Star Hotels

EUGENIE SEHOLM

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- Data Manipulations
- Exploratory Data Analysis (EDA)
- Key Findings and Insights
- Model Overview and Performance Summary
- Business Recommendations

Business Problem Overview

- Core business idea:
  - Create a Machine Learning based solution to help predict which bookings are likely to be canceled
- Problem to tackle:
  - Find which factors have a high influence on booking cancelations
  - Build a predictive model that can predict which bookings are going to be canceled in advance
  - Help in formulating profitable policies for cancelations & refunds

Business Problem Overview

- Financial implications:
  - False positives (identification of cancelation when booking is not canceled)
    - can lead to overbooking
    - may create negative feelings for guests
    - would require human resources to make alternate arrangements for guests who cannot check in to overbooked rooms
    - may require vouchers or perks for guests in future
  - False negative (identification of no cancelation when booking is canceled)
    - revenue may be lost when the hotel cannot resell the room
    - may incur additional costs by increasing commissions or paying for marketing to help sell rooms
    - may need to lower prices so the hotel can resell the room, resulting in reducing profit margin
- We will maximize Recall & F1 to minimize false positives & false negatives

Business Problem Solution Approach

Logistic Regression Model

- Logistic Regression with Sklearn library
  - newton-cg solver faster for high-dimensional data
  - with/without square root transformation on continuous variables
- Logistic Regression with statsmodels library
  - remove multicollinearity using Variance Inflation Factor
  - remove variables with p-value > 0.05
- Convert coefficients to odds to interpret
- Optimal threshold using ROC-AUC curve
- Use Precision-Recall curve to try finding better threshold

Decision Tree Model

- use DecisionTreeClassfier function, default 'gini' criteria to split
  - overfitted model
  - complex
  - difficult to interpret
- pre-prune: use GridSearch to compute optimal values of hyperparameters
  - best recall/f1 scores
  - much less complex
  - can be challenging to interpret manually
- post-prune: Cost Complexity Pruning
  - choose alpha
  - not overfitted
  - easy to interpret
  - better accuracy, reacall, f1 scores than original tree on test set

Data Overview

56926 rows (observations)

18 columns (variables)

- 3 indicator
- 2 continuous
- 6 discrete numerical
- 3 categorical

No missing values

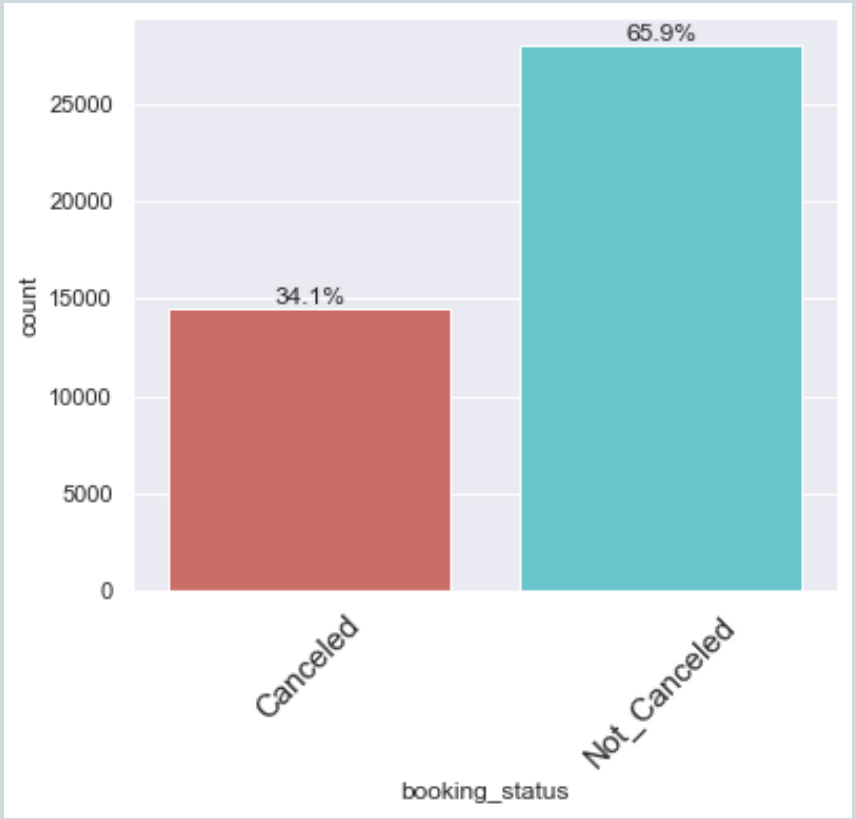
Variable	Description
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday/Sunday) guest stayed or booked to stay
no_of_week_nights	Number of weeknights (Mon-Fri) guest stayed or booked to stay
type_of_meal_plan	Type of meal plan booked by guest
required_car_parking_space	Whether or not guest requires a car parking space
room_type_reserved	Type of room reserved. Values are ciphered by Star Hotels Group
lead_time	Number of days between date of booking and 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	Repeated guest or not
no_of_previous_cancellations	Number of previous bookings canceled by customer prior to current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by customer prior to current booking
avg_price_per_room	Average price per day of reservation
no_of_special_requests	Total number of special requests made by guest
booking_status	Flag indicating if booking was canceled or not

Data Manipulations

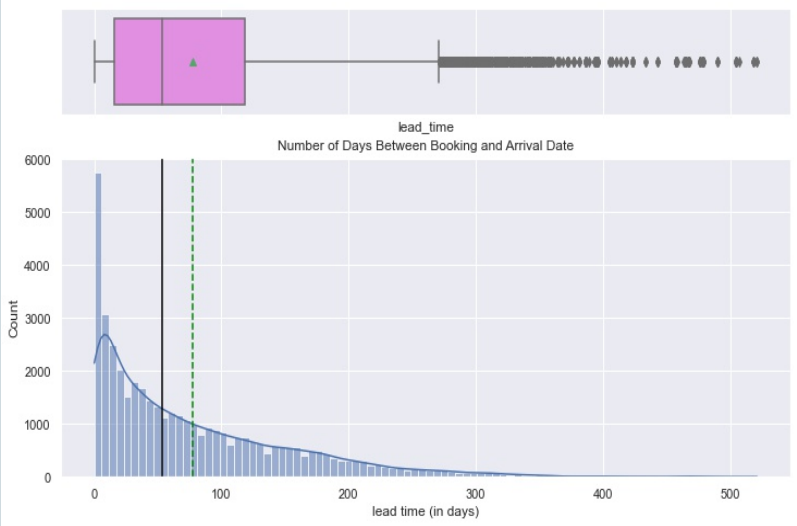
- 102 rows with 0 weekend & week nights
  - Assumed to be due to data entry error
  - Deleted
- 14347 duplicate rows
  - Assumed to be due to data entry error
  - Deleted
- Square root transformation on ***lead_time*** and ***avg_price_per_room*** to determine if it improves logistic regression

EDA: Booking Status (Target)

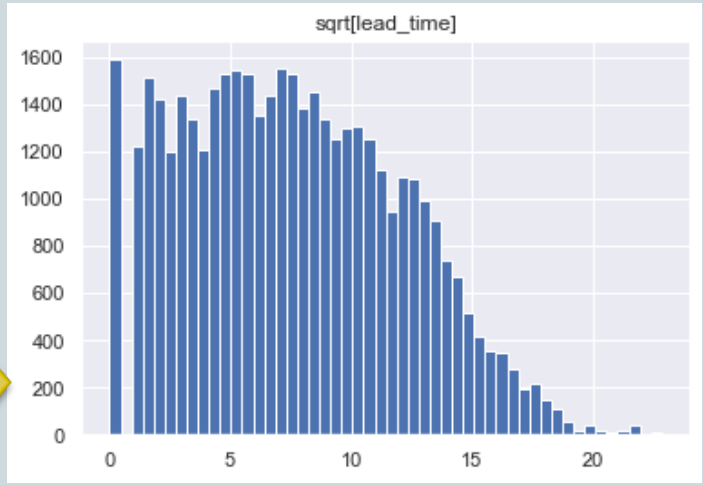
- Out of 42477 bookings, more than a third were canceled.
  - 27992 were not canceled (65.9%)
  - 14485 were canceled (34.1%)



EDA: Lead Time



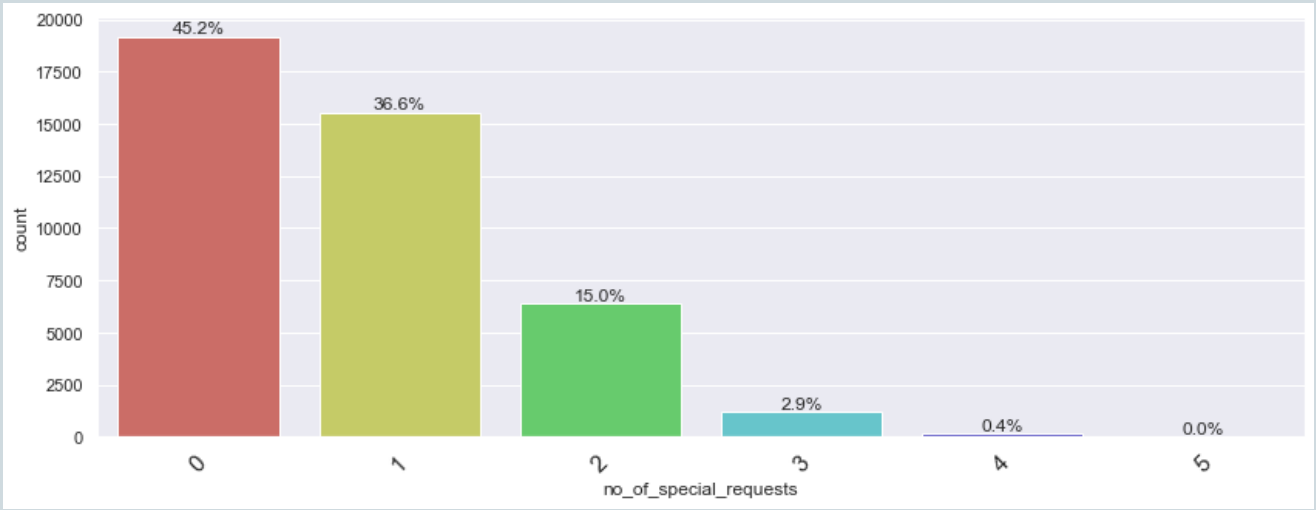
SQUARE ROOT TRANSFORMATION



- Heavily skewed in the positive direction
- More than half of the bookings were booked within 2-3 months of arrival, most likely because they may not know their plans far out in advance
- Some people book more than a year ahead of arrival.
- Square root transformation makes distribution less skewed, fewer outliers

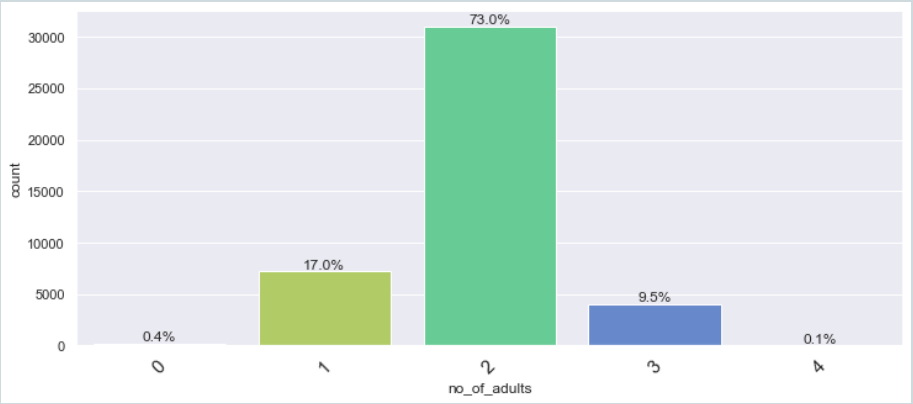
EDA: Number of Special Requests

- 45.2% bookings have no special requests
- 36.6% bookings have 1 special request

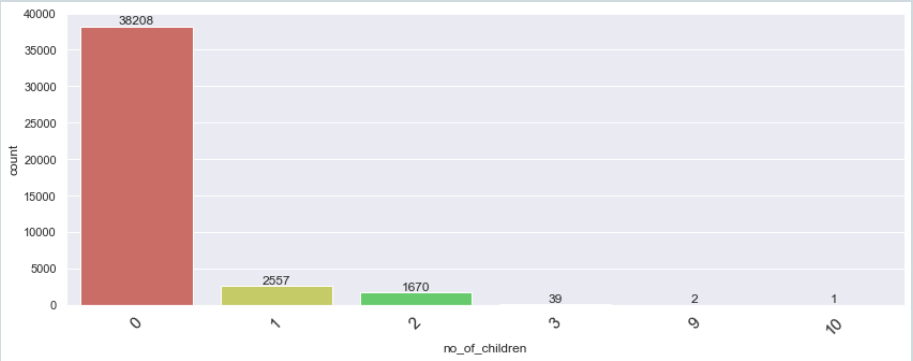


EDA: Number of Guests (Adults & Children)

- Adults
  - Majority of bookings for 2 or 1 adults

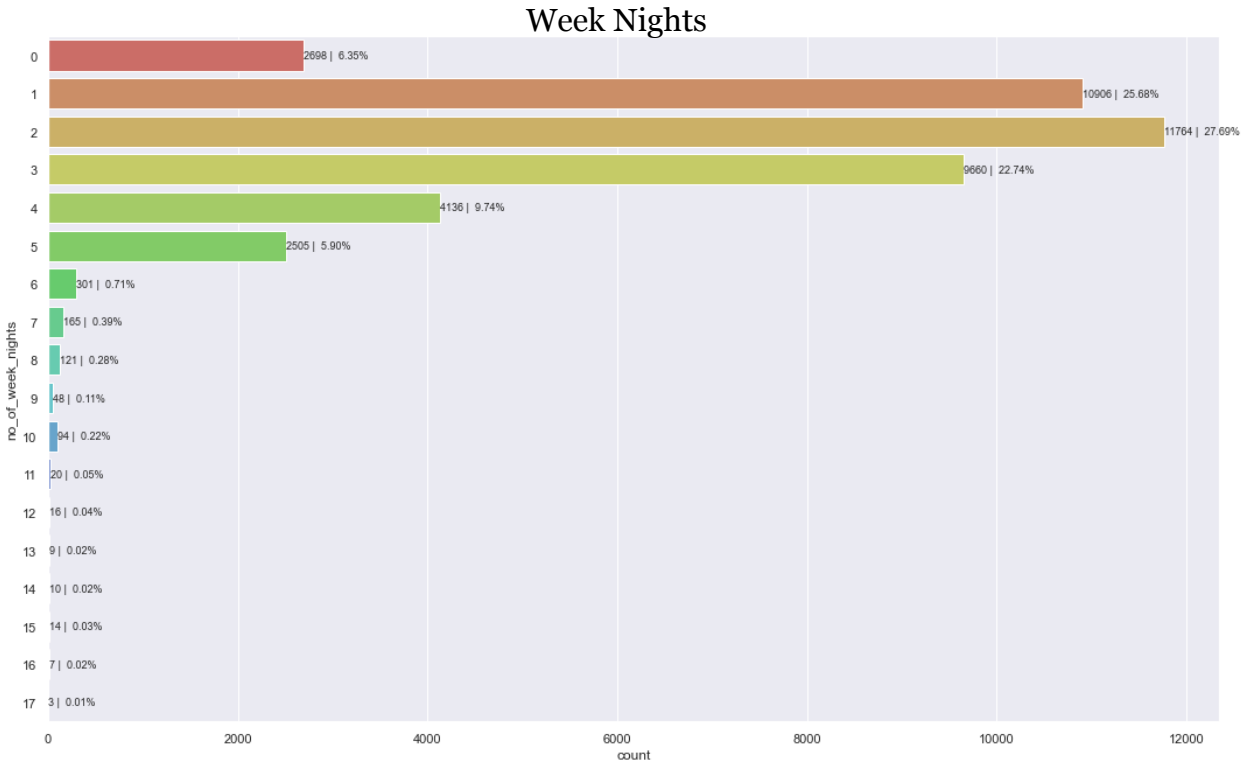


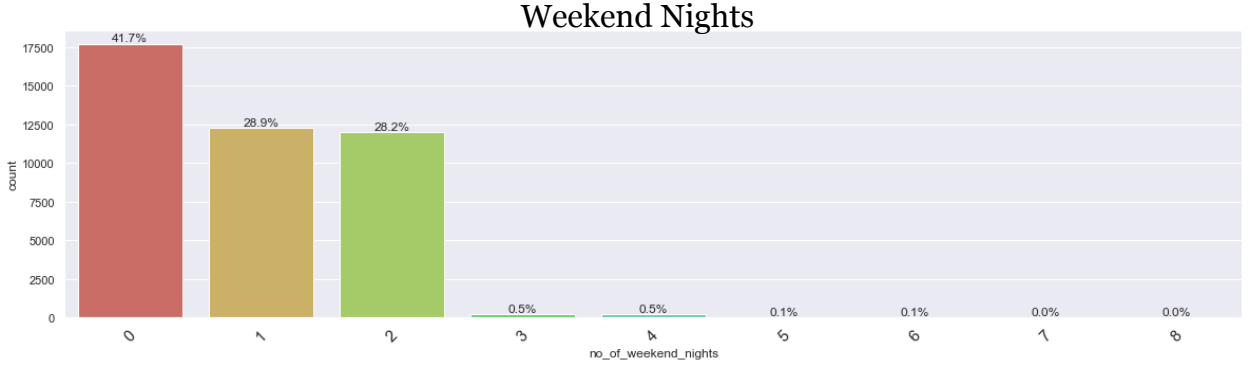
- Children
  - Majority of bookings for 0 children
  - When children are booked, it is usually 1 child
  - Beyond 3 children, few bookings for larger groups involving children
  - Bookings with children have more cancellations than those without



EDA: Nights (Weekend & Weeknights)

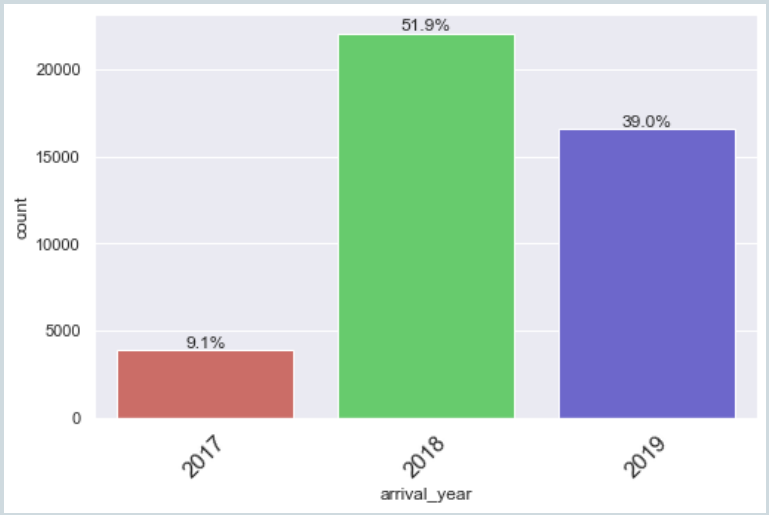
- 45.1% of bookings have no weekend nights
- ~99% of bookings are one week or shorter
- Most common bookings involve 2 weeknights and/or no weekend nights
- Very few people stay more than half a week

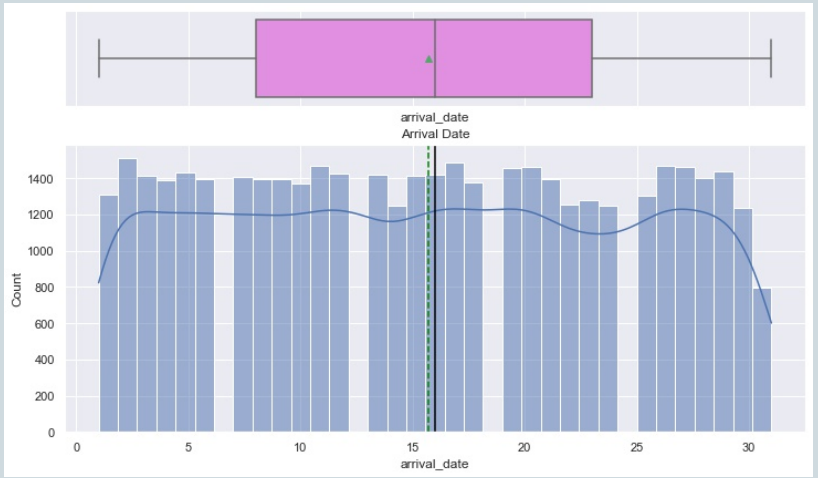




EDA: Arrival Year & Date

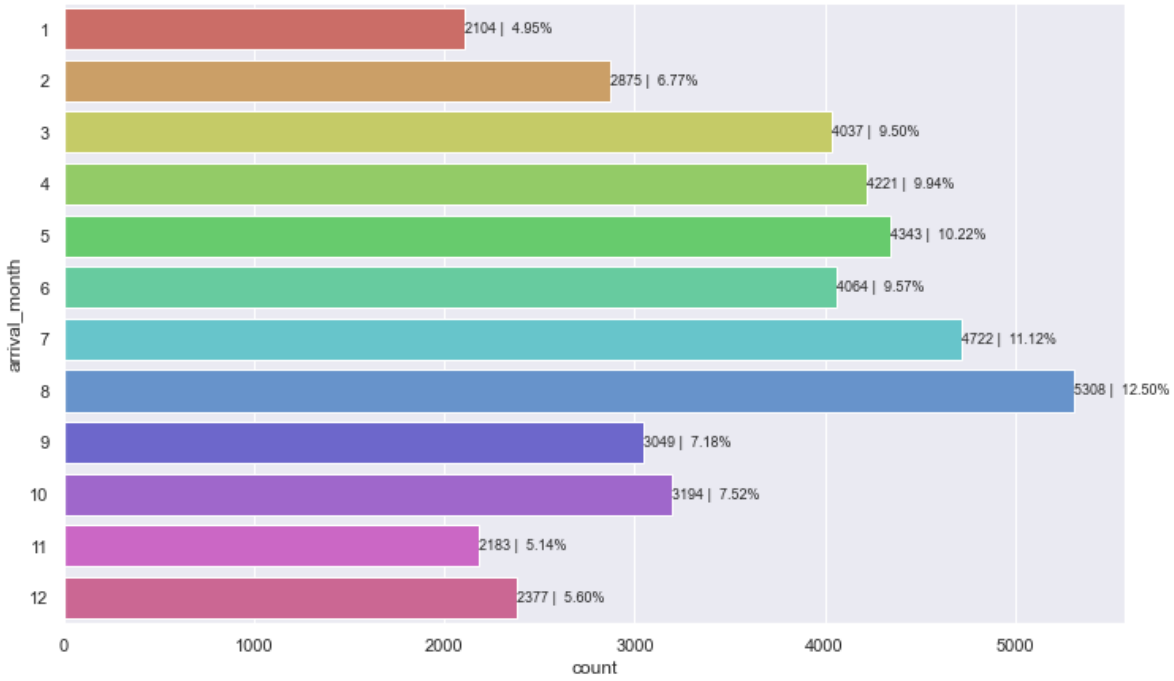
- 51.9% of data is for bookings for 2018
- 39% of data is for bookings for first half of 2019
- 9.1% of data is for bookings for second half of 2017
- arrival date is approximately uniformly distributed
- fewer bookings on 31st as there are only about half the number of months with 31 days





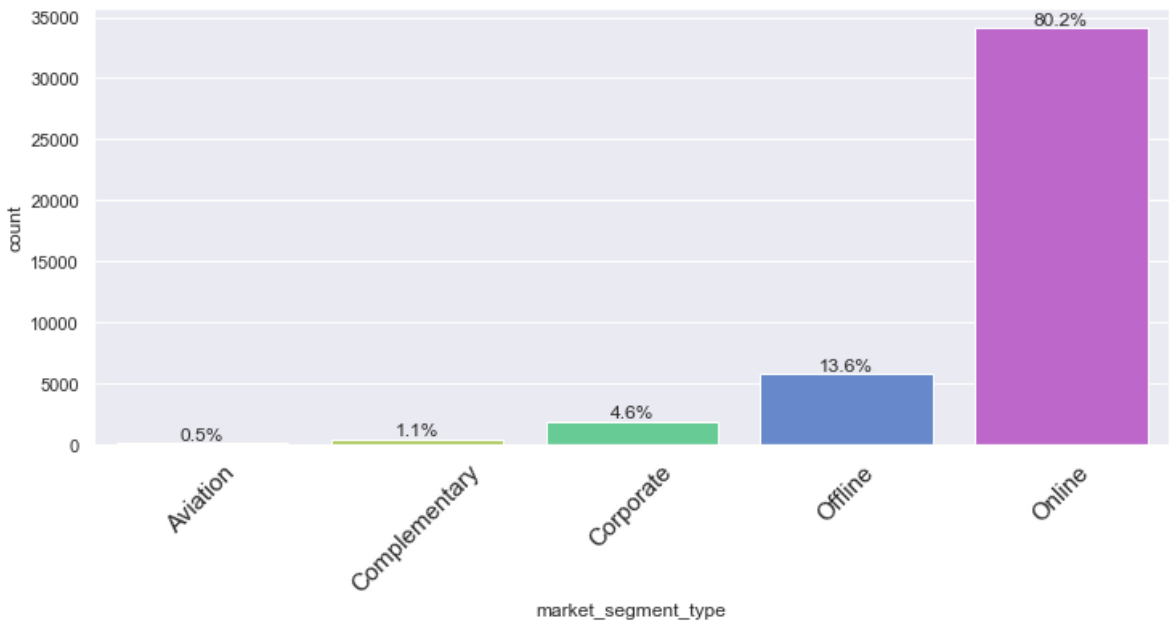
EDA: Arrival Month

- Bookings start to pick up in the spring
- Most bookings in summer months
- Fewer bookings in late fall/early winter



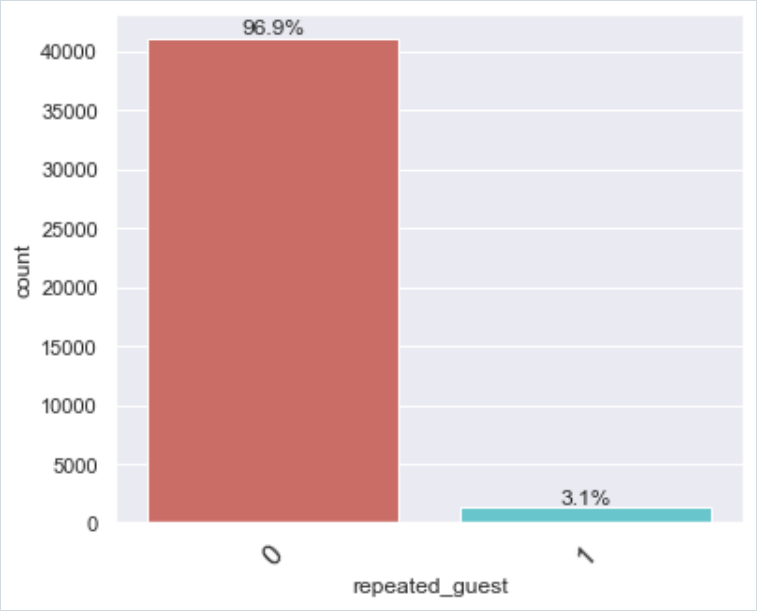
EDA: Market Segment Type

- Most (80.2%) bookings are from online segment
- Very few from aviation
- 1.1% complimentary

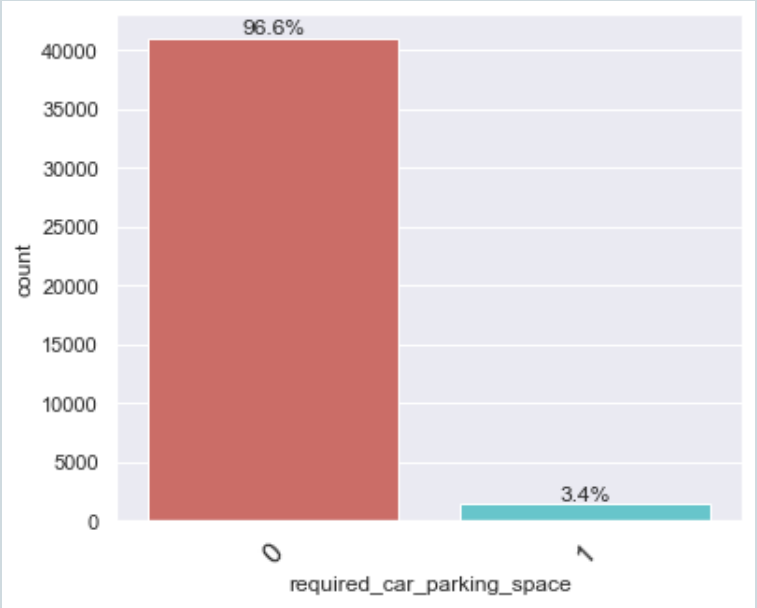


EDA: Repeated Guest & Car Parking Space

- Large majority (~97%) not repeat guests
- Very few (~3%) returning guests



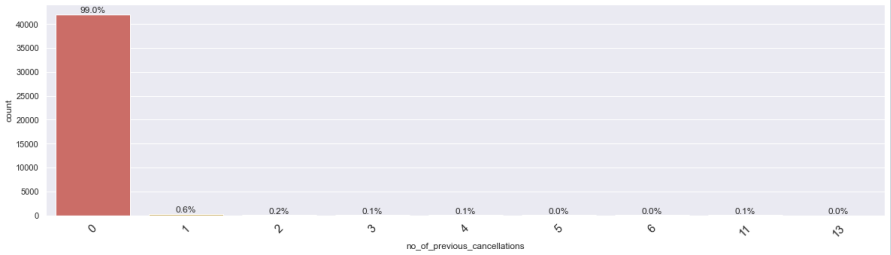
- Large majority (~97%) do not require parking
- Very few (~3%) require parking



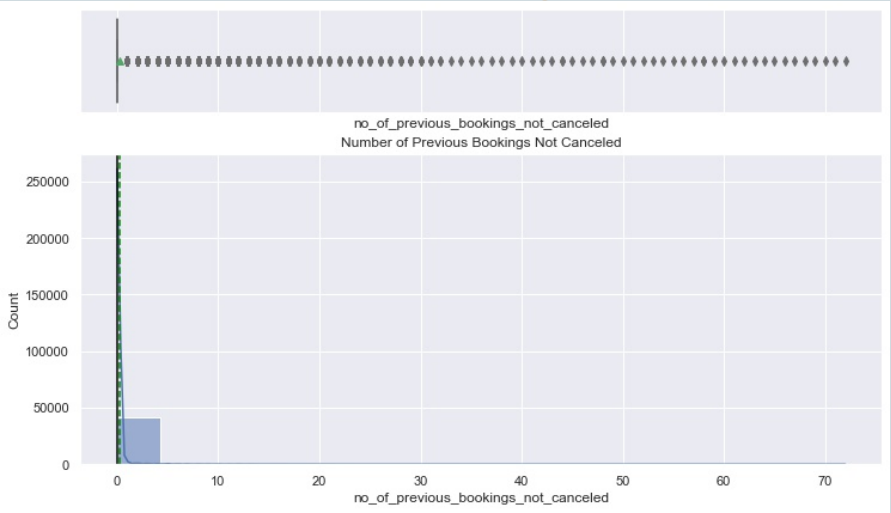
EDA: Previously Cancelled/Not Cancelled

- It is fairly rare for a booking with someone with a previous cancellation
- There are some with as many as 13 previous cancellations
- Most of the bookings were not from repeat guests
- There are some bookings by guests with 70 previous bookings not cancelled

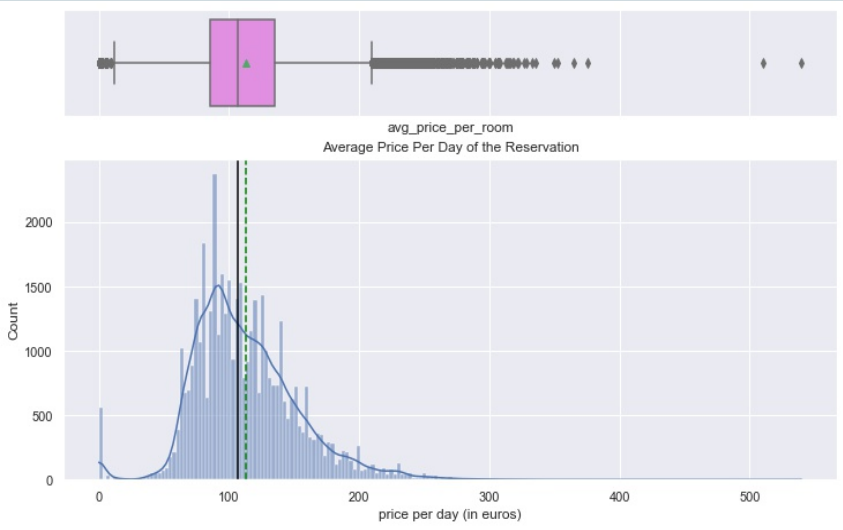
Number of Previous Cancellations



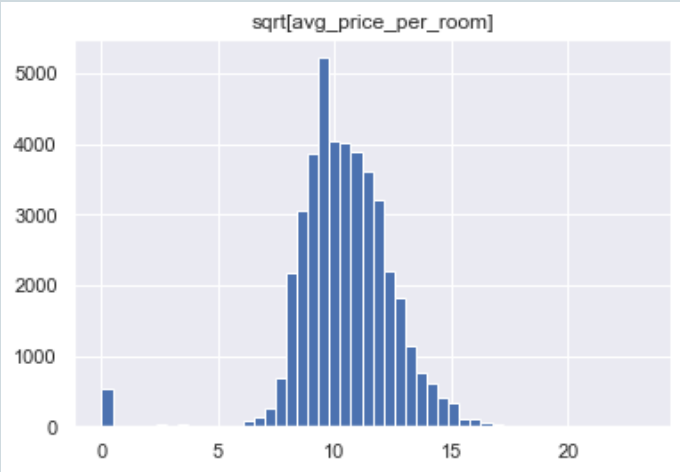
Number of Previous Bookings Not Cancelled



EDA: Average Price Per Room of Reservation

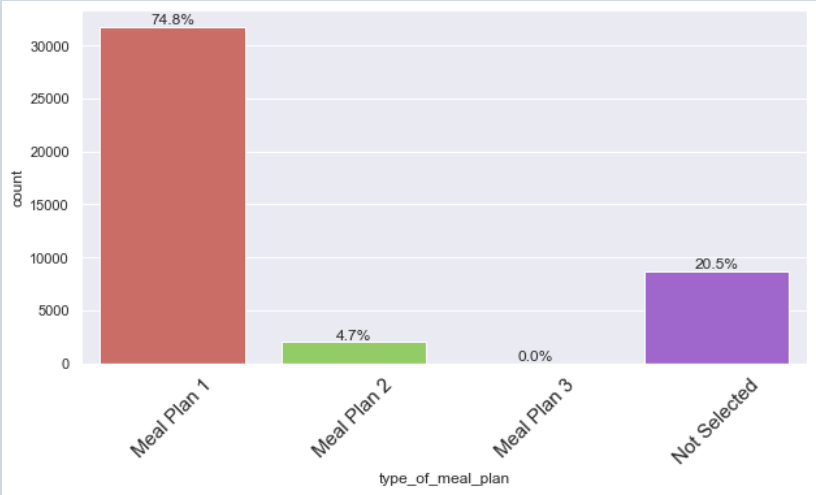


SQUARE ROOT TRANSFORMATION



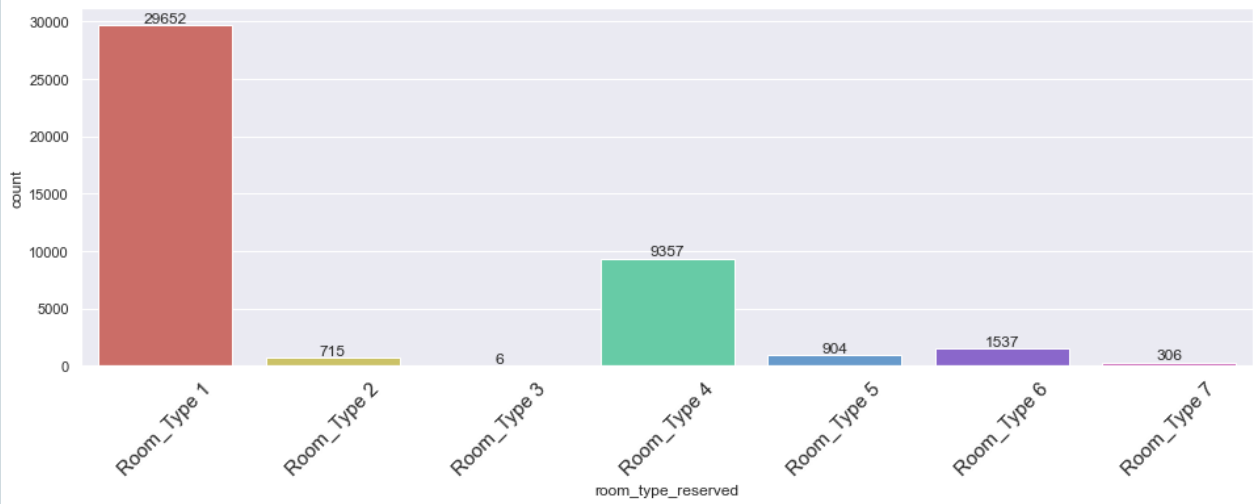
- There are many outliers in both directions
- Slight positive skew
- Notable number of free rooms
- Some rooms are 5 times the price of the median/ mean
- Square root transformation makes distribution less skewed, fewer outliers

EDA: Meal Plan & Room Type



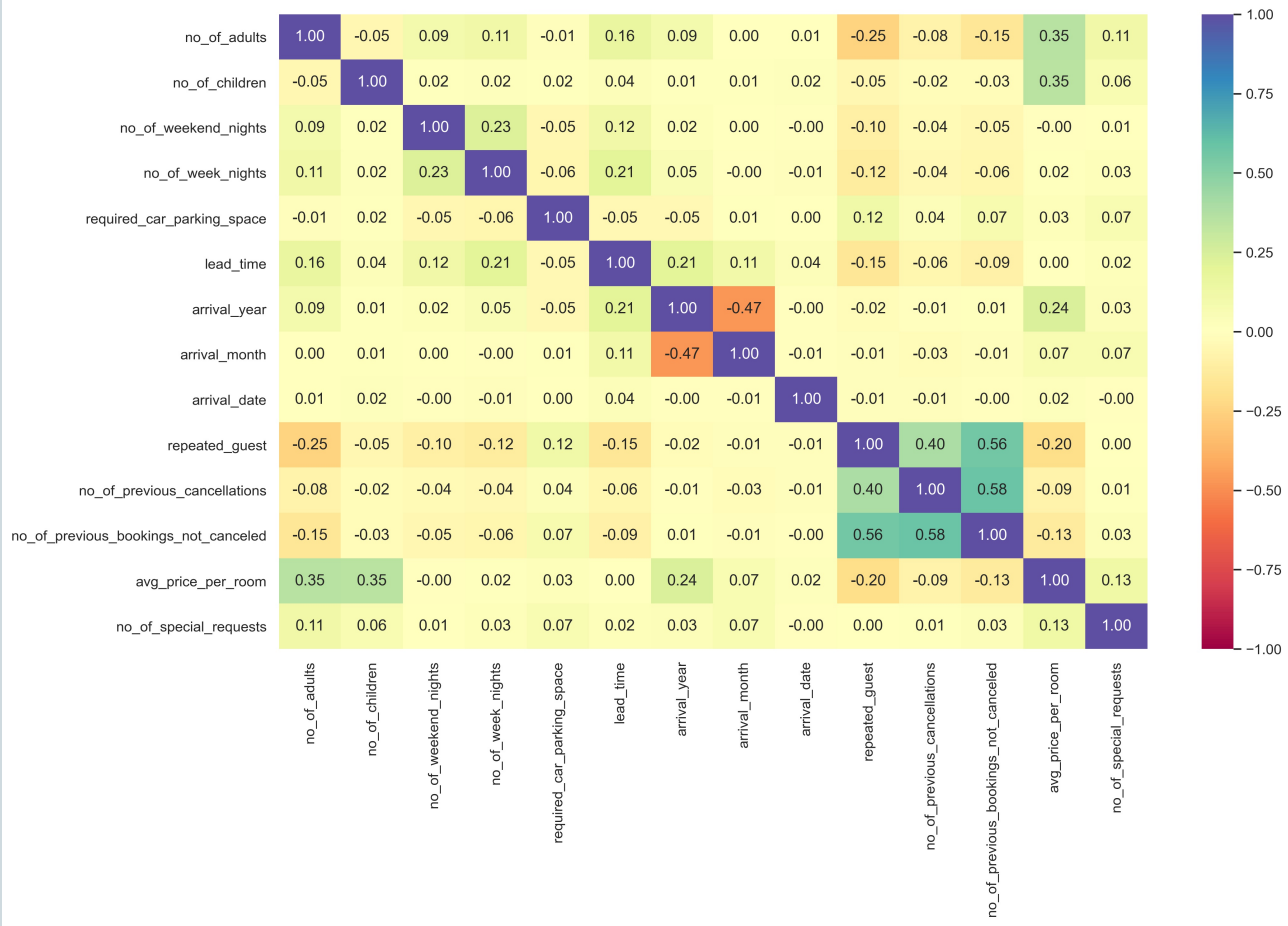
- Meal plan 1 is most commonly selected (~75%)
- No selection for meal plan ~21%

- Room type 1 is most common
- Room type 4 is second most common
- Room type 3 rarely selected

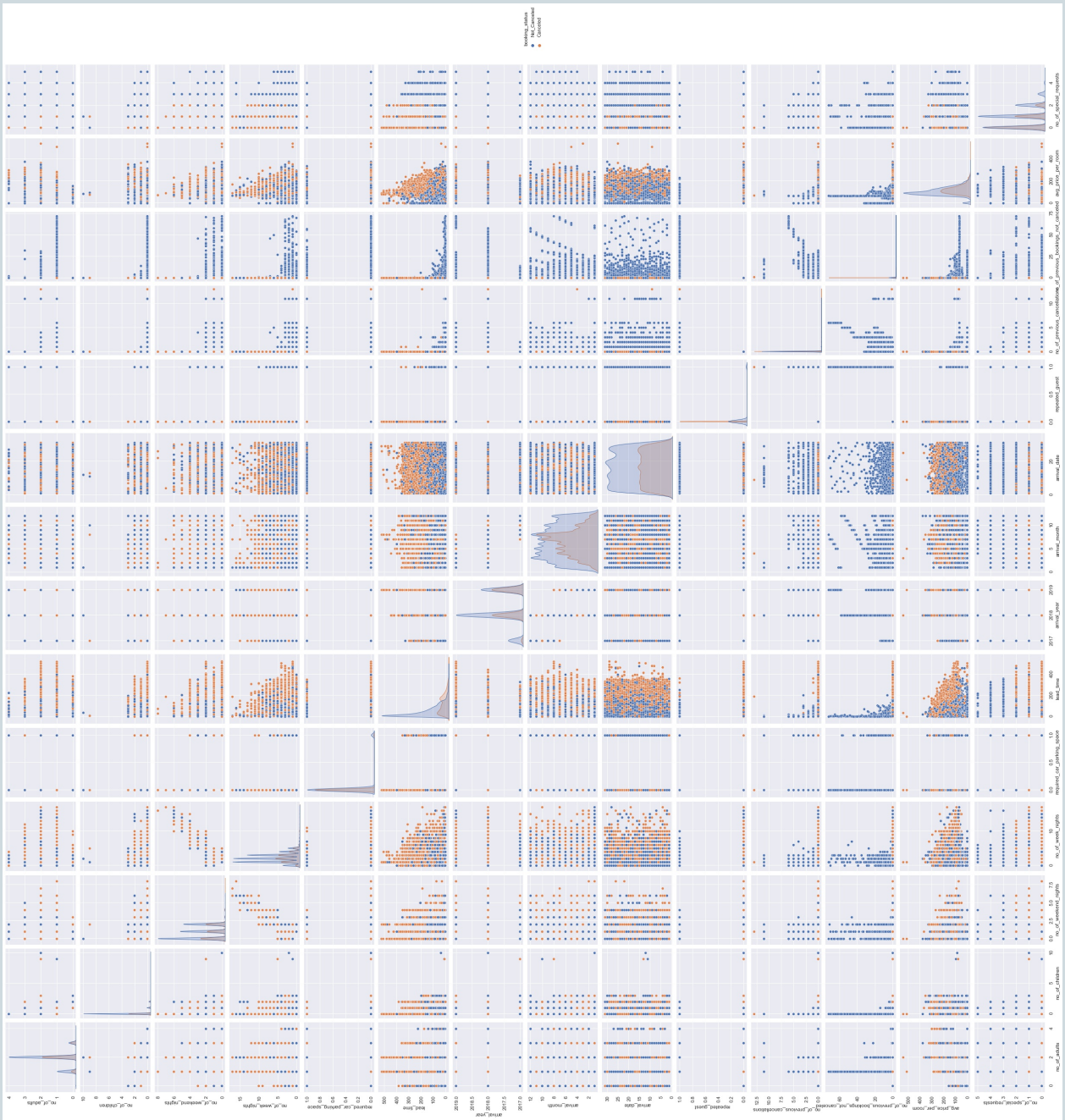


EDA: Correlation Heatmap

- No strong correlations
- Weak to moderate positive correlation between variables involving return guests & previous bookings/stays
- Weak positive correlation between average price per room and number of guests



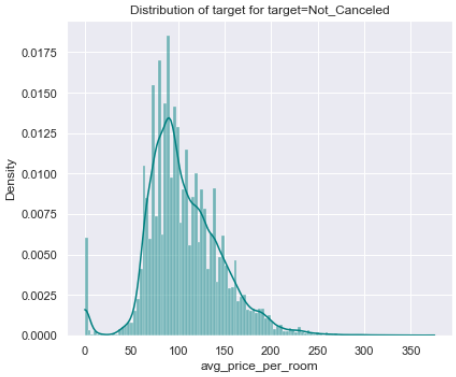
EDA: Pairwise relationship between numerical variables: cancelled/not cancelled

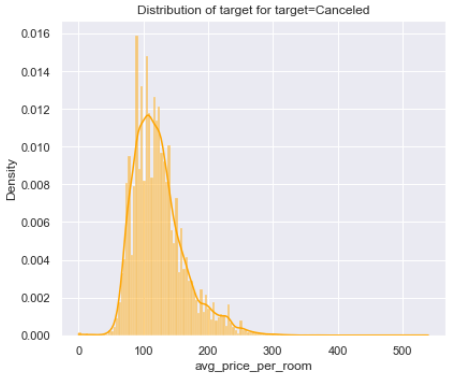


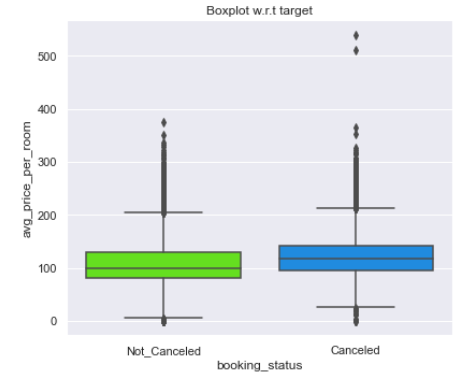
May be more cancellations with longer lead time/longer stays, except for bookings with many special requests. Repeat guests tend not to cancel

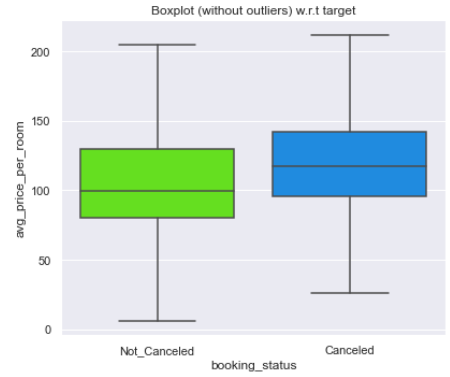
EDA: Booking Status vs Lead Time/Price

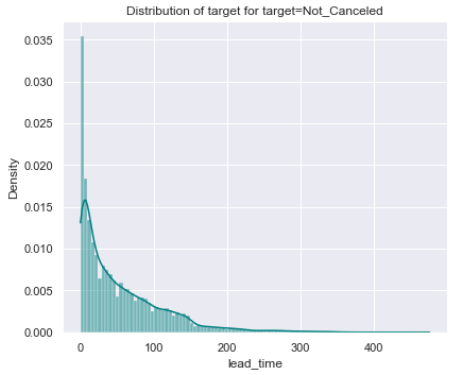
- Cancelled bookings had longer lead time and higher average price per room

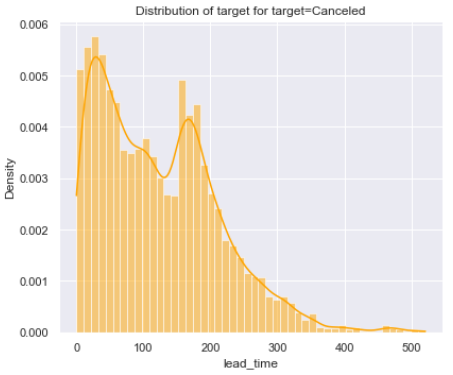


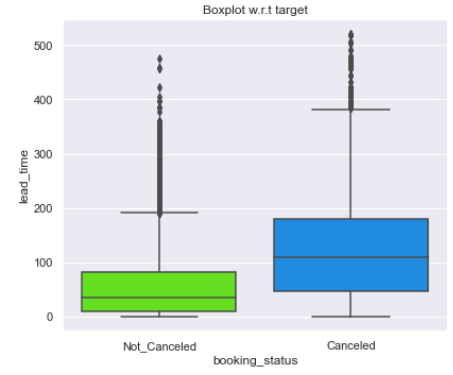


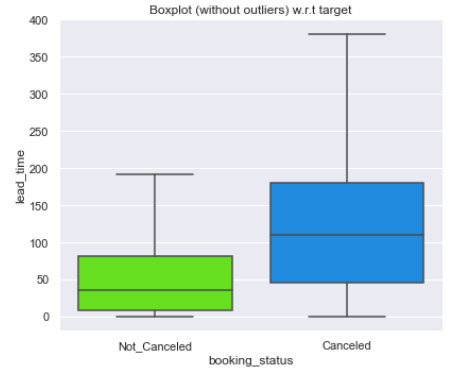






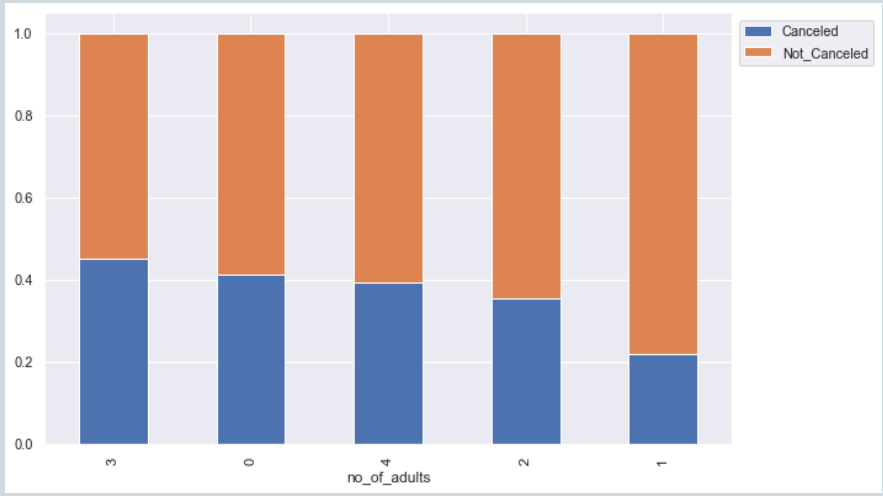


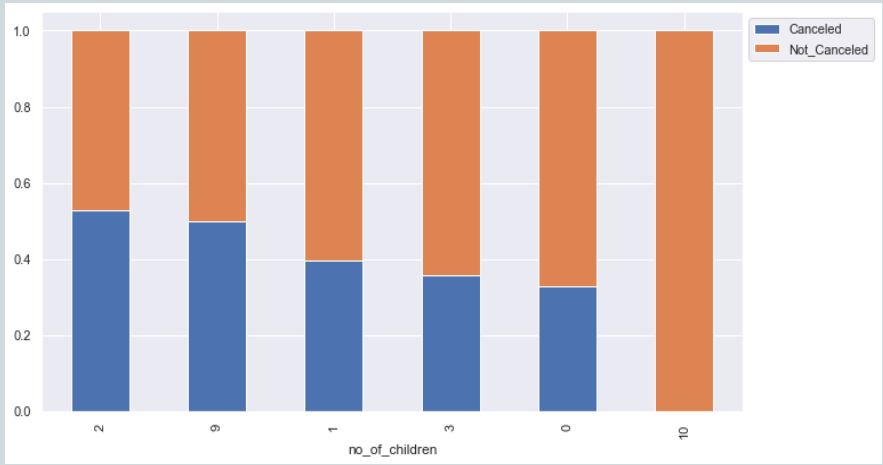




EDA: Number of Guests vs Status

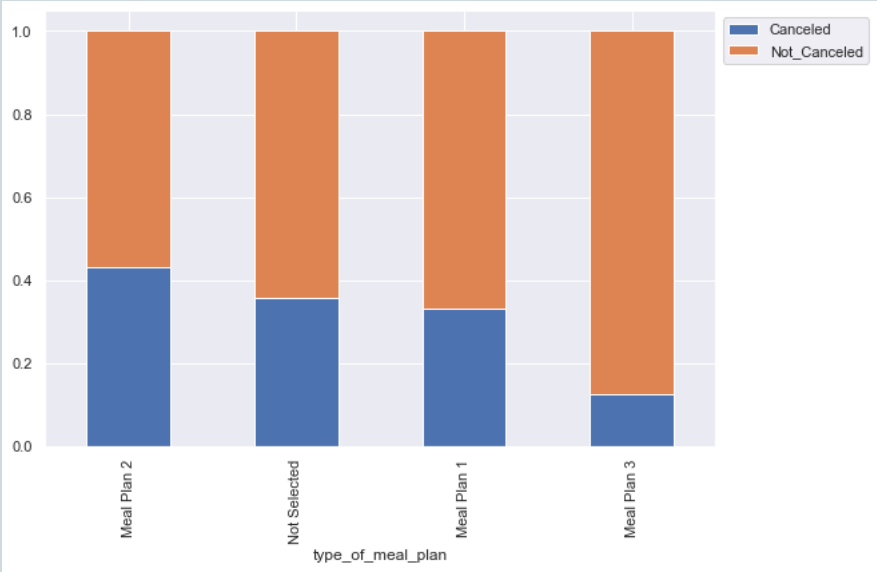
- Adults
  - Bookings with 3 adults have highest proportion of cancellations
  - Bookings with 0 adults also have high proportion of cancellations
  - However, recall most reservations are for 1 or 2 adults
- Children
  - Bookings with children have higher proportion of cancellation than bookings with no children
  - Only one case of booking with 10 children

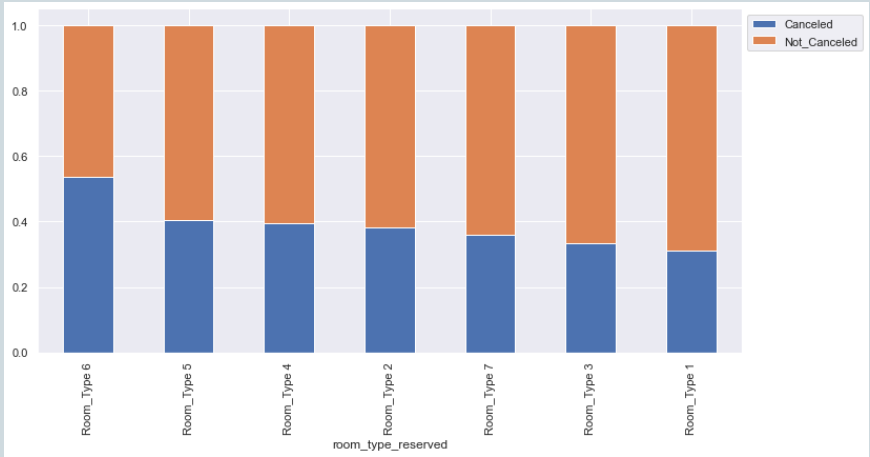




EDA: Status vs Meal/Room

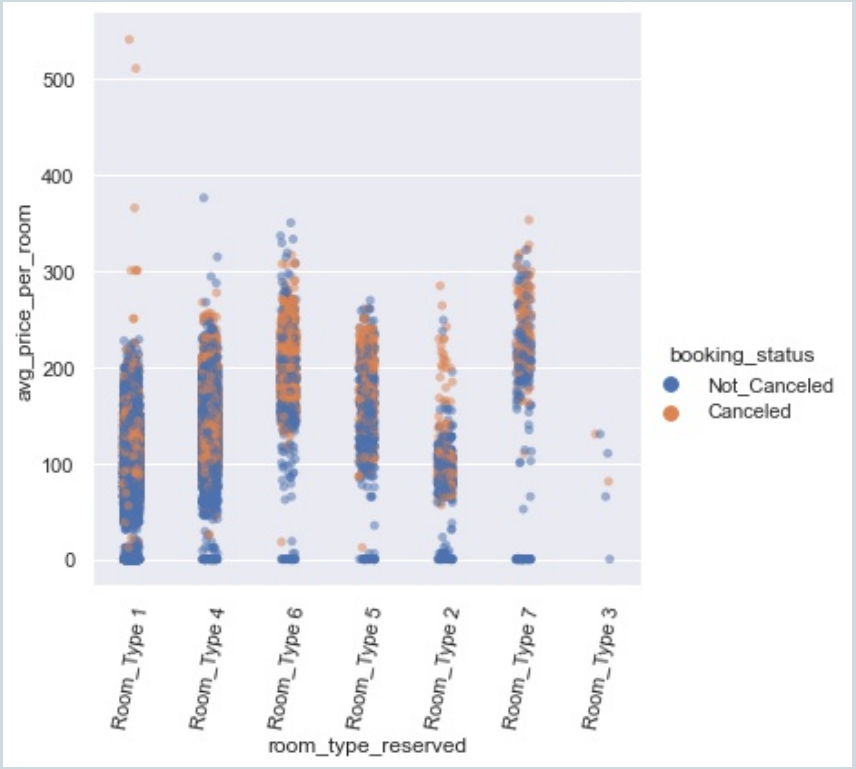
- Those who selected meal plan 2 and who did not select a meal plan had high proportion of cancellation





- Room types 1, 3, 7 have lowest cancellation but types 3 & 7 are not popular
- Room types 4, 6, 5 are booked frequently and have high proportion of cancellation

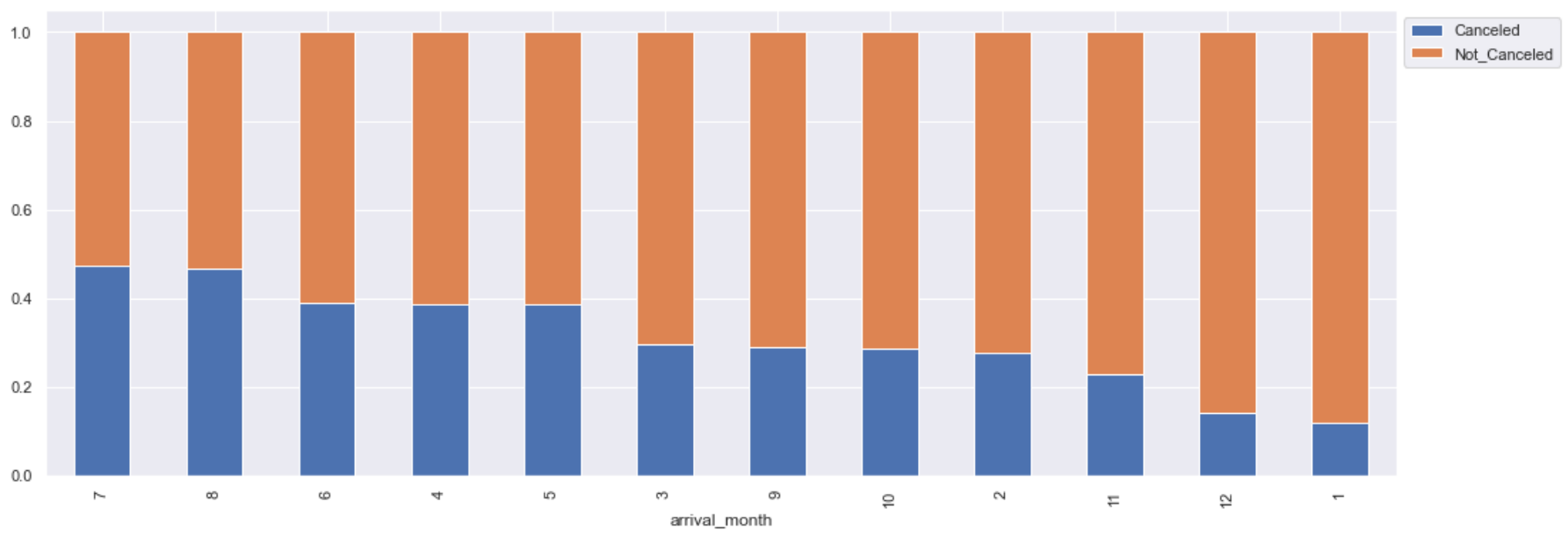
EDA: Price vs Room Type (vs Status)



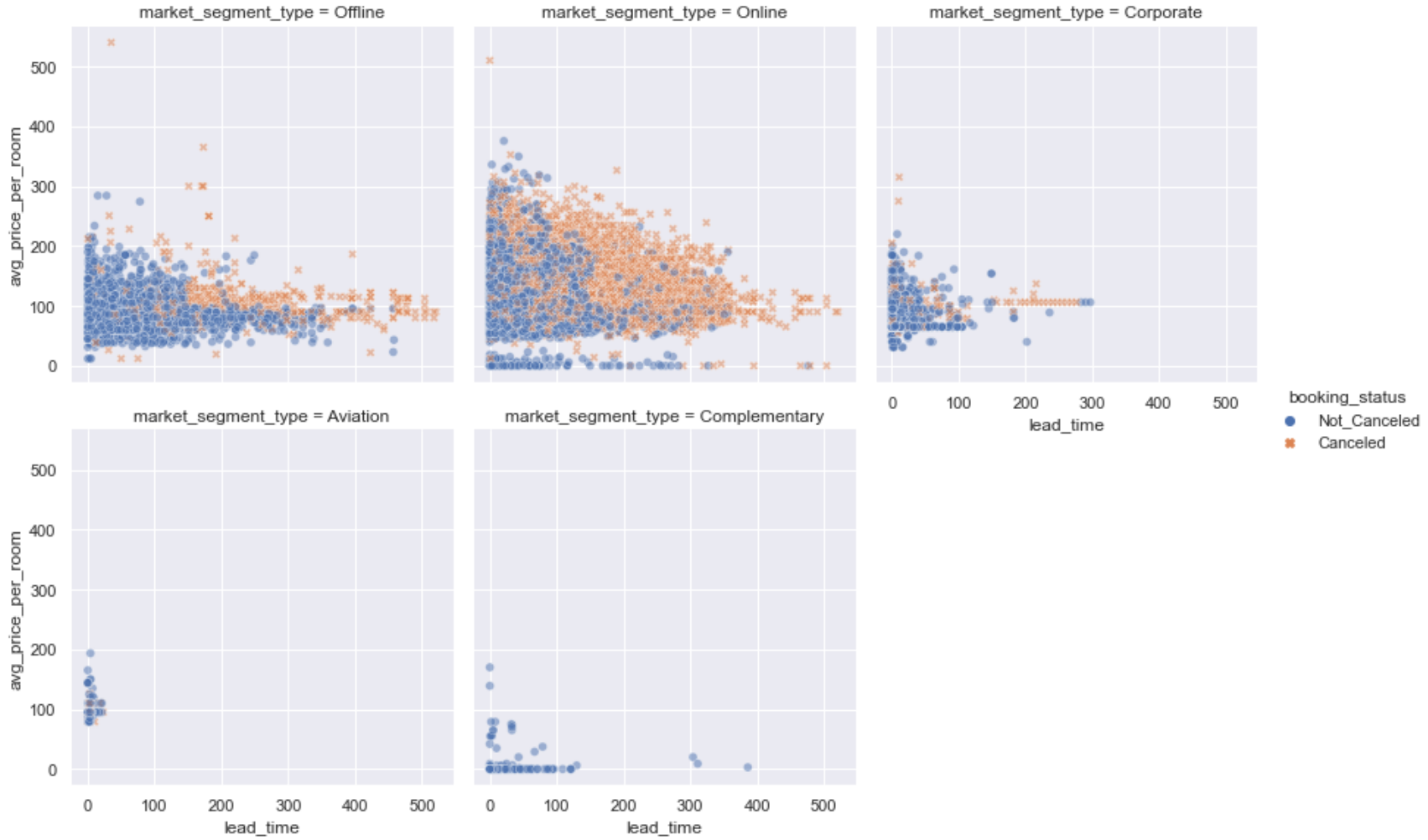
- Type 1 has lowest cancellation and is reserved frequently
- Types 4, 5, 6, 7 are more expensive
- Types 4, 5, 6 have higher proportion of cancellation
- More expensive rooms have higher proportion of cancellation, except room type 7, which is possibly specialty room

EDA: Status vs Month

- Although bookings begin to increase beginning in March and reach their height in summer months, the summer months also contribute to most cancellations
- Busiest months by bookings not cancelled in descending order: March, August, May, April, July, June
- Busiest months by number of reservations are in different order

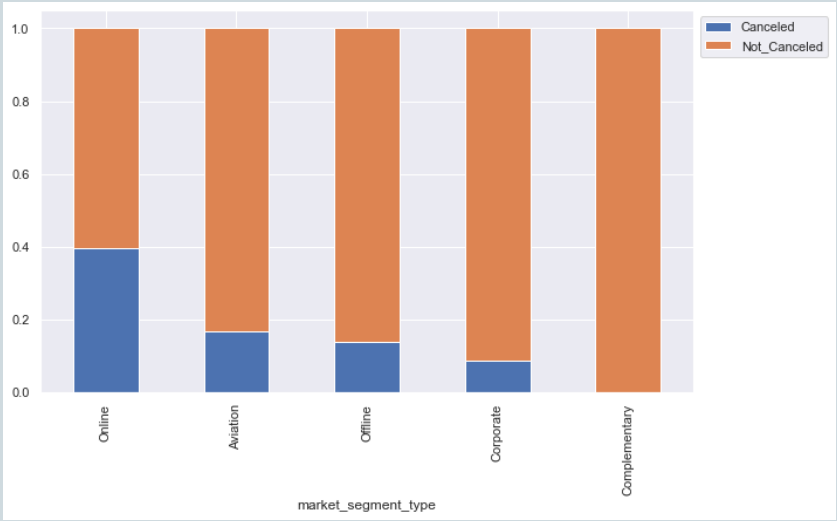


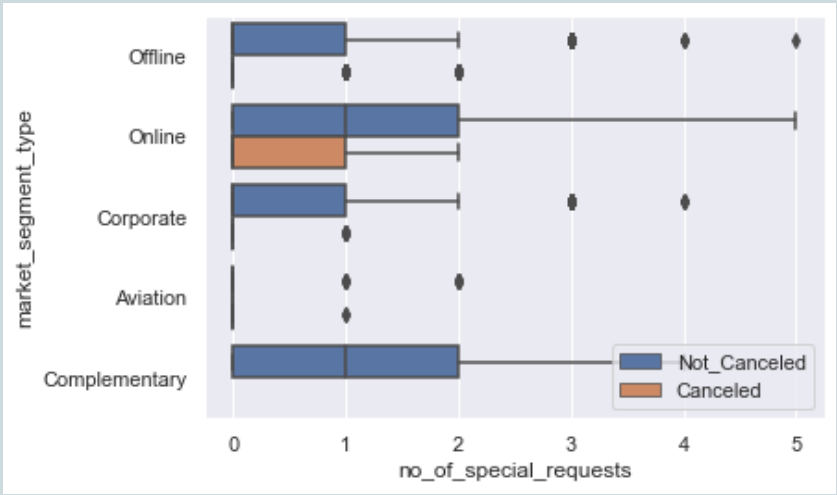
EDA: market vs price vs lead time vs status



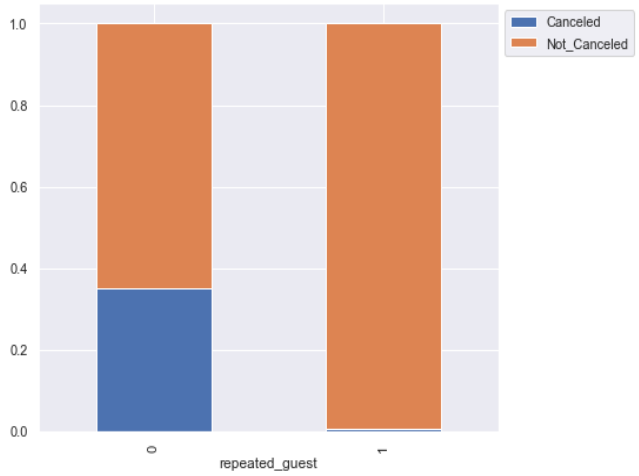
EDA: market vs status (vs special requests)

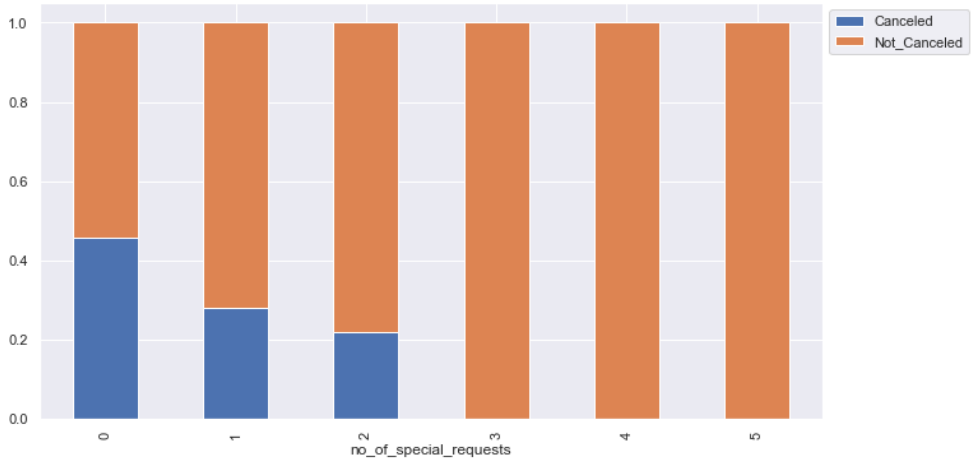
- Online market segment
  - highest proportion of cancellations
  - largest number of reservations by far
  - most reservations not cancelled (of all bookings)
  - most expensive prices per room, even with same lead time
  - average prices trend downward with more lead time
  - some free rooms from this segment, with a few free rooms cancelled
  - more lead time & higher prices have more cancellations
- Corporate market segment (Ignoring complimentary market segment)
  - cheapest on average (median)
  - smallest proportion of cancellations
  - prices start to increase with more lead time
- Offline market segment (Ignoring complimentary market segment)
  - second lowest average price/room
  - second lowest proportion of cancellations
  - prices do not have an upward/downward trend with lead time
  - more lead time and higher prices have more cancellations
- Aviation market segment
  - fewest bookings
  - second highest proportion of cancellations
  - second highest average price per room
  - very little lead time
- Complimentary market segment
  - no cancelled bookings



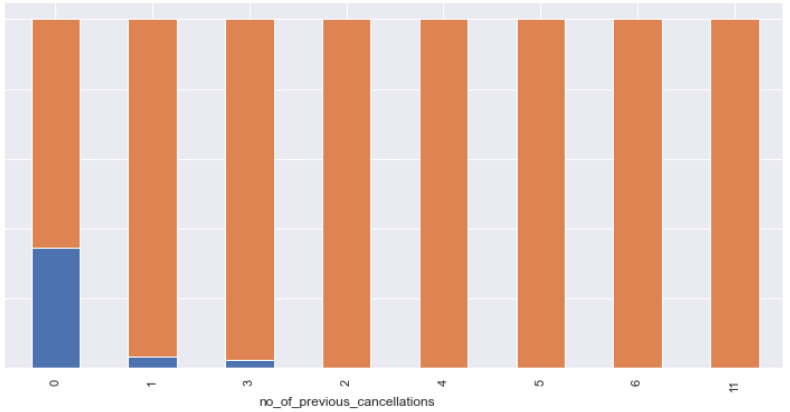


EDA: Status vs Repeat/Requests/Cancellations





- Less than 1% repeated guests cancelled
- Over one-third non-repeat guests cancelled
- Reservations with 3 or more special requests were not cancelled
- As number of special requests increases, proportion of cancellation decreases
- Generally speaking, most cancellations came from guests with no previous cancellations, likely non-repeat guests
- There were some but few cancellations from those with 1 and 3 cancellations, and a cancellation from someone with 13 previous cancellations

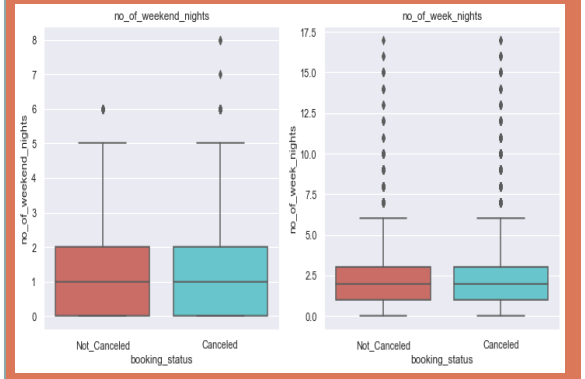


EDA: Nights (Weekend & Weeknights)

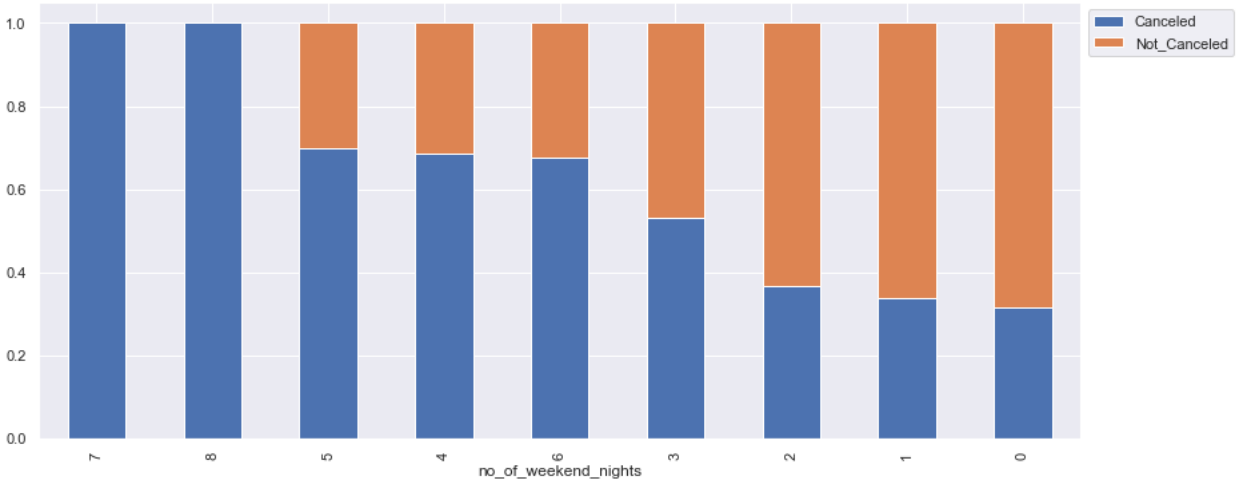
Weekend-only stays are canceled more frequently than weeknight-only stays

Bookings longer than just a couple nights have higher proportion of cancellation

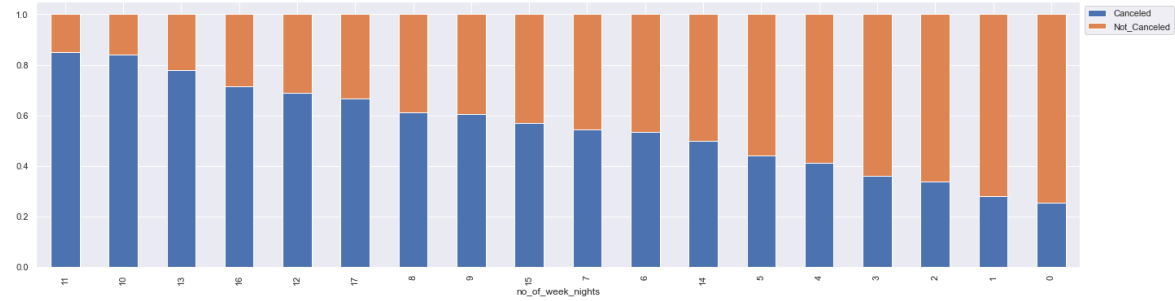
Number of nights alone does not indicate cancellation



Weekend nights



Weeknights



Key Findings and Insights

- Higher Likelihood of Cancellation:
  - Longer stays
  - Long lead time
  - Higher average price per room
  - Fewer/no special requests
  - No meal plan
  - Online
  - March-August months
  - Child guest
  - Weekend-only stay
  - Non-repeat guest
- Online market is the largest, which means there are the most bookings from the Online market, but it also has the largest proportion of cancellations. This may be due in part to the lack of connection customers have when making reservations online.

Model Overview & Performance Summary

Logistic Regression Model

- Logistic Regression with Sklearn library
  - Training Recall: 0.6239
  - Test Recall: 0.6204
  - Training F1: 0.6723
  - Test F1: 0.6725
- Logistic Regression with Sklearn on square root transformed set
  - Training Recall: 0.6382
  - Test Recall: 0.6310
  - Training F1: 0.6763
  - Test F1: 0.6755
- Logistic Regression with statsmodels library
  - Training Recall: 0.6235
  - Test Recall: 0.6184
  - Training F1: 0.6727
  - Test F1: 0.6716
- Optimal threshold using ROC-AUC curve
  - Training Recall: 0.7978
  - Test Recall: 0.7936
  - Training F1: 0.7066
  - Test F1: 0.7067
- Use Precision-Recall curve to try finding better threshold
  - Training Recall: 0.6972
  - Test Recall: 0.6862
  - Training F1: 0.6942
  - Test F1: 0.6909

Interpretation of Coefficients

- no_of_weekend_nights: Holding all other features constant, increasing the booking by 1 weekend night will increase the odds of the booking being cancelled by 1.05 times or a 4.95% increase
- no_of_week_nights: Holding all other features constant, increasing the booking by 1 week night will increase the odds of cancellation by 1.11 times or 11.26% increase
- required_car_parking_space: Holding all other features constant, a booking with a car parking space required decreases the odds of cancellation by 78.23%
- lead_time: Holding all other features constant, increasing the lead time by a day will increase the odds of cancellation 1.71%
- repeated_guest: Holding all other features constant, the odds of a cancellation from a repeat guest is 97.82% less likely
- no_of_previous_cancellations: Holding all other features constant, increasing the number of previous cancellations by one unit will increase the odds of cancellation by 41.81%
- avg_price_per_room: Holding all other features constant, increasing the average price per room by 1 unit will increase the odds of cancellation by 1.86%
- no_of_special_requests: Holding all other features constant, increasing the number of special requests by 1 unit will decrease the odds of cancellation by 73.57%
- type_of_meal_plan: Holding all other features constant, selecting meal plan 2 will decrease the odds of cancellation by 18.9% while no selection of a meal will increase the odds by 40.71%
- type_of_room_reserved: Holding all other features constant, reservation of room types 4, 5, 6, and 7 each decrease the odds of cancellation by 16.89%, 38.96%, 35.54%, and 67.63%, respectively
- market_segment_type: Holding all other features constant, Corporate and Offline market types will each decrease the odds of cancellation by 44.78% and 89.36%, respectively
- arrival_month: Holding all other features constant, arriving in the months of January, May, June, and December each decrease the odds of cancellation by 49.82%, 20.19%, 13.42%, and 68.44%, respectively. Arriving in the months of February, March, and November increase the odds of cancellation by 96.42%, 40.24%, and 51.88%, respectively

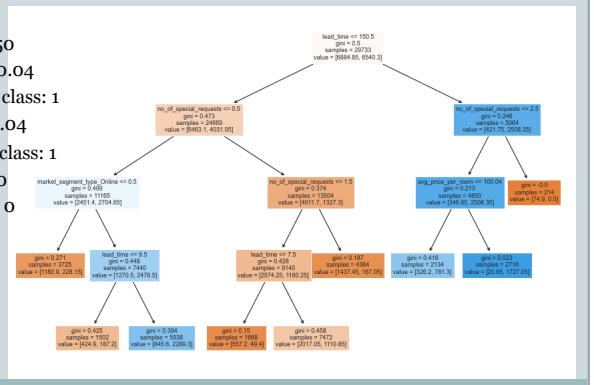
Model Overview & Performance Summary

Decision Tree Model

- use DecisionTreeClassfier function, default 'gini' criteria to split
  - Training Recall: 1.0000
  - Test Recall: 0.6946
  - Training F1: 0.9944
  - Test F1:0.6946
- pre-prune: use GridSearch to compute optimal values of hyperparameters
  - Training Recall: 0.7747
  - Test Recall: 0.7658
  - Training F1: 0.7459
  - Test F1: 0.7382
- post-prune: Cost Complexity Pruning
  - Training Recall: 0.7336
  - Test Recall: 0.7386
  - Training F1: 0.7080
  - Test F1: 0.7125
- Lead time, number of special requests are the top two most important features to predict cancellations

Post-Pruned Tree

```
|--- lead_time <= 150.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- weights: [1180.90, 228.15] class: 0
|   |   |--- market_segment_type_Online >  0.50
|   |   |   |--- lead_time <= 9.50
|   |   |   |   |--- weights: [424.90, 187.20] class: 0
|   |   |   |--- lead_time >  9.50
|   |   |   |   |--- weights: [845.60, 2289.30] class: 1
|   |--- no_of_special_requests >  0.50
|   |   |--- no_of_special_requests <= 1.50
|   |   |   |--- lead_time <= 7.50
|   |   |   |   |--- weights: [557.20, 49.40] class: 0
|   |   |   |--- lead_time >  7.50
|   |   |   |   |--- weights: [2017.05, 1110.85] class: 0
|   |   |--- no_of_special_requests >  1.50
|   |   |   |--- weights: [1437.45, 167.05] class: 0
|--- lead_time >  150.50
|   |--- no_of_special_requests <= 2.50
|   |   |--- avg_price_per_room <= 100.04
|   |   |   |--- weights: [326.20, 781.30] class: 1
|   |   |--- avg_price_per_room >  100.04
|   |   |   |--- weights: [20.65, 1727.05] class: 1
|   |--- no_of_special_requests >  2.50
|   |   |--- weights: [74.90, 0.00] class: 0
```



General Recommendations

- Keep prices competitive (low) to avoid cancellation
- Make special request and low-cost meal plans look enticing and easy to select from website. Customers may be more hesitant to cancel if booking is personalized
- Limit lead time for online reservation
  - Call or fill out a form to be contacted
- Set up app to send SMS for reservations with long lead time 1-2 months prior to arrival date
  - Changes to reservation
  - Confirm/cancel
- Repeat guests have much lower odds of cancellation
  - Free loyalty program providing increasingly more perks (that are low-cost to hotel)
  - Offer complimentary rooms (with very short lead time) for frequent guests when there may be rooms that cannot be sold quickly, rather than sitting empty. Aforementioned app can be used to alert a number of loyal customers of special deals/ complimentary rooms for short lead time
- Implement typical 24-hour cancellation window for no refund
- Offer online specials for non-refundable discounted rate on short lead time deals, which are prominent on the website

Further Recommendations

Logistic Regression Model

- During months with increased odds of cancellation (Feb, Mar, Nov), strategically double-book rooms
- During months with decreased odds of cancellation (Jan, May, Jun, Dec), avoid double-booking many rooms

Decision Tree Model

- Example of specifics: a) If a customer's lead time is greater than 150 days, there is a high chance the customer will cancel the booking, unless the customer is from a market segment besides Online. If Online, lead time of at least 10 days will likely lead to cancellation. b) If a customer's lead time is no more than 150 days and has at least one special request, that guest will likely not cancel the booking.