# Hotel Reservation Cancellation Prediction

## Interim Report

*Team*:

| | |
|---|---|
| Batch details | PGPDSE CHN Jan23 (Group – 2) |
| Team members | 1. Praveen Arokia Raj A<br>2. G.Raghav<br>3. Sathish J<br>4. Harisri Karthikeyan<br>5. Balachandar P<br>6. Mohana Devan R |
| Domain of Project | Tourism Industry |
| Proposed project title | Hotel Reservation Cancellation Prediction |
| Group Number | Group 2 |
| Team Leader | Balachandar P |
| Mentor Name | Mr Jayveer Nanda |

# 1 Table of Contents

**Abstract:** ..... 4

**Summary of the problem statement, Data and findings** ..... 10

1.1. Problem Statement ..... 10

1.2. Project Objectives: ..... 10

**2. EDA** ..... 12

2.1. Approach: ..... 12

2.2. Analysis: ..... 13

# Abstract:

This report provides an extensive review of the hotel industry, focusing on the current practices related to booking a room for stay. It presents background research to understand the dynamics of the industry, including key players, market trends, and technological advancements. The report aims to offer insights into the booking process, various options available to customers, and the challenges faced by the industry. This information will assist in understanding the industry landscape and identifying opportunities for improvement and innovation.

## Introduction:

The hotel industry offers the convenience of booking rooms in advance for stays, providing travelers with the ability to secure accommodations ahead of their travel dates. This section presents key aspects of the hotel industry related to booking rooms in advance.

## Current Practices on Booking a Room for Stay:

## Online Booking Platforms:

With the advancement of technology, online platforms have become the primary method for booking hotel rooms in advance. Numerous websites and mobile apps specialize in facilitating hotel bookings. These platforms typically provide search functionality, allowing users to specify their desired destination, travel dates, and other preferences to find suitable hotel options.

## Hotel Websites and Mobile Apps:

Many hotels have their own websites and mobile apps that allow customers to book rooms directly. These direct booking channels often provide benefits such as best available rates, loyalty program rewards, and the opportunity to communicate specific requirements or preferences directly with the hotel. Hotels may offer exclusive deals and discounts to encourage customers to book directly.

## User Reviews and Ratings:

Before making a booking, customers often rely on user reviews and ratings to gain insights into the quality and experience of a hotel. Websites and apps typically provide customer reviews and ratings for each hotel, allowing prospective guests to make informed decisions based on the feedback shared by previous visitors.

## Price Comparison and Deals:

Customers frequently compare prices across different platforms to find the best deals and value for their money. Online booking platforms often display prices from multiple sources, enabling customers to compare rates and choose the most competitive option. Price comparison websites and tools further assist customers in finding the lowest available prices.

## Flexible Booking Policies:

To provide customers with more flexibility, many hotels offer various booking policies. Flexible booking options may include free cancellation up to a certain date, the ability to modify reservations without penalties, or the choice between refundable and non-refundable rates. Such policies cater to travelers who value flexibility and may need to change their plans.

## Integration with Travel Apps and Services:

Hotel booking platforms and hotel chains often integrate with popular travel apps and services to provide a seamless experience for travelers. For example, integration with airline booking platforms allows customers to book flights and hotel accommodations together, and integration with ride-sharing services facilitates transportation arrangements.

## Personalization and Loyalty Programs:

Many hotels aim to provide personalized experiences for their guests. They collect guest data and preferences to offer tailored recommendations, room upgrades, and personalized amenities. Loyalty programs are also prevalent, offering exclusive benefits, discounts, and rewards to frequent guests.

## Secure Payment Options:

Online booking platforms and hotel websites ensure secure payment options for customers. They employ encryption and secure payment gateways to protect customers' financial information during the booking process. Common payment methods include credit/debit cards, digital wallets, and bank transfers.

# Scope of the Hotel Industry:

The scope of the hotel industry is extensive and encompasses various aspects related to accommodations, hospitality services, and related sectors. This section outlines the key elements that define the scope of the hotel industry.

## Accommodation Services:

The hotel industry primarily revolves around providing accommodation services to travelers. This includes hotels, resorts, motels, bed and breakfast establishments, and other lodging options. The industry caters to different types of travelers, ranging from budget-conscious tourists to luxury-seeking guests.

## Hospitality Services:

In addition to accommodations, hotels offer a range of hospitality services to enhance the guest experience. This includes concierge services, room service, housekeeping, laundry facilities, fitness centers, swimming pools, and spa amenities. The level of hospitality services provided varies across different types of hotels.

## Event and Conference Facilities:

Many hotels provide event and conference facilities, catering to business meetings, conferences, weddings, and other social gatherings. These venues offer meeting rooms, banquet halls, audiovisual equipment, and catering services. Event and conference facilities contribute significantly to the revenue of hotels.

## Food and Beverage Operations:

Hotels often have on-site restaurants, cafes, bars, or lounges that provide food and beverage services to guests. These operations may include breakfast buffets, à la carte dining, room service, and specialty cuisines. Food and beverage operations contribute to the overall guest experience and revenue generation for hotels.

## Travel and Tourism Integration:

The hotel industry is closely linked to the travel and tourism sector. Hotels collaborate with travel agencies, tour operators, and online travel agencies to attract guests and promote their services. They often participate in tourism campaigns and develop partnerships to tap into the potential of the tourism industry.

## Employment and Economic Impact:

The hotel industry plays a significant role in employment generation and economic growth. It provides employment opportunities for a wide range of roles, including front desk staff, housekeeping personnel, chefs, servers, and managerial positions. Hotels contribute to local economies through job creation, tax revenue, and tourism-related expenditures.

## Technological Advancements and Innovations:

Technological advancements have greatly influenced the hotel industry. Hotels continuously embrace innovations to improve operations, enhance guest experiences, and streamline processes. This includes the adoption of mobile technology, artificial intelligence, smart room features, energy-efficient systems, and sustainability initiatives.

# Literature Survey: Hotel Room Reservation - Trends and Practices

## Introduction:

Hotel room reservation is a critical aspect of the hotel industry, enabling customers to secure accommodations in advance. This literature survey examines the existing research on hotel room reservation, focusing on the trends and practices in the industry. It explores the evolving landscape of reservation methods, technological advancements, customer preferences, and challenges faced by the hotel industry in the reservation process.

## Online Booking Platforms:

The advent of online booking platforms has transformed the hotel reservation landscape. Research by Wang et al. (2018) highlighted the increasing preference for online platforms due to their convenience and accessibility. It revealed that customers value the ability to search and compare different hotels, view real-time availability, and make reservations directly through websites or mobile apps.

## Direct Booking with Hotels:

Many hotels offer their own websites and mobile apps for customers to book rooms directly. Research by Sigala and Gretzel (2019) emphasized the benefits of direct booking, such as competitive rates, loyalty program rewards, and the opportunity to communicate specific preferences directly with the hotel. It also highlighted the importance of hotel websites' usability and user experience in facilitating direct bookings.

## Dynamic Pricing and Personalization:

Dynamic pricing is a common practice in hotel room reservation, with prices varying based on factors such as demand, seasonality, and availability. Research by Xie et al. (2017) examined the impact of dynamic pricing on customer behavior and found that personalized pricing strategies, tailored to individual preferences and booking patterns, can enhance customer satisfaction and increase booking conversion rates.

## Mobile Technology and Mobile Booking:

The proliferation of mobile technology has led to the rise of mobile booking in the hotel industry. Studies by Buhalis and Tan (2015) and Law et al. (2017) explored the adoption and usage patterns of mobile booking apps among travelers. They found that mobile booking

offers convenience, real-time access, and personalized features, making it a popular choice for customers seeking instant reservation capabilities.

## User Reviews and Ratings:

User reviews and ratings play a significant role in the hotel reservation process. Research by Li and Wang (2019) highlighted the influence of online reviews on customers' decision-making. It demonstrated that positive reviews and high ratings significantly impact customers' perceptions of hotel quality, trustworthiness, and booking intentions. However, the study also emphasized the importance of managing fake reviews and ensuring their credibility.

## Distribution Channels and Online Travel Agencies (OTAs):

The presence of online travel agencies (OTAs) as intermediaries in the hotel reservation process has been extensively studied. Research by Xiang et al. (2015) examined the relationship between hotels and OTAs, exploring the advantages and challenges associated with OTA distribution. It found that while OTAs provide a wide reach and access to a broad customer base, hotels must carefully manage their distribution channels to maintain brand identity and profitability.

## Cancellation Policies and Customer Flexibility:

Hotel reservation policies, particularly cancellation policies, impact customers' decision-making and flexibility. Research by Cao et al. (2016) examined the effect of flexible cancellation policies on customer satisfaction and loyalty. It revealed that lenient cancellation policies, including free cancellation or flexible modification options, positively influence customer perceptions of hotel service quality and increase repeat bookings.

## Conclusion:

The literature survey highlights the evolving trends and practices in hotel room reservation. It underscores the growing importance of online booking platforms, direct booking with hotels, dynamic pricing, mobile technology, user reviews, distribution channels, and flexible cancellation policies. Understanding these trends and practices is crucial for hotels to enhance their reservation processes, meet customer expectations, and remain competitive in the evolving hospitality industry.

# Summary of the problem statement, Data and findings

## 1.1. Problem Statement

Develop a classification model to predict the booking status (canceled or not canceled) based on the provided booking information. The model should use features such as the number of adults and children, length of stay, meal plan, parking requirements, room type, lead time, and previous booking and cancellation history to accurately classify whether a booking will be canceled or not. The objective is to develop a reliable model that can help hotel management to identify potential cancellations in advance and take appropriate measures to reduce the impact on their business.

The model's performance should be evaluated based on metrics such as accuracy, precision, recall, and F1 score. The model should be able to provide actionable insights to the hotel management team for better decision-making and business planning."

## 1.2. Project Objectives:

The business objectives for the Hotel Reservations Classification Dataset include maximizing revenue, improving customer satisfaction, reducing cancellations, and streamlining operations. By analyzing the data, hotels can identify trends and patterns that can inform pricing and revenue management strategies, address negative guest experiences, predict and prevent cancellations, and streamline operations to enhance efficiency and profitability. The ultimate goal is to improve the overall guest experience while optimizing hotel operations for maximum success.

## 1.3. Data & Findings :

- Details about the data and dataset files are given in below link, https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset
- Data Dictionary:
    1. Booking_ID: unique identifier of each booking
    2. no_of_adults: Number of adults
    3. no_of_children: Number of Children
    4. no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
    5. no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
    6. type_of_meal_plan: Type of meal plan booked by the customer:

7. required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
8. room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
9. lead_time: Number of days between the date of booking and the arrival date
10. arrival_year: Year of arrival date
11. arrival_month: Month of arrival date
12. arrival_date: Date of the month
13. market_segment_type: Market segment designation.
14. repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
15. no_of_previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
16. no_of_previous_bookings_not_cancelled: Number of previous bookings not cancelled by the customer prior to the current booking
17. avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
18. no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc.)
19. booking_status: Flag indicating if the booking was cancelled or not.

# 2. EDA

## 2.1. Approach:

**Checking for the summary statistics of the dataset**

```
df_hotel.describe().T
```

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| no_of_adults | 36275.0 | 1.844962 | 0.518715 | 0.0 | 2.0 | 2.00 | 2.0 | 4.0 |
| no_of_children | 36275.0 | 0.105279 | 0.402648 | 0.0 | 0.0 | 0.00 | 0.0 | 10.0 |
| no_of_weekend_nights | 36275.0 | 0.810724 | 0.870644 | 0.0 | 0.0 | 1.00 | 2.0 | 7.0 |
| no_of_week_nights | 36275.0 | 2.204300 | 1.410905 | 0.0 | 1.0 | 2.00 | 3.0 | 17.0 |
| required_car_parking_space | 36275.0 | 0.030986 | 0.173281 | 0.0 | 0.0 | 0.00 | 0.0 | 1.0 |
| lead_time | 36275.0 | 85.232557 | 85.930817 | 0.0 | 17.0 | 57.00 | 126.0 | 443.0 |
| arrival_year | 36275.0 | 2017.820427 | 0.383836 | 2017.0 | 2018.0 | 2018.00 | 2018.0 | 2018.0 |
| arrival_month | 36275.0 | 7.423653 | 3.069894 | 1.0 | 5.0 | 8.00 | 10.0 | 12.0 |
| arrival_date | 36275.0 | 15.596995 | 8.740447 | 1.0 | 8.0 | 16.00 | 23.0 | 31.0 |
| repeated_guest | 36275.0 | 0.025637 | 0.158053 | 0.0 | 0.0 | 0.00 | 0.0 | 1.0 |
| no_of_previous_cancellations | 36275.0 | 0.023349 | 0.368331 | 0.0 | 0.0 | 0.00 | 0.0 | 13.0 |
| no_of_previous_bookings_not_canceled | 36275.0 | 0.153411 | 1.754171 | 0.0 | 0.0 | 0.00 | 0.0 | 58.0 |
| avg_price_per_room | 36275.0 | 103.423539 | 35.089424 | 0.0 | 80.3 | 99.45 | 120.0 | 540.0 |
| no_of_special_requests | 36275.0 | 0.619655 | 0.786236 | 0.0 | 0.0 | 0.00 | 1.0 | 5.0 |

## Inference:

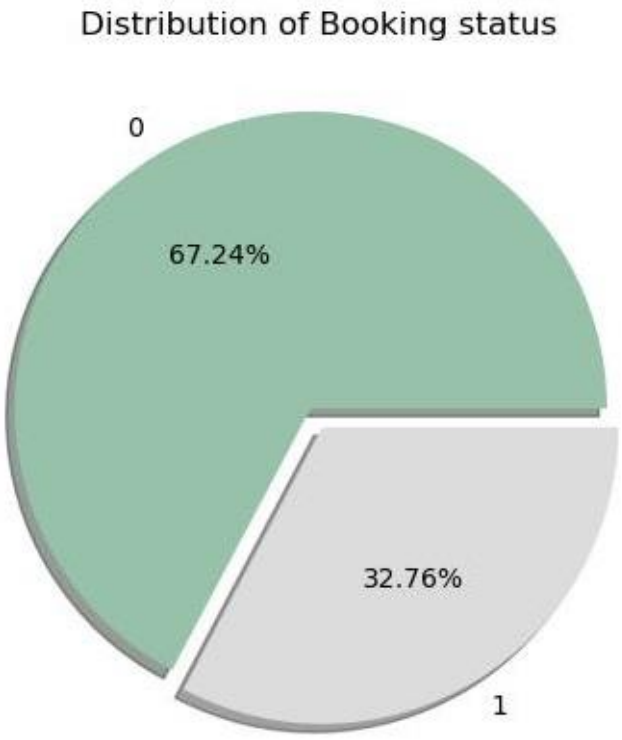
From above statistics we get to know:

- The maximum number of guests who occupied the hotel room over the age of 18 is 4.
- The maximum number of guests who occupied the hotel room below the age of 18 is 10.
- On an average 1 weekend night (Saturday or Sunday) the guest has stayed or booked to stay at the hotel.
- The request for car parking space has majorly been asked for as a requirement (again this is a categorical variable)
- The Lead time between the date of arrival and the date of booking on an average is 85 days and a maximum of 443 days.
- The data collected is for the years 2017, 2018.
- The arrival month we can get to see which month has the highest bookings, 12, December.
- With the arrival date we can classify or infer the following
    - The time of the booking, month end or the mid-month, here most frequent bookings happened for the month end.

- Weekday or the Weekend, where the cancellations happen the most
- In the collected data, this organisation has had more of repeated guest than the new visits.
- The maximum number of previous bookings that were cancelled and not cancelled by the customer prior to the current booking is 13 and 58 respectively.
- Price of the room ranges from $80 to $540 , however the average booking price has been ~$103.
- Comparatively there were not special requests made by the customer (e.g. high floor, view from the room, etc), the maximum number itself is 5.

## 2.2. Analysis:

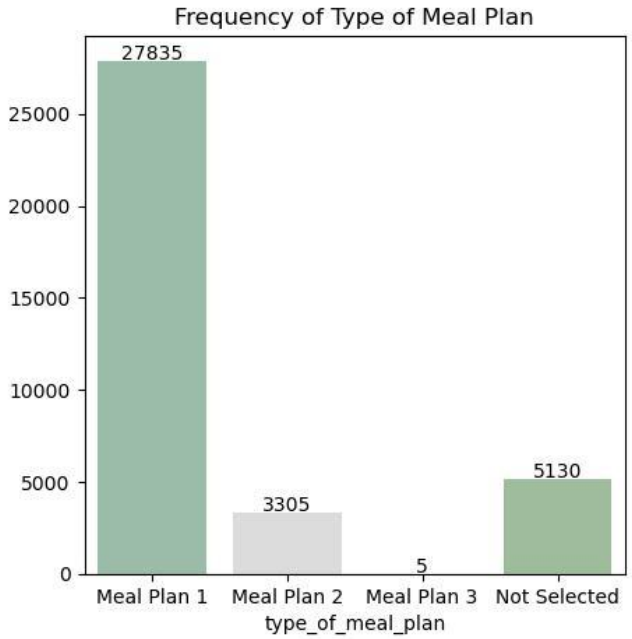
- Target Variable analysis:
  - The prediction in the case study is on the booking status of the hotel reservation, has the following distribution of the class.




- In the target variable we can see it is a binary class and there is class imbalance. Majority of the customers has honoured the reservation, which is 67 per cent of the total bookings.
- Dropping of irrelevant columns from the dataset
  - The variable booking id is a unique identifier of each booking. This variable can be dropped as it might not contribute much to the prediction booking status or might impute noise to the model.

### 1.1.1.1 Univariate Analysis:

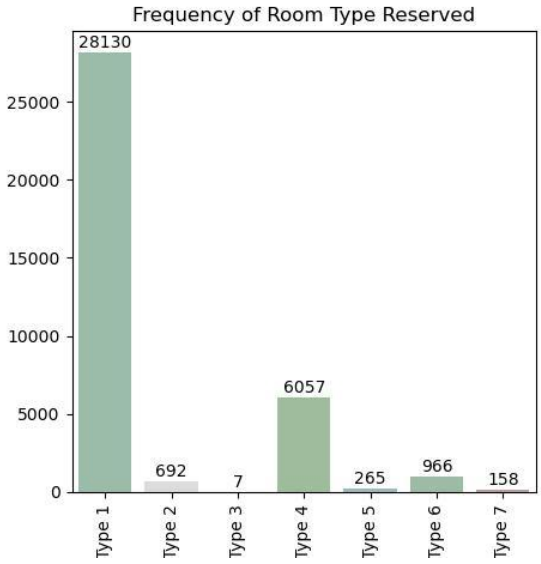
- **Type of meal plan**:




**Inference:**

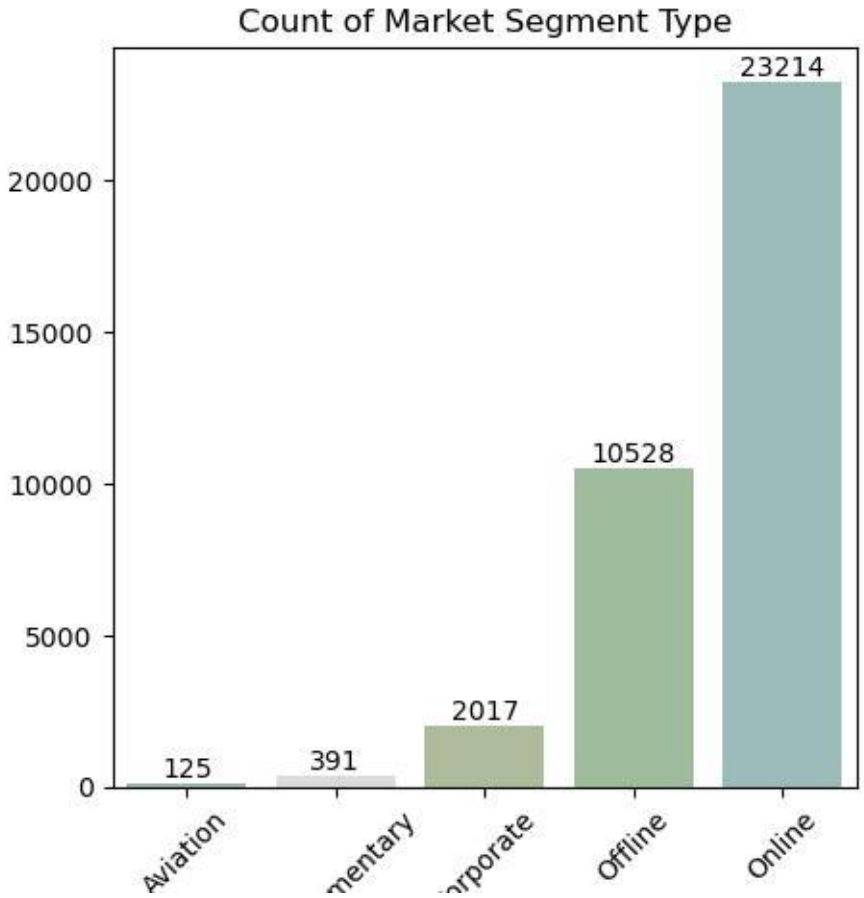
From above count plot it is clearly evident that meal plan 1 is preferred by most of the customers. It is followed by not selected and meal plan 2 with count of 5130 and 3305. Only a few customer prefer meal plan 3

- **Room type reserved**



Inference: Room type 1 is the most preferred type; second highest is the room type 4.
The average price for the room type 1 is ~$95

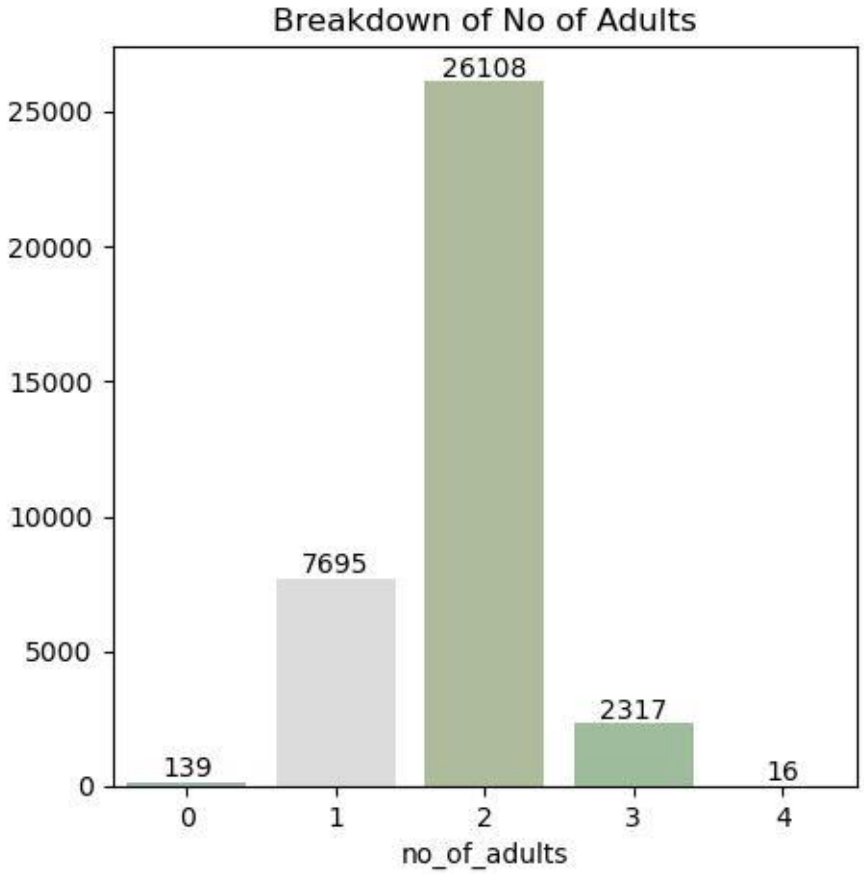
- **Market Segment Type:**




## Inference:

From above plot it is clearly evident that customers who reserved rooms through online modes is higher compared to other modes of reservation. The industry is driven by the online bookings off late, and more thoughts on the cancellation of online cancellation should be given a thought.

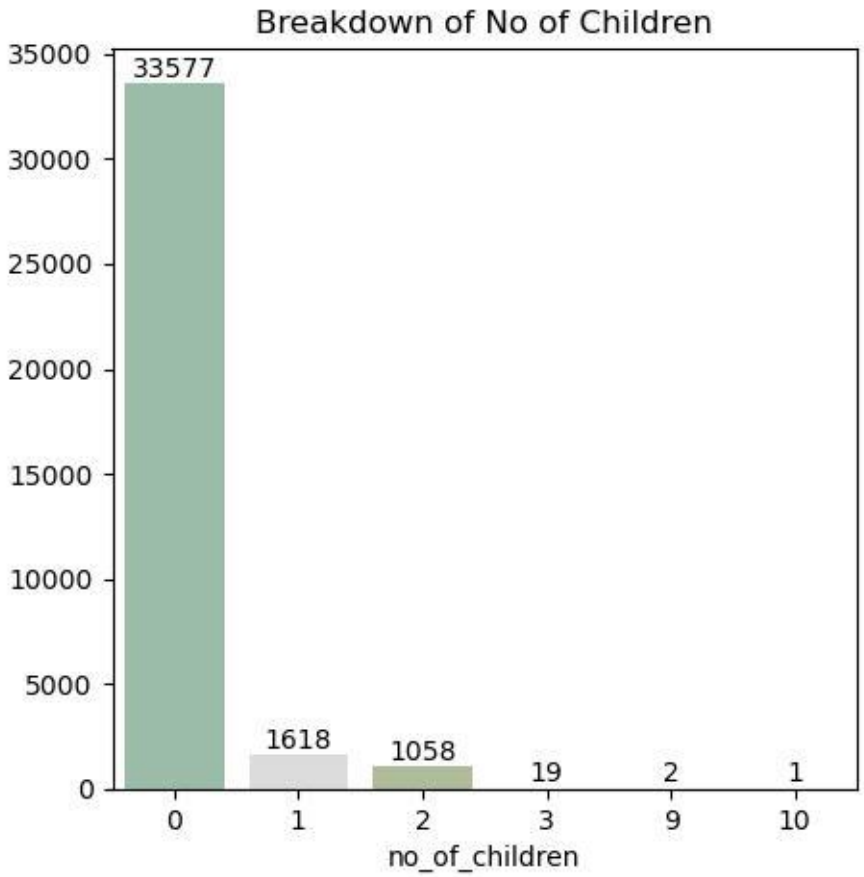
- **No of Adults:**




## Inference:

For the given data, the highest booking has happened for the 2 adults. That is the occupancy has been for 2 adults maximum followed by single occupancy has topped the table.

- **No of Children:**

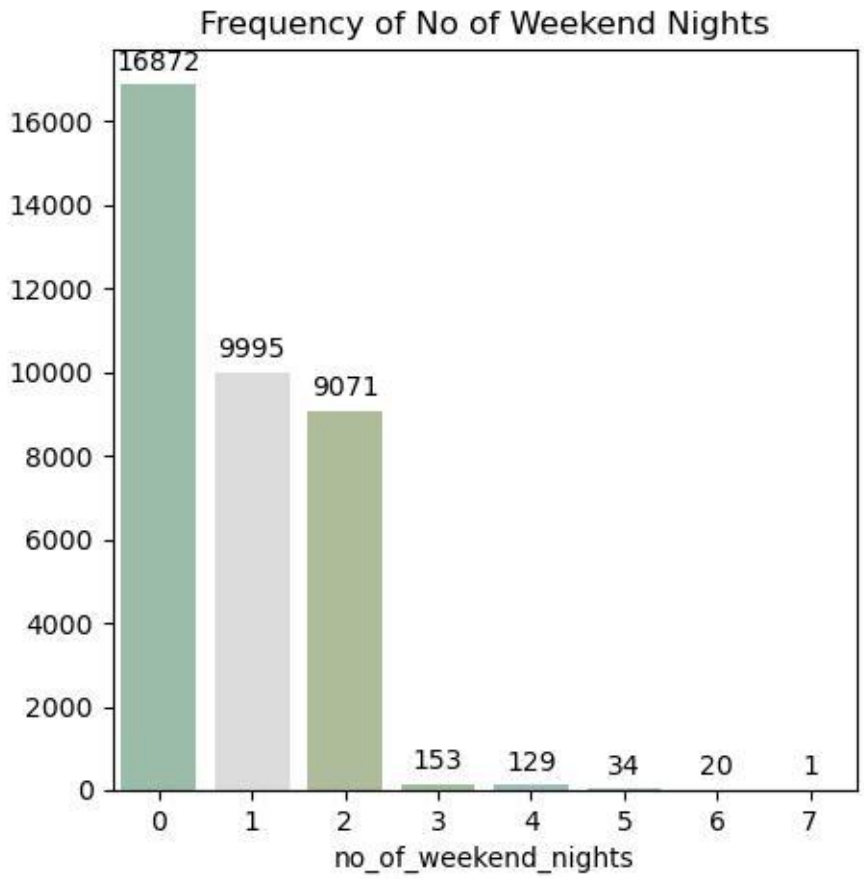



## Inference:

The highest booking is done with zero kids, so it is clearly evident that the booking pattern is highest for 2 adults

- **No of Weekend Nights:**

The "no_of_weekend_nights" variable refers to the number of weekend nights (i.e., Saturday, or Sunday nights) that a guest will be staying at the hotel as part of their reservation.

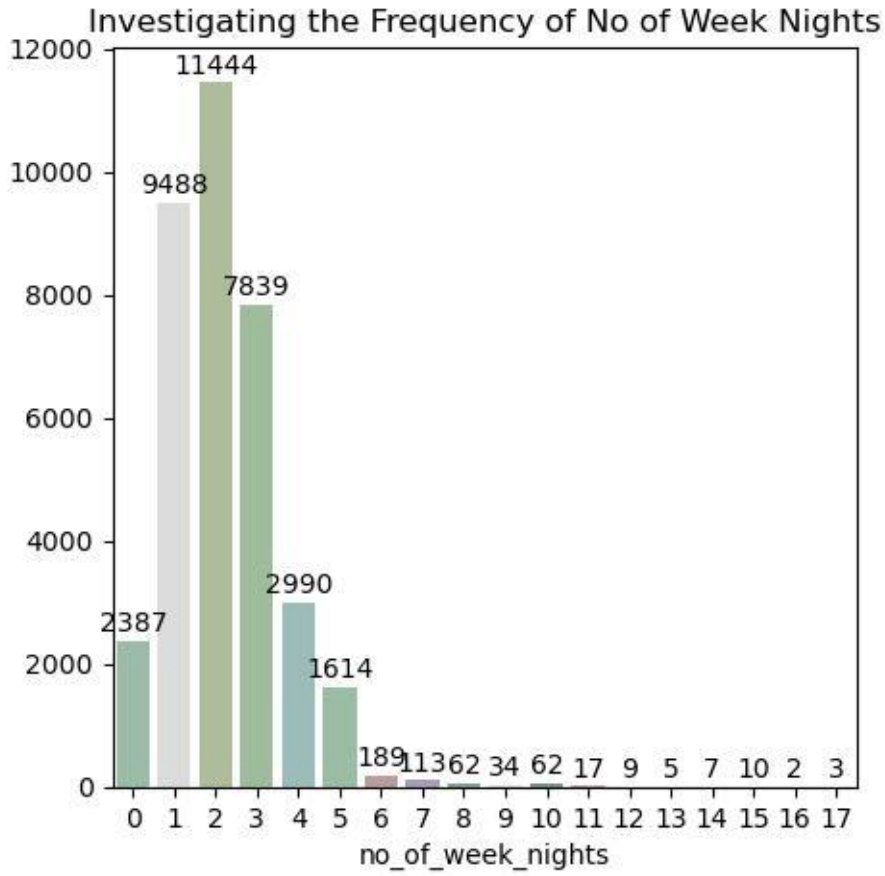



## Inference:

From the above plot it is clearly evident that, least of bookings have happened with 0 no of weekend nights

- **No of week nights:**

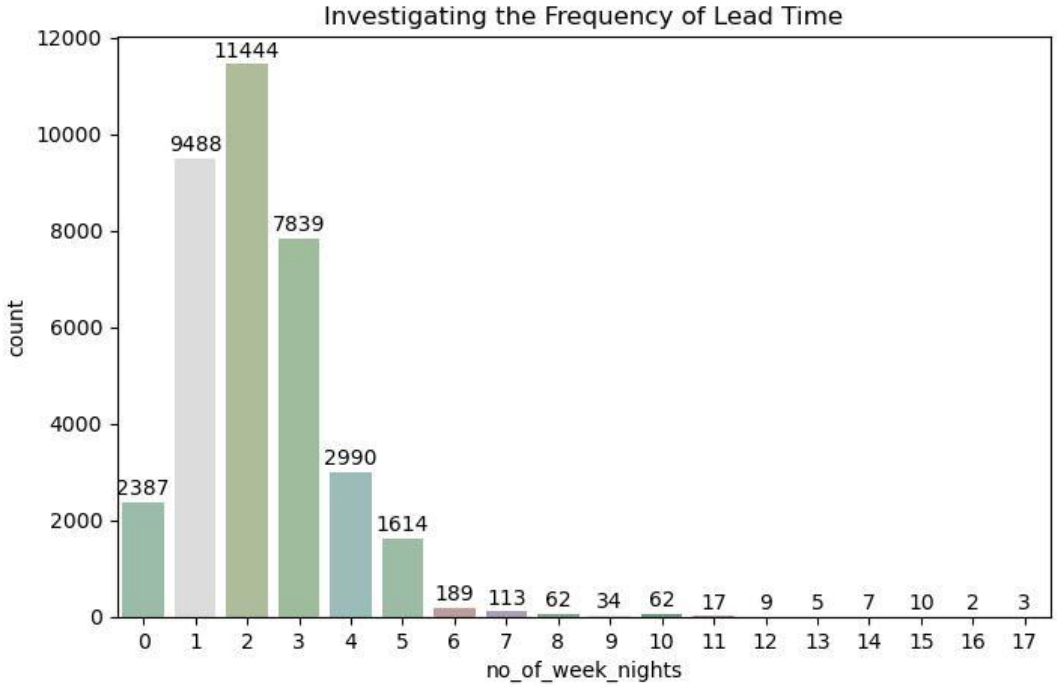
It is the number of weekday nights (i.e., Monday to Friday nights) that a guest will be staying at the hotel as part of their reservation.




## Inference:

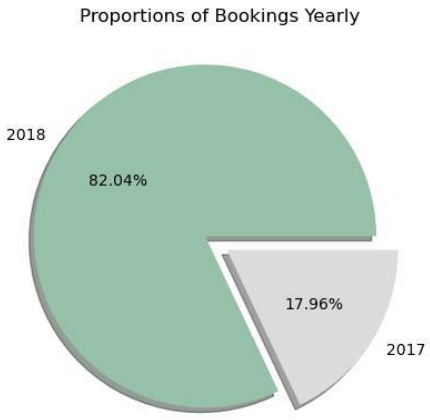
From the above visuals it is clearly seen that the reservation for the weekday nights are most for 1- 3 days

- **Lead Time:**



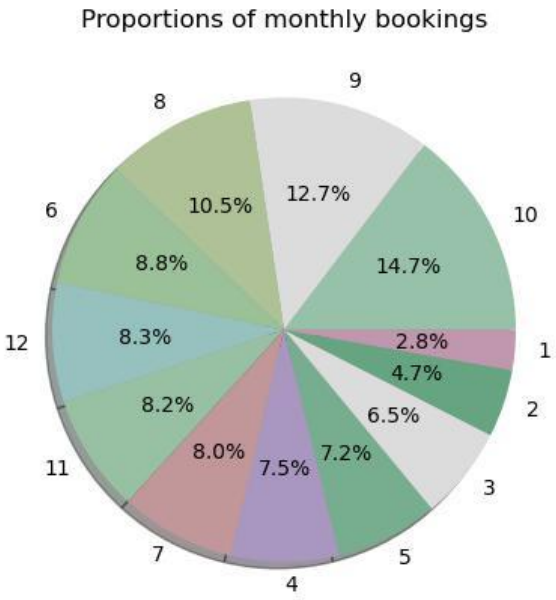

**Inference:** The difference between the booking time and the arrival time is the lead time which has had a trend of 1 to 3 days prior booking; however there are outliers with a maximum of 418 days, which is almost a year prior arrival date.

- **Arrival year**



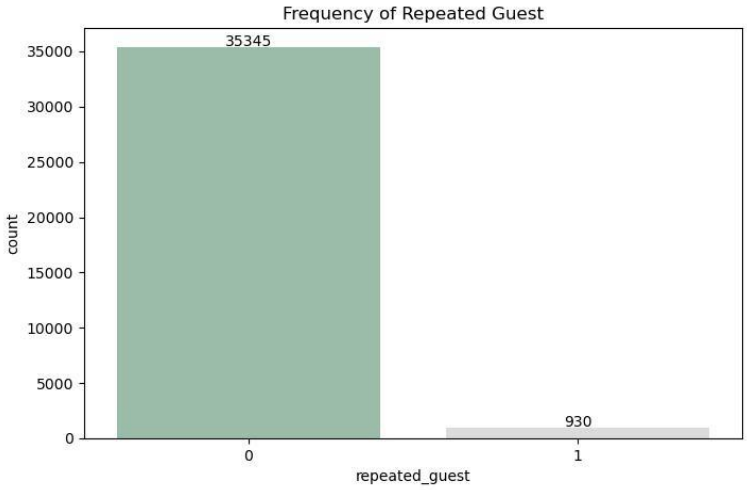

**Inference:** The given data Set is dominated by the 2018 data, or the bookings have boosted high in the year 2018

- **Arrival Month**



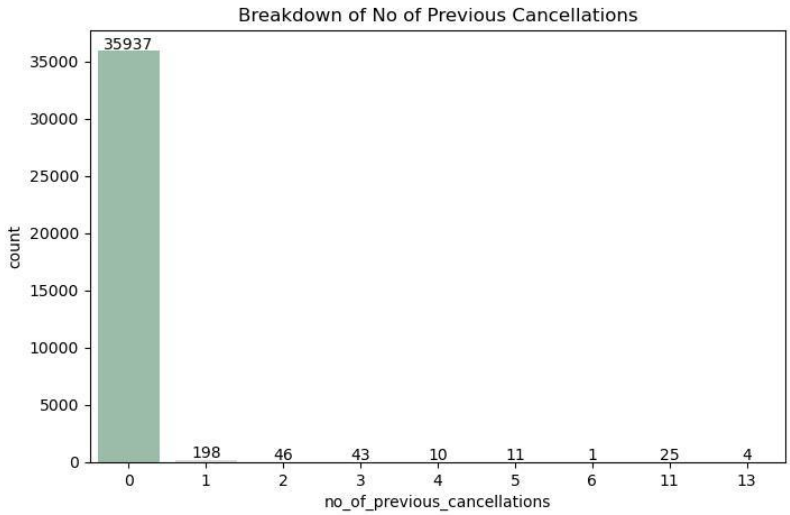

**Inference:** 8,9,10 have the highest no bookings. This is August, September and November. So in the financial year perspective 2nd and 3rd has the highest footfall / bookings. There is a scope for feature engineering for this column, to bin the months quarter wise or bin it seasons wise to establish better pattern with the target variable

- **Repeated guest**



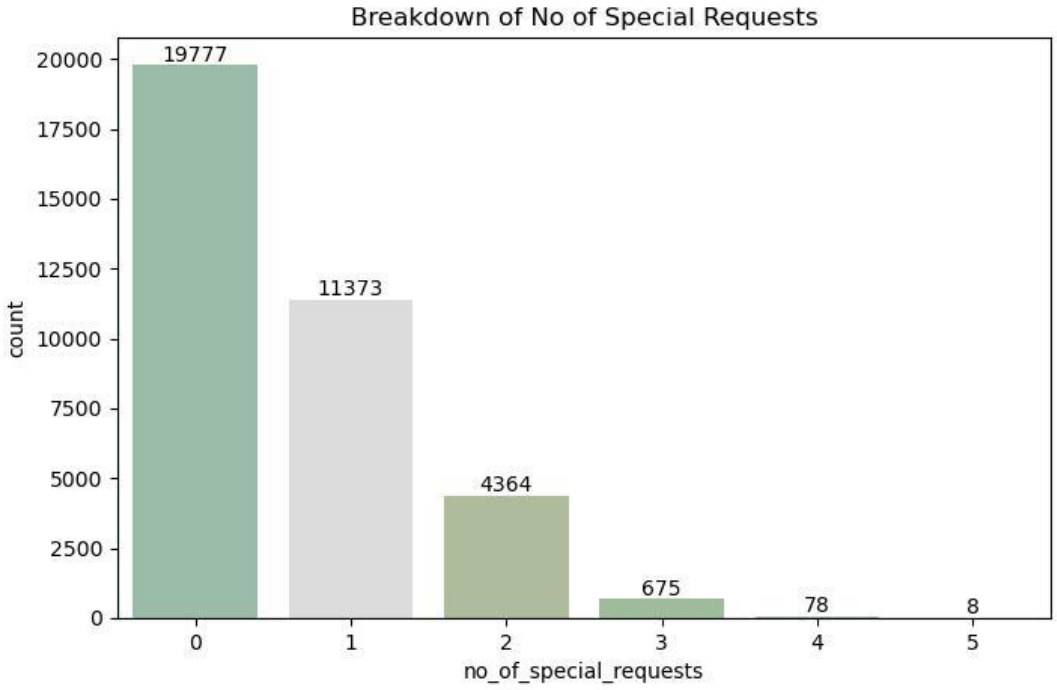
**Inference:** This hotel has a loyal customer base, and also scope for increasing the footfall of new users.

- **No of previous cancellations:**




**Inference:** From the history of bookings, it is clear that the previous bookings have been honored duly by the customer base. This data clearly seems to have data imbalance.

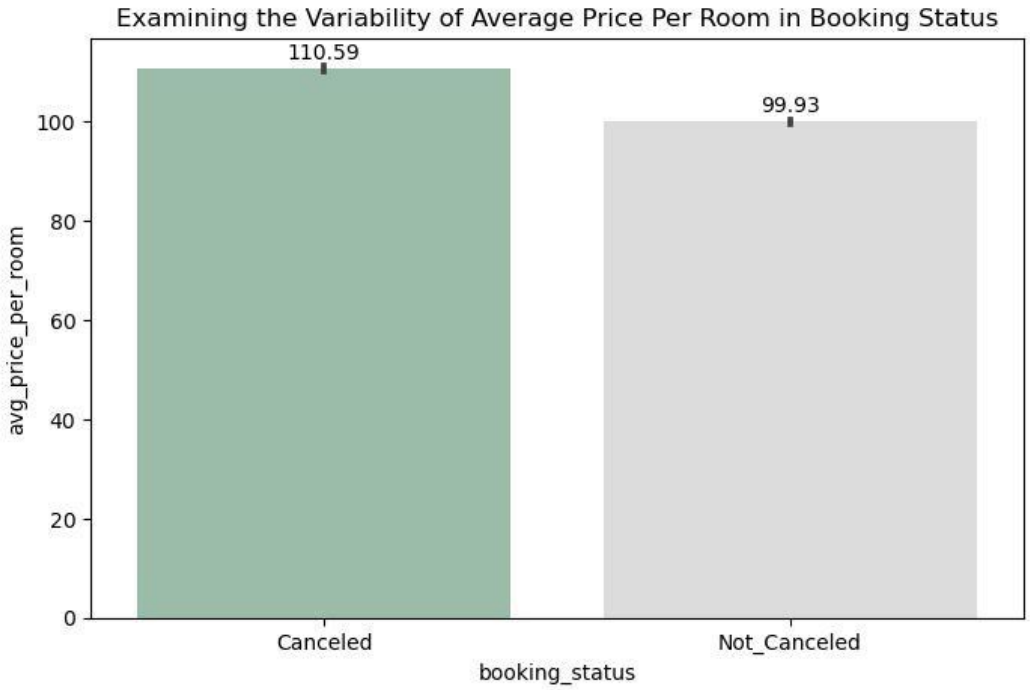
- **No of special requests:**




**Inference:** The data is highly positively skewed and there were no much ask for special requests as well.

### 2.2.2 Bivariate Analysis:

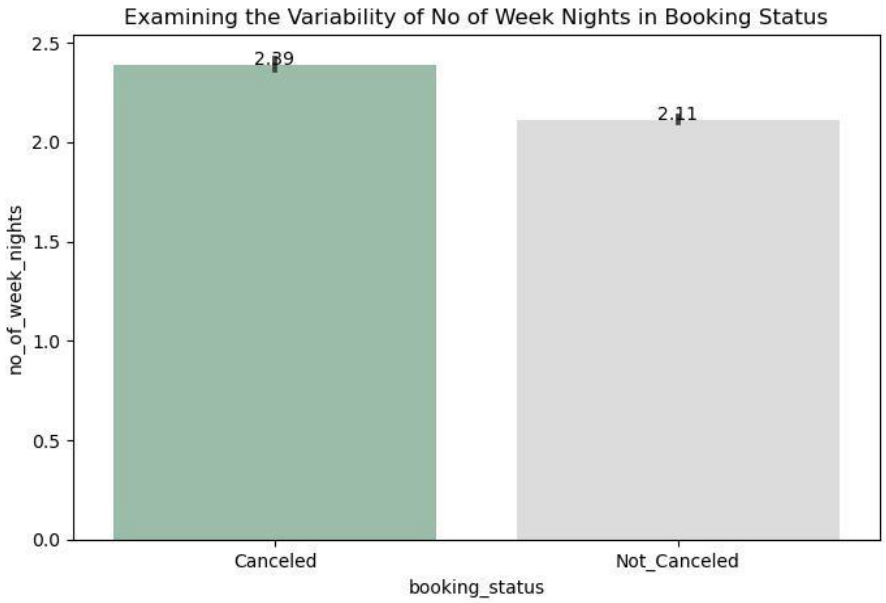
**Examining the Variability of Average Price per Room in Booking Status**




## Inference:

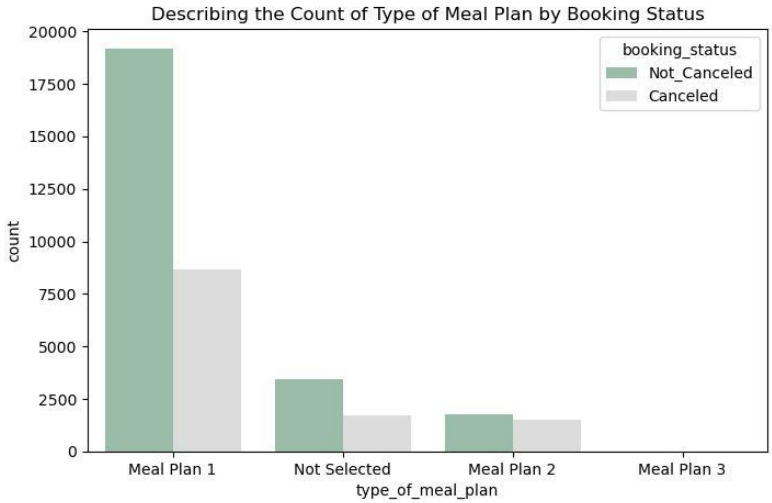
The average booking costs for both cancelled and now - cancelled status are almost the same but the not cancelled comparatively has a higher cost of booking

## Examining the Variability of No of Week Nights in Booking Status



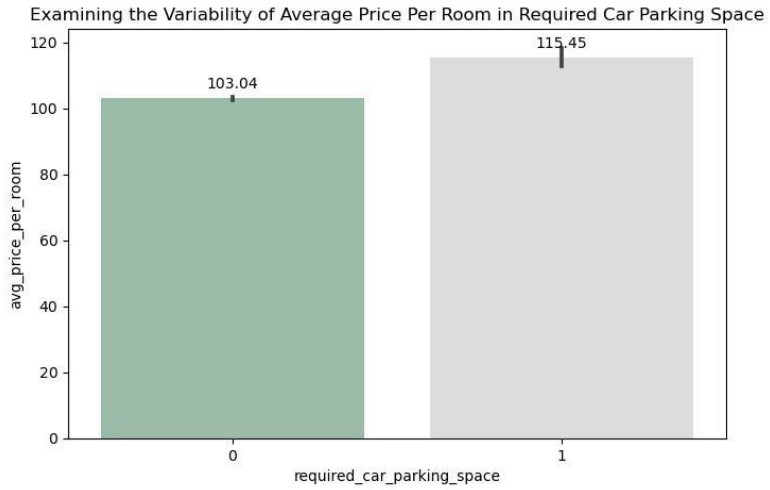

**Inference:** This plot can help us visualize how booking status varies with the length of the stay. For example, we might find that bookings with longer weeknight stays are more likely to result in a cancellation

## Describing the Count of Type of Meal Plan by Booking Status



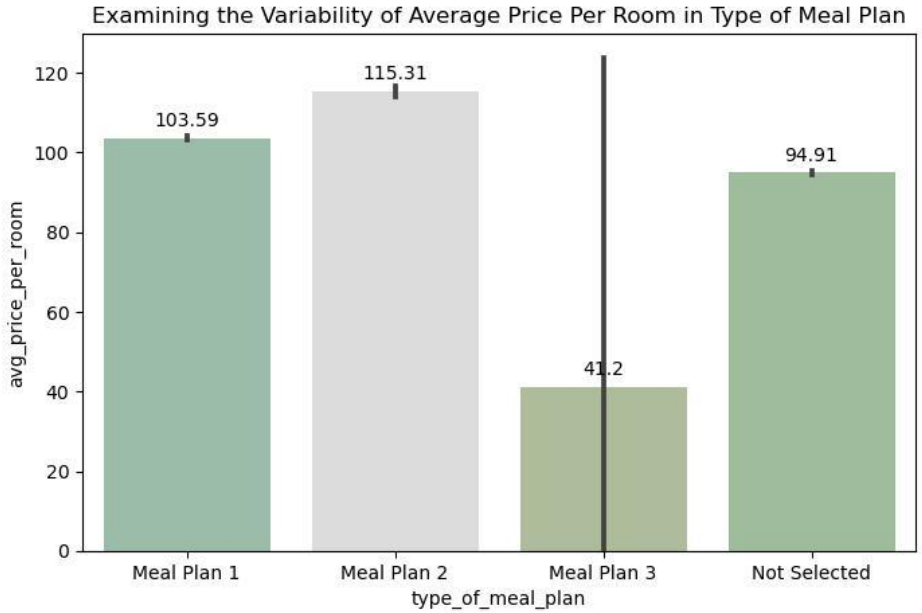

**Inference:** The meal plan 3 has merely been chosen by the customers, meal plan 1 is the most frequently asked for type.

## Examining the Variability of Average Price per Room in Required Car Parking Space



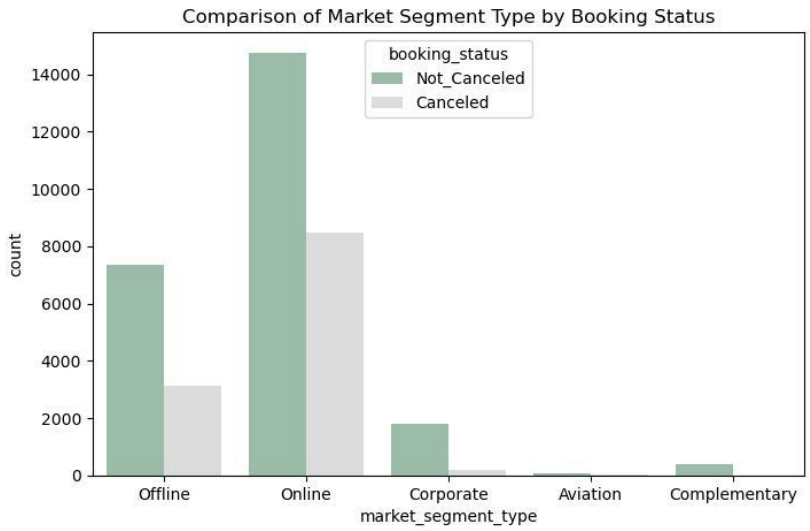

**Inference:** The average price of the room is significantly higher for those with the request of car parking space

### Examining the Variability of Average Price per Room in Type of Meal Plan




**Inference:** Average price of meal plan 2 category is the highest of all and meal plan 1 and not selected meal plan has almost similar pricing pattern, whereas the meal plan 3 has the lowest price

## Comparison of Market Segment Type by Booking Status

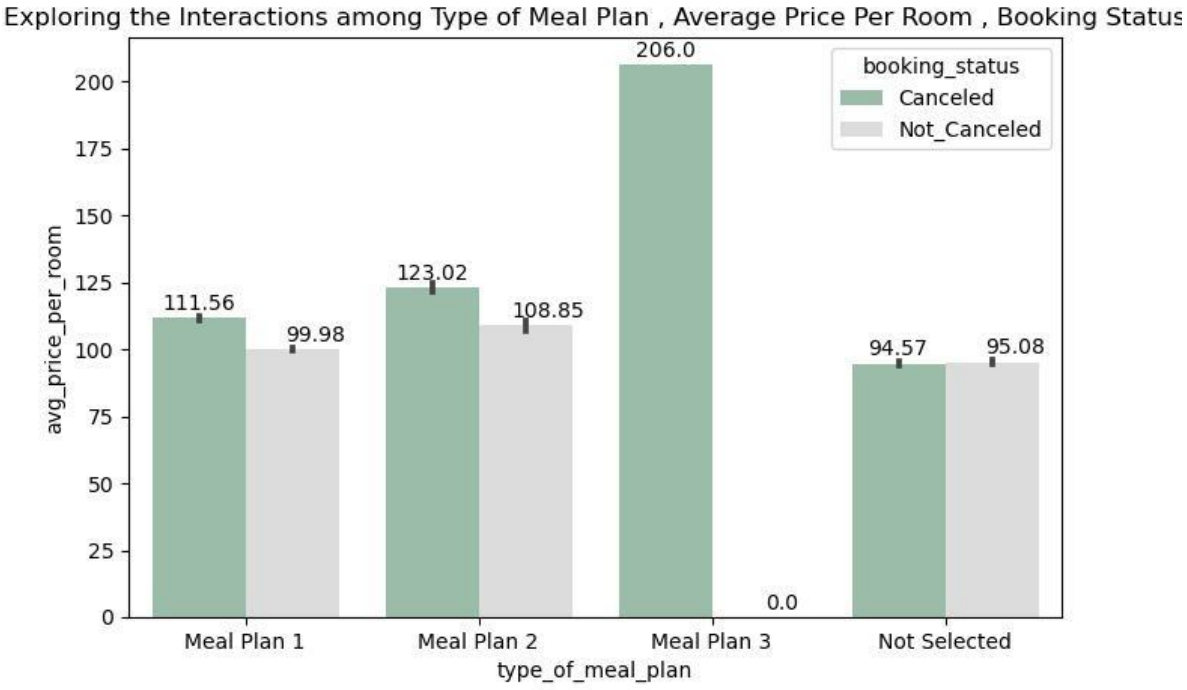



## Inference:

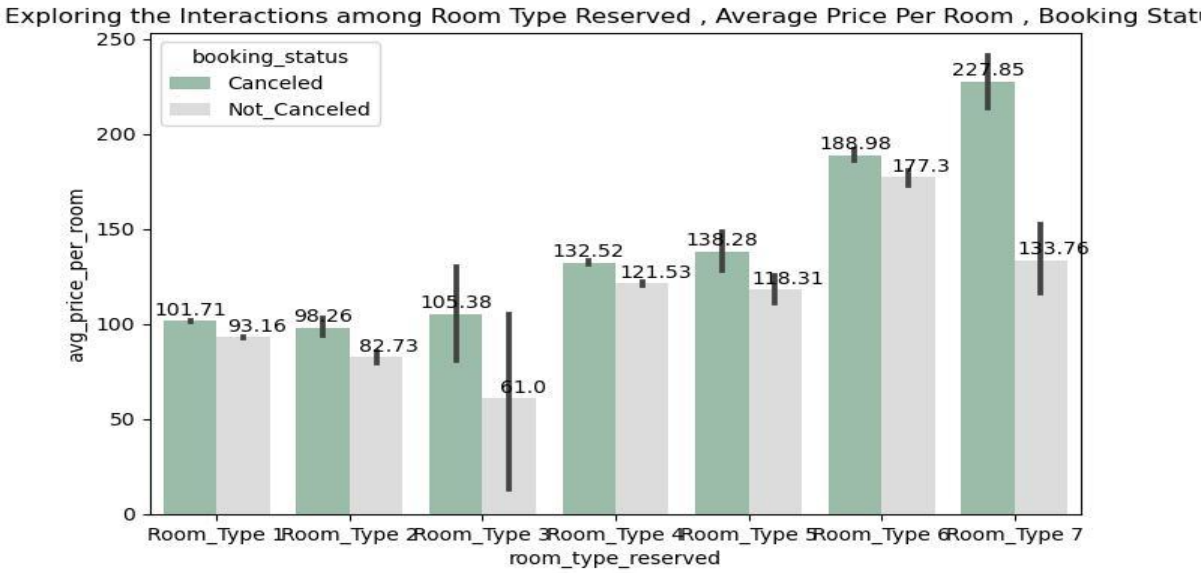
Out of all the markets the online has the highest booking as well as the cancellations, while offline bookings have a better non cancelled proportion

### 2.2.3 Multivariate Analysis:

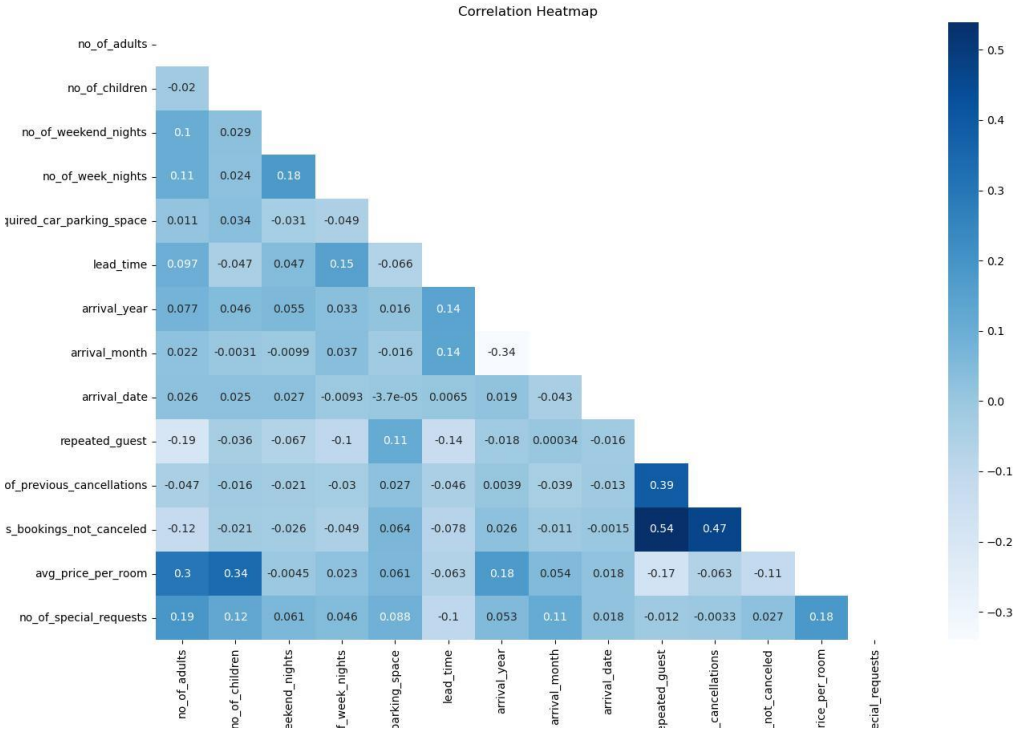
**Exploring the Interactions among Type of Meal Plan, Average Price per Room, Booking Status**



**Exploring the Interactions among Room Type Reserved, Average Price per Room, Booking Status**




# Heat map:




**Inference**: From the correlation matrix it is found that not much of the independent variable has correlation with each other. Highest correlation exist between "no of previous bookings not cancelled, and repeated guest, no of previous cancellations"

# VIF

| | Predictor | VIF |
|---|---|---|
| 6 | no_of_previous_bookings_not_canceled | 1.298973 |
| 7 | avg_price_per_room | 1.258702 |
| 0 | no_of_adults | 1.212540 |
| 5 | no_of_previous_cancellations | 1.207131 |
| 4 | lead_time | 1.137136 |
| 1 | no_of_children | 1.115342 |
| 8 | no_of_special_requests | 1.082633 |
| 3 | no_of_week_nights | 1.073780 |
| 2 | no_of_weekend_nights | 1.027494 |

## Inference

From the VIF it is seen that there is no much multi-collinearity between the numerical variables. As of now it is not required to drop any of the variables.

# Statistical Tests:

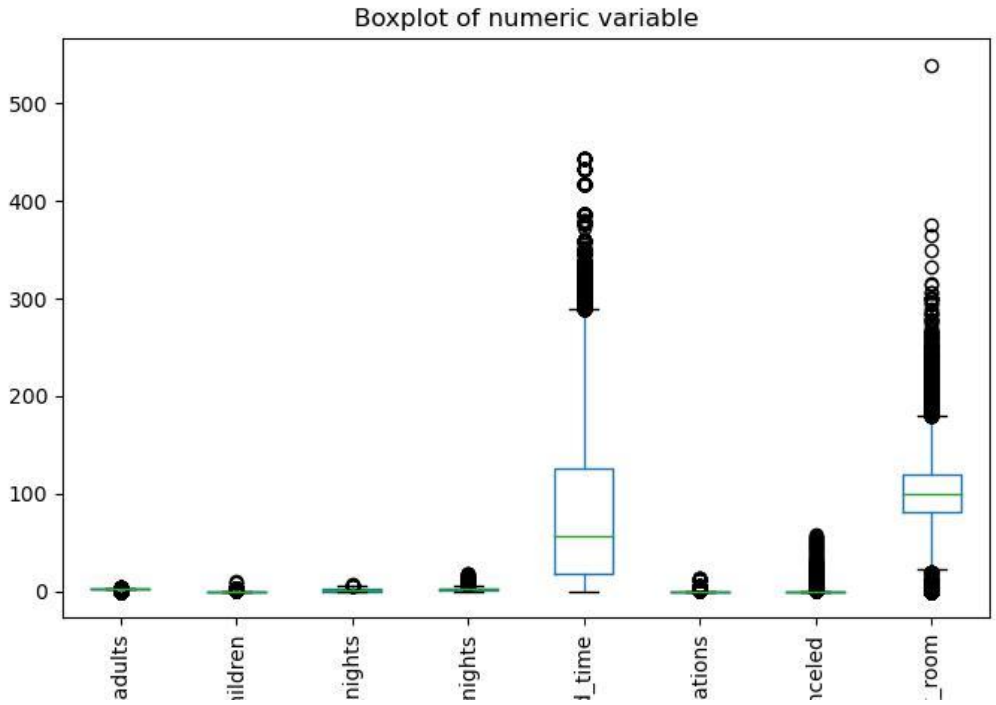
| Testing the relationship of every feature with the predicted variable using statistical tests | | | | | |
|---|---|---|---|---|---|
| **S.No** | **Statistical Test** | **Type of Variables** | **Feature Name** | **P Val** | **Inference** |
| 1 | f_oneway | Num Vs Cat | no_of_adults | 0 | Significant |
| 2 | f_oneway | Num Vs Cat | no_of_children | 0 | Significant |
| 3 | f_oneway | Num Vs Cat | no_of_weekend_nights | 0 | Significant |
| 4 | f_oneway | Num Vs Cat | no_of_week_nights | 0 | Significant |
| 5 | chi2_contingency | Cat Vs Cat | type_of_meal_plan | 0 | Significant |
| 6 | chi2_contingency | Cat Vs Cat | required_car_parking_space | 0 | Significant |
| 7 | chi2_contingency | Cat Vs Cat | room_type_reserved | 0 | Significant |
| 8 | f_oneway | Num Vs Cat | lead_time | 0 | Significant |
| 9 | chi2_contingency | Cat Vs Cat | market_segment_type | 0 | Significant |
| 10 | chi2_contingency | Cat Vs Cat | repeated_guest | 0 | Significant |
| 11 | f_oneway | Num Vs Cat | no_of_prev_cancellations | 0 | Significant |
| 12 | f_oneway | Num Vs Cat | no_of_prev_bookings_not_canceled | 0 | Significant |
| 13 | f_oneway | Num Vs Cat | avg_price_per_room | 0 | Significant |
| 14 | f_oneway | Num Vs Cat | no_of_special_requests | 0 | Significant |

# Inference:

From above performed statistical tests we can conclude that none of the columns has failed to reject null hypotheses which means none of the columns are insignificant to the target variable booking_status. At this stage we can't drop any variable. On further progress after building few models and checking for their metrics if the performance is considerably low we can drop some columns based on the feature_importance score.

Also the arrival date, arrival month, arrival year are the date features are not tested with the statistical tests which are held for feature engineering. However, for base model none of these columns were removed or treated.

# Outlier and distribution of the numerical variable:




**Inference:** Most of the numerical variable has outliers and are skewed. Since IQR treatment will result in loss of data, we choose to perform power transformation for the data to treat the skewness before introducing the data to the model.

# Base Model Summary

```
print(model_lr.summary())
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25374
Method:                           MLE   Df Model:                           17
Date:                Thu, 18 May 2023   Pseudo R-squ.:                 0.3090
Time:                        09:39:40   Log-Likelihood:                -11084.
converged:                      False   LL-Null:                       -16041.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                         -1175.2949    118.305     -9.934      0.000   -1407.168    -943.422
no_of_adults                     -0.0251      0.019     -1.332      0.183      -0.062       0.012
no_of_children                   -0.0111      0.018     -0.624      0.533      -0.046       0.024
no_of_weekend_nights              0.0370      0.017      2.185      0.029       0.004       0.070
no_of_week_nights                -0.0446      0.018     -2.536      0.011      -0.079      -0.010
type_of_meal_plan                 0.0914      0.040      2.265      0.024       0.012       0.171
required_car_parking_space       -1.4355      0.135    -10.655      0.000      -1.700      -1.171
room_type_reserved                0.1681      0.021      7.947      0.000       0.127       0.210
lead_time                         1.3424      0.024     56.831      0.000       1.296       1.389
arrival_year                      0.5820      0.059      9.927      0.000       0.467       0.697
arrival_month                    -0.0344      0.006     -5.439      0.000      -0.047      -0.022
arrival_date                      0.0049      0.002      2.570      0.010       0.001       0.009
market_segment_type              -0.7225      0.023    -31.850      0.000      -0.767      -0.678
repeated_guest                  -13.1832    114.101     -0.116      0.908    -236.817     210.450
no_of_previous_cancellations      1.1475     10.962      0.105      0.917     -20.338      22.633
no_of_previous_bookings_not_canceled  0.2079  0.130      1.598      0.110      -0.047       0.463
avg_price_per_room                0.6652      0.026     25.466      0.000       0.614       0.716
no_of_special_requests           -1.0254      0.020    -50.797      0.000      -1.065      -0.986
=================================================================================================
```

## Classification Report for train and test data

```
print(f'Train report : \n{classification_report(ytrain,y_pred_train)}\n')
print(f'Testreport : \n{classification_report(ytest,y_pred_test)}')
```

```
Train report :
              precision    recall  f1-score   support

           0       0.80      0.93      0.86     17099
           1       0.79      0.52      0.63      8293

    accuracy                           0.80     25392
   macro avg       0.79      0.73      0.74     25392
weighted avg       0.80      0.80      0.78     25392


Testreport :
              precision    recall  f1-score   support

           0       0.80      0.93      0.86      7291
           1       0.79      0.52      0.63      3592

    accuracy                           0.80     10883
   macro avg       0.80      0.73      0.75     10883
weighted avg       0.80      0.80      0.78     10883
```

# Summary:

From above report we can conclude that our base model has performed well in both train and unseen data with accuracy of almost 80%. On further progress we try to improve our performance by building other models, tuning their hyper parameters and selecting columns based on feature importance score

The target variable has class imbalance, therefore there is further scope of improvement I the recall and precision.

# Scope for improvement:

## Feature Engineering:

**arrival_date** : It can be segregated to weekday and weekend and do further analysis, on how the variable behaves with the target

**arrival_month** : The arrival month can be segregated in to different seasons and then check the behavior with the target variable

# Model Improvement:

The model is built using the statistic models, to further improve the model we would boost the model using xgboost. We would further improve the model using the Decision tree, random forest algorithms to improve the model performance.

The Feature selection techniques like recursive feature selection, sequential feature selection and lasso to be performed to understand the import feature. This will help us in giving the best business recommendation.

The hyper parameter tuning can also be used extensively to understand the model performance at each level of hyper parameters.