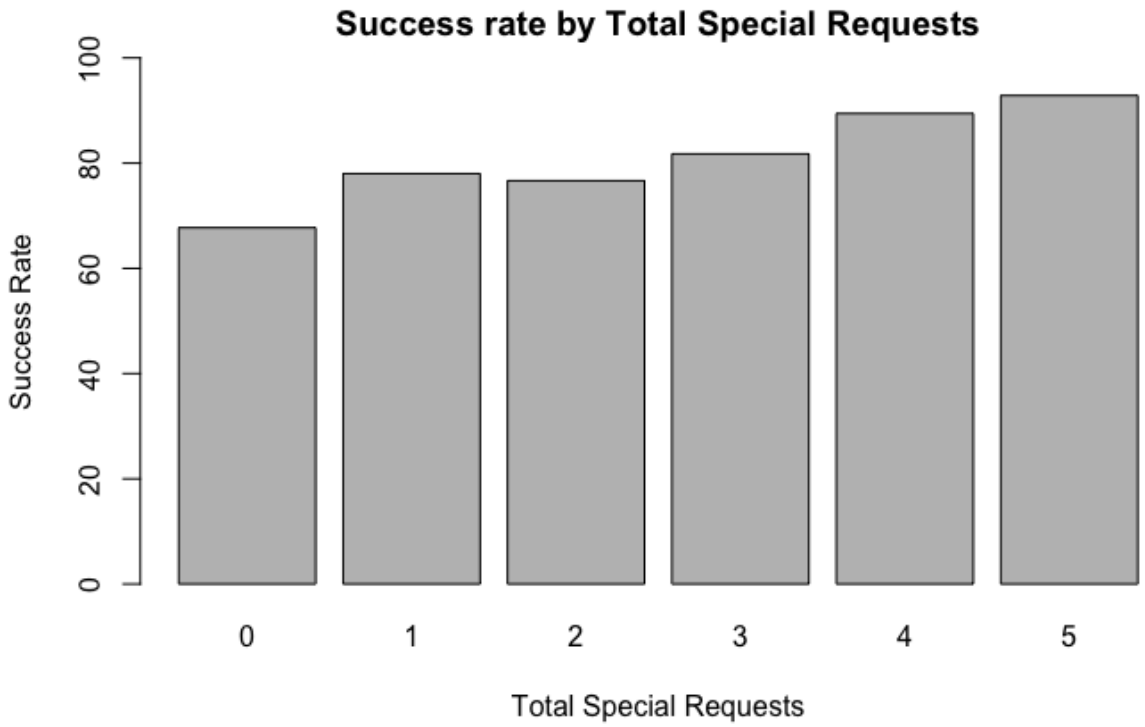
Analysis: Transient performs slightly worse than the other customer types.

Required Car Parking Spaces:
Required Parking Spaces has a mean of 93.57 and a median of 100 . The range is from 67.83 to 100.



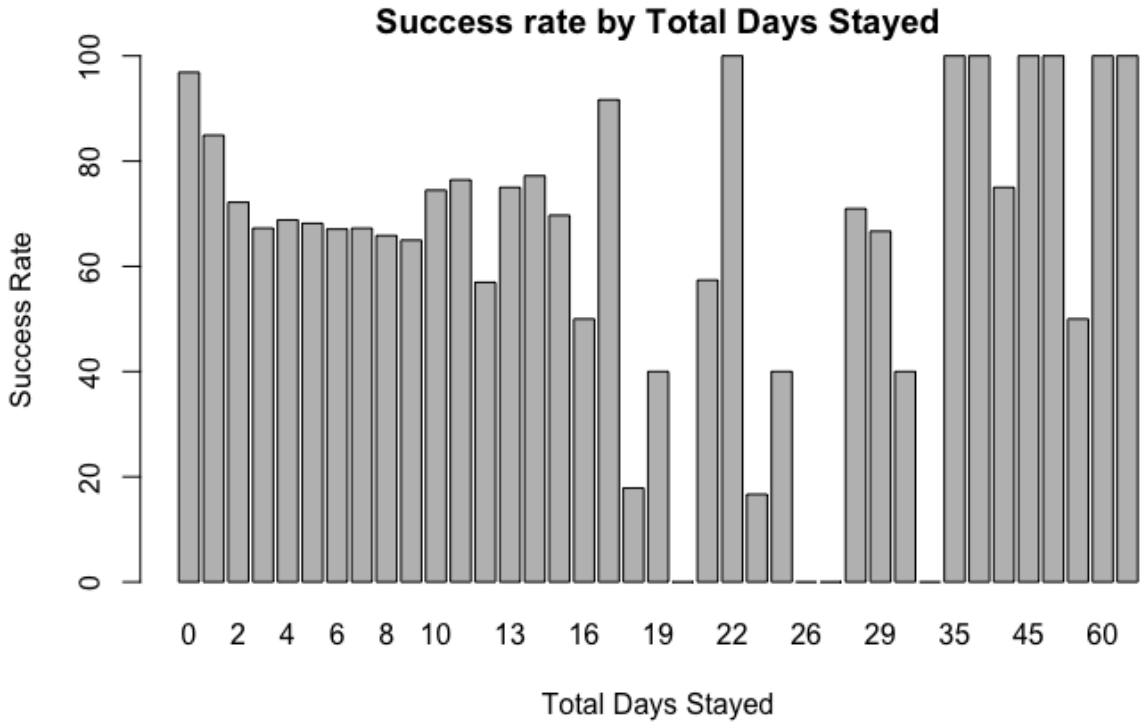
Analysis: We have a 100% success rate whenever a parking space is requested, as opposed to the 67% success when no parking spaces are requested. The number of booking requesting parking spaces represents 13.7% of our business.

Total of Special Requests:
Total Special Requests has a mean of 81.07 and a median of 79.88 . The range is from 67.73 to 92.86.



Analysis: We see an upward trend with special requests. As the customer makes more requests, they are more likely to show up.

Total Days Stayed:
Total Special Requests has a mean of 63.32 and a median of 68.17. The range is from 0 to 100.



Analysis: As mentioned above, we consider a value of 0 for total days stayed to be invalid. Our analysis shows a troff of success, where success is highest when under 16 days and over 35 days. This is not consistently linear, as there is high variability. However, we should note the number of visitors staying for extended periods of time are small. For total days stayed between 15 and 69 days, we have no more than 56 bookings representing any given column.

Long Weekend:
Long weekends only account for about 7.86% of our business. Non-long weekend business cancels at a 27.2% rate, while long weekenders cancel at a rate of 34.43%.

Extended Stays:
Extended stays only account for 8.01% of our business. Non-extended stays cancel at a 27.65% rate, while extended stays cancel at a rate of 29.01%.

Modeling

Our machine learning analysis started with an unsupervised model, in an attempt to find correlating factors with cancellations. We first created a model using apriori, with a support of at least 0.085 and confidence of 0.7. This resulted in 34 rules generated. The most frequent correlations were two adults, an origin of Portugal, room type A assigned, no booking changes and no required car spaces. This model has a confidence of around 0.71 and a lift of at least 2.5.

We then dove into finding out what correlated with bookings that did not cancel. We choose a support of 0.17 and confidence of 0.9, and produced an apriori model. This generated 26 rules. The most frequent correlations include a room change, no children or babies and no deposit. This model has a confidence of at least 0.95 and a lift of at least 1.3.

Based on this analysis, we decided to dive deeper into ways to reduce the number of cancellations from Portuguese guests. Portugal represents around 44% of all booking. Reducing the number of cancellations from Portugal by even a small percentage would yield large results, sheerly by the demographic's size.

An apriori model was created with just Portuguese guests who cancelled. We used a support of 0.1 and a confidence of 0.9. This model produced 52 rules. The most frequent correlations were a market segment of 'groups', no room changes, no parking spaces requested, no special requests and a party size of 2. This model has a confidence of at least 0.90 and a lift of at least 2.14.

Knowing what factors are present for the guests that cancel, we decided to check on the correlating factors that caused Portuguese guests to successfully retain their bookings. An apriori model was generated with just Portuguese guests who stayed. We used a support of 0.13 and a confidence of 1. This model produced 16 rules. The most frequent correlations were requiring a parking space and no deposit needed. This model has a confidence of 1 and a lift of at least 1.73.

We explored if lead time, by itself, was a significant factor in cancellations. A supervised machine learning model was constructed. 70% of the booking were used for training, the remaining 30% for testing. A ksvm model was created with the sample data. The model has an accuracy of 0.72, but the no information rate was 98%. This model was discarded.

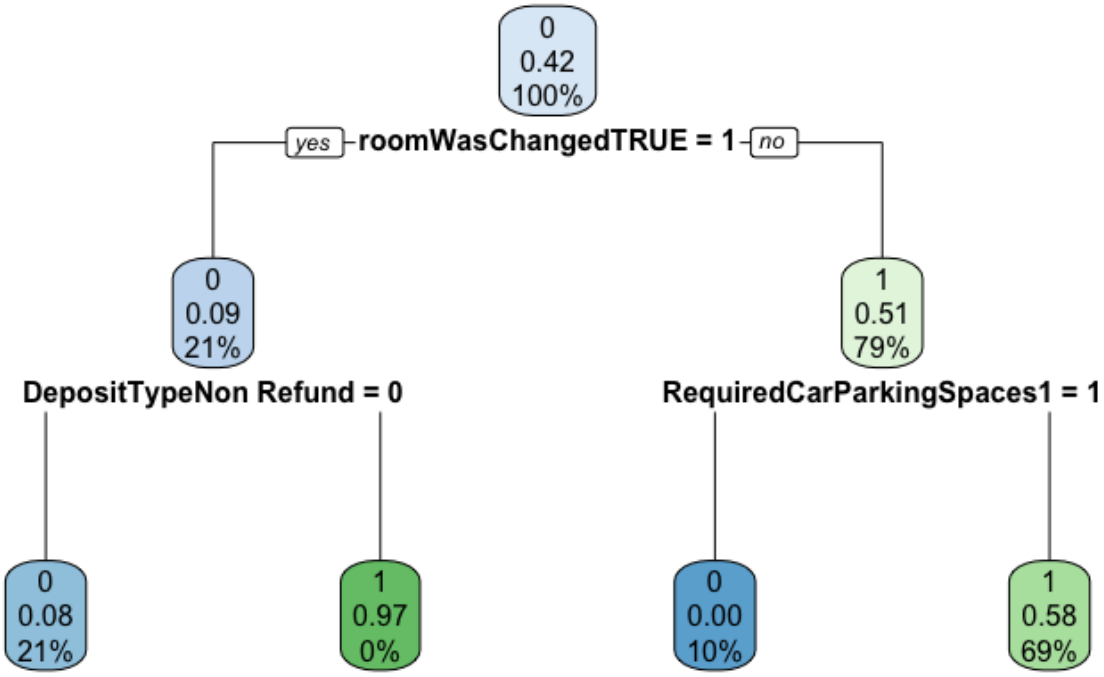
Another kvsm model was created using all variables. This model produced an accuracy rate of 88% with a no information rate of 73%. We opted to pair down the model to just country, market segment, repeated guest and required parking spaces. We again ran the model with a 70/30 split. This model had an accuracy rate of 78% and a no information rate of 73%.

We continued our supervised modeling by creating an rpart model. Using all available variables, we constructed a model that had an accuracy of 0.85 and a no information rate of 0.75, with a p-value of 2.2e-16. The top factors in the model were required parking spaces, room changes, Portuguese guests and a non refundable deposit.

Continuing the trend of investigating Portuguese travels, we repeated the model with just Portuguese guests. All variables were included. The result was a model that had an accuracy of 0.83 and a no info rate of 0.58. The p-value was less than 2e-16. The top most factors included deposit type, days stayed, required car parking spaces, market segment and room change.

We decided to narrow down the model by focusing on just required parking space, days stayed, room change, deposit type and market segment. This resulted in a model that had an accuracy of 0.78 and a no information rate of 0.53. The p-value was under 2.2e-16.

The above models were all felt to be statistically significant and have a high confidence rating. We made one last model with the three variables that pop up most often: required car parking spaces, room change and deposit type. The resulting decision tree showed the heavy favorable bias towards room changes and parking spaces.



Course of Action

After analyzing the data, we have found several correlations with cancellations. The following are our recommendations for protecting, converting and growing the business. We should note that revenue figures were not provided in our data set. As such, our analysis does not factor in the potential for growth or loss of revenue due to implementing these changes. We strongly recommend analyzing the financial impact of these recommendations before full implementation.

Convert Prospects To Customers

Hype campaign 2 to 6 months in advance of booking

We have found that the rate of cancellations trend upwards after 7 weeks and under 6 months in advance. After 6 months, the chance of cancellation becomes very likely. We recommend promotional marketing for bookings that are inside the 2 to 6 month range. This campaign would promote our hotel, activities and should generally be used to get the customer excited about their stay.

Eliminate or discourage the full board meal plan

Approximately 60% of prospects with the full board (FB) meal plan cancel. Only 4% of guests with no meal plan (SC) cancel. All other meal options see a cancellation rate of about 30%. The full board plan has nearly double the number of cancellations. If maintaining a meal plan is desired or marketable, we recommend shifting these customers over to the half board or bed and breakfast plans.

Offer a room switch

In nearly all our models, customers who switch rooms were less likely to cancel. Only 5% of customers with a switched room cancelled. We should note that our data analysis cannot determine if the room switch resulted in the customer getting a better (more expensive) room, or when in the process the switch occurred. It is conceivable the switches resulted favorably towards the customer, but at our detriment if they were occupying a more profitable room. The data analytics team would like to explore this data point further and collect additional information. We see high potential here, but the lack of transparency and price impact makes us cautious to fully recommend it without another data study.

Remove the non-refundable deposit option

The non-refundable option is a small portion of our business. It occurs in only about 4% of our bookings. However, it yields a 96% cancellation rate. Our normal cancellation rate is closer to 20%. Given this wide disparity, and the potential for customer dissatisfaction by losing their deposit, we recommend shifting these customers to a refundable deposit option. Guests with refundable deposits arrive 84.51% of the time.

Encourage parking spaces and car travel

We see tremendous success when a customer requests a parking space. We have a 100% success rate whenever a parking space is requested, as opposed to the 67% success when no parking spaces are requested. The number of booking requesting parking spaces represents 13.7% of our business. Given the sheer success, we recommend encouraging customers to drive to our properties. When booking the room, we can partner with a car rental firm. Bundle discounts may be offered, or just the convenience of booking a car rental at the same time. We need to ensure the properties have sufficient parking space to handle increases in car travel. We should also evaluate the feasibility of installing electric car charging stations to our guests, given the prevalence of European visitors.

Protect The Business

Target British visitors

Great Britain is our second largest source of guests. These visitors have a cancellation rate of just 13%, which is below the average of 20%. Given the volume of visitors, we believe the British represents a reliable source of income. While we do foresee the possibility of growing the business by targeting non-western European visitors, we also acknowledge that recent market volatility in the hospitality industry makes it prudent to have a strong base to fall back upon.

Encourage repeated guests

The success rate of the first time guest is 71.2%, while the repeat guest is 93.8%. Given the high success rate of people returning to our hotel, we recommend a campaign to encourage repeated visitors. This could include promotional campaigns to prior guests or offering to book their next stay during checkout. For the later suggestion, we do need to note that bookings over six months in advance have a very poor success rate, so this type of program would be more applicable to frequent guests.

Encourage special requests

There is a correlation between the number of special requests and the likelihood of arrival. When no requests are made, the success rate is 67.73%. This rate steadily rises to 92.86% with 5 requests made. We recommend promoting that special requests are possible and outlining what requests could be made. We also believe it is important to coach our customer service representatives into understanding that special requests are important to the business, and should not be viewed as a burden. We want to encourage the guest to make these requests.

Grow The Business

Target more non-western European visitors

Almost 90% of our business comes from western European nations. We believe there is market potential to tap visitors from other countries and continents. China has a low number of bookings, but has a cancellation rate of less than 7%. The United States has a cancellation rate of around 15%. This is far below the average of 20%. The top 20 countries does not include any nations from Africa, the middle east or APAC (except for China) and all of Latin and South America (except for Brazil).

Offer last minute stays

Nearly 27% of our guests book their stays within the 2 weeks of arriving, with a 92% success rate. Only 8% book the day of arrival, with a 95% success rate. If we have extra inventory, any ‘day and date’ bookings are virtually guaranteed to be fulfilled. We do acknowledge that last minute bookings like this present unique business challenges, like ensuring proper staffing levels.