HOTEL RESERVATION CASE STUDY

Submitted by: Eshah Arshad Khan

TABLE OF CONTENTS

Introduction

Summary

This study looks at using data analysis to predict customer actions and chances of them canceling hotel bookings. It uses hotel booking data from 'INN Hotels.' The work involves cleaning the data, doing a detailed analysis (EDA), and using two methods called K-Nearest Neighbors (KNN) and Decision Tree to make predictions. The study deals with fixing data issues like errors, missing information, and uneven data, using a technique called SMOTE. The results show that these methods are good at guessing booking cancellations, showing how useful data analysis can be for managing hotels. This report also carries 15 questions from our study to ensure better understanding of the dataset at hand.

Importance

Predictive analytics is very important for hotels because it helps them understand customers behavior. By looking at past information, hotels can guess if people will cancel their bookings, what they might want, and when they are likely to visit. This helps hotels plan better, offer services that customers like, and set prices that attract more guests. It also helps hotels fix problems before they get big. When hotels know what their customers want ahead of time, they can make their stay better, keep them happy, and encourage them to come back.

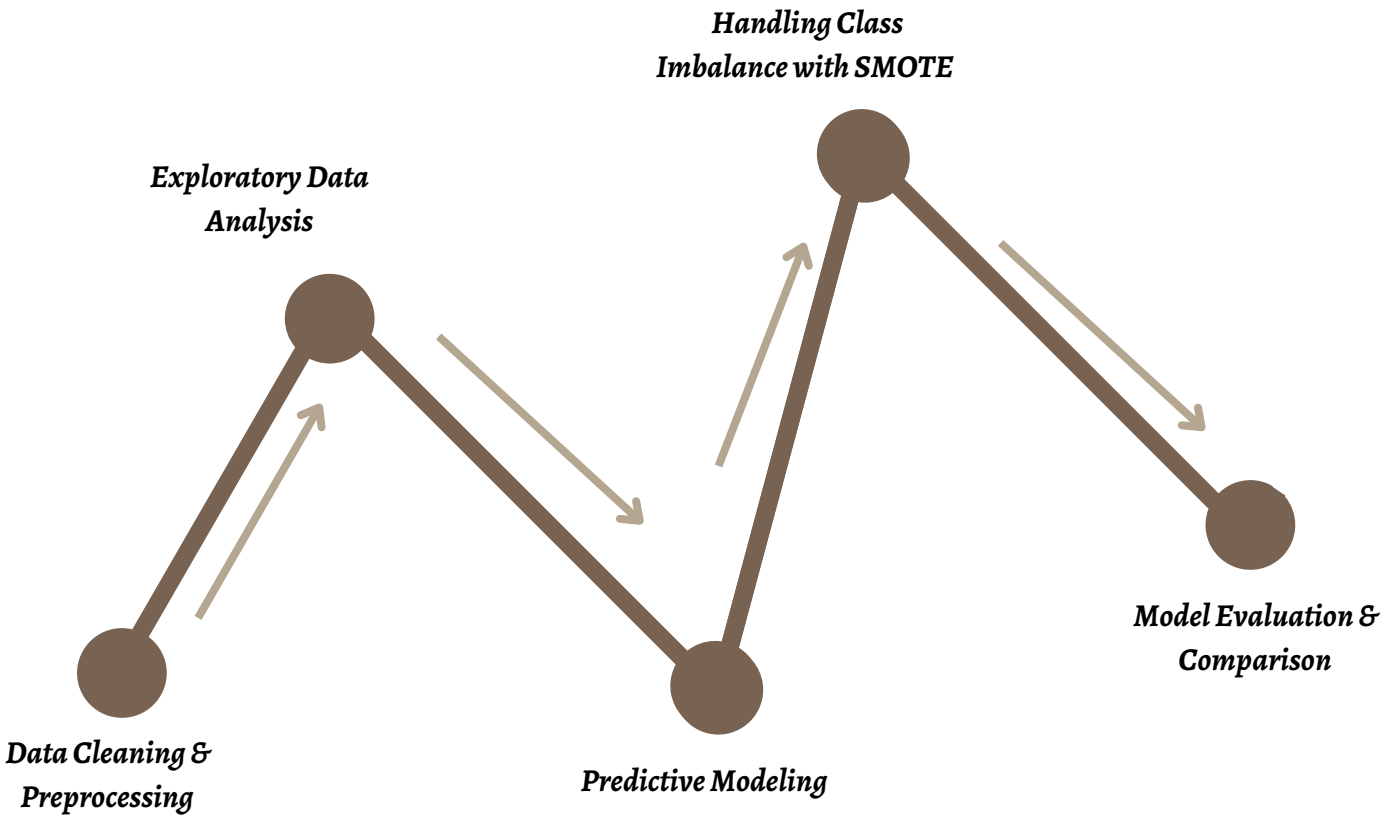
Data Overview

Data Dictionary

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

Methodology

With the “INN Hotel” dataset we have a classification problem at hand. To make a successful predictive model the following road map was implimented.

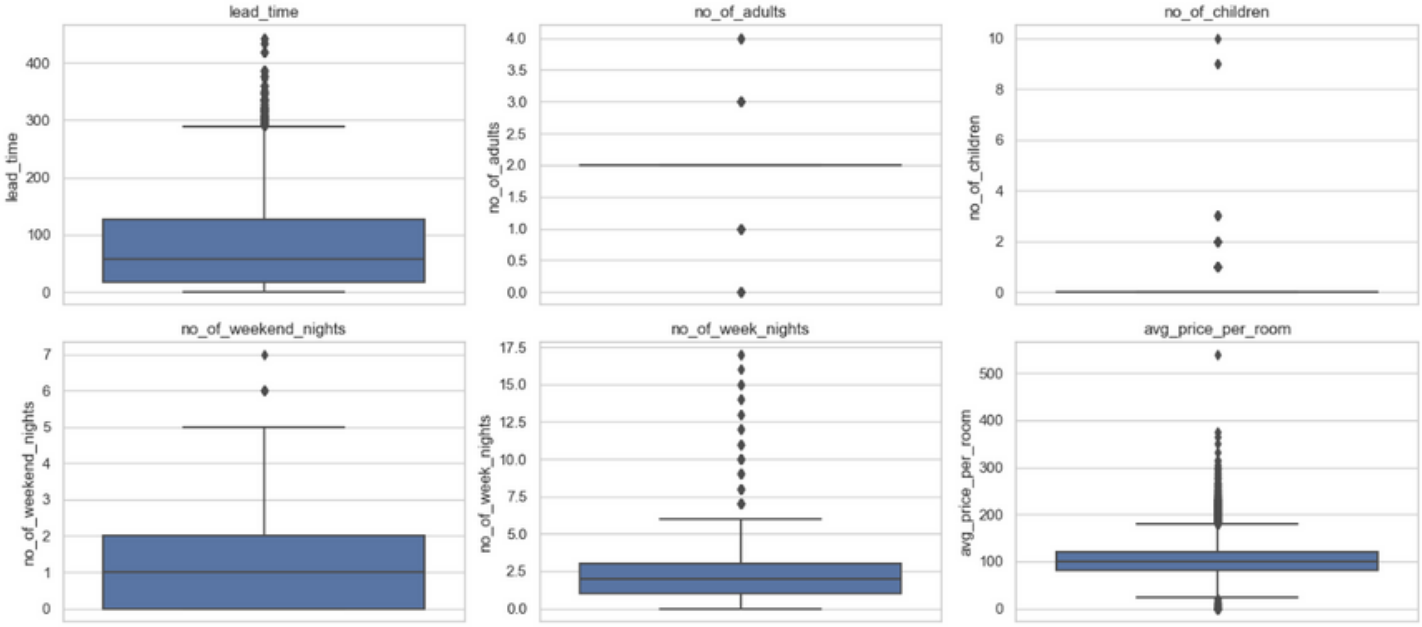


- **Data Cleaning and Preprocessing:** The dataset was checked for duplicates and missing values, with outliers assessed and retained for accuracy.
- **Exploratory Data Analysis (EDA):** Insights into customer behaviors and booking trends were uncovered through statistical analysis and visualizations.
- **Predictive Modeling:** K-Nearest Neighbors (KNN) and Decision Trees were used for their effectiveness in predicting booking cancellations.
- **Handling Class Imbalance with SMOTE:** Class imbalances were corrected using SMOTE to ensure model accuracy.
- **Model Evaluation and Comparison:** Models were evaluated and compared using key metrics to assess the impact of balancing the dataset.

Data Cleaning and Pre-Processing

The dataset was found to be free of duplicates and missing values, making it complete and ready for analysis.

Box plots were used to look for any unusual data points (outliers)



To work these outliers out, correlation analysis for each of these columns was carried out against other features. which showed results which indicated that the values were indeed real and fit the context properly.

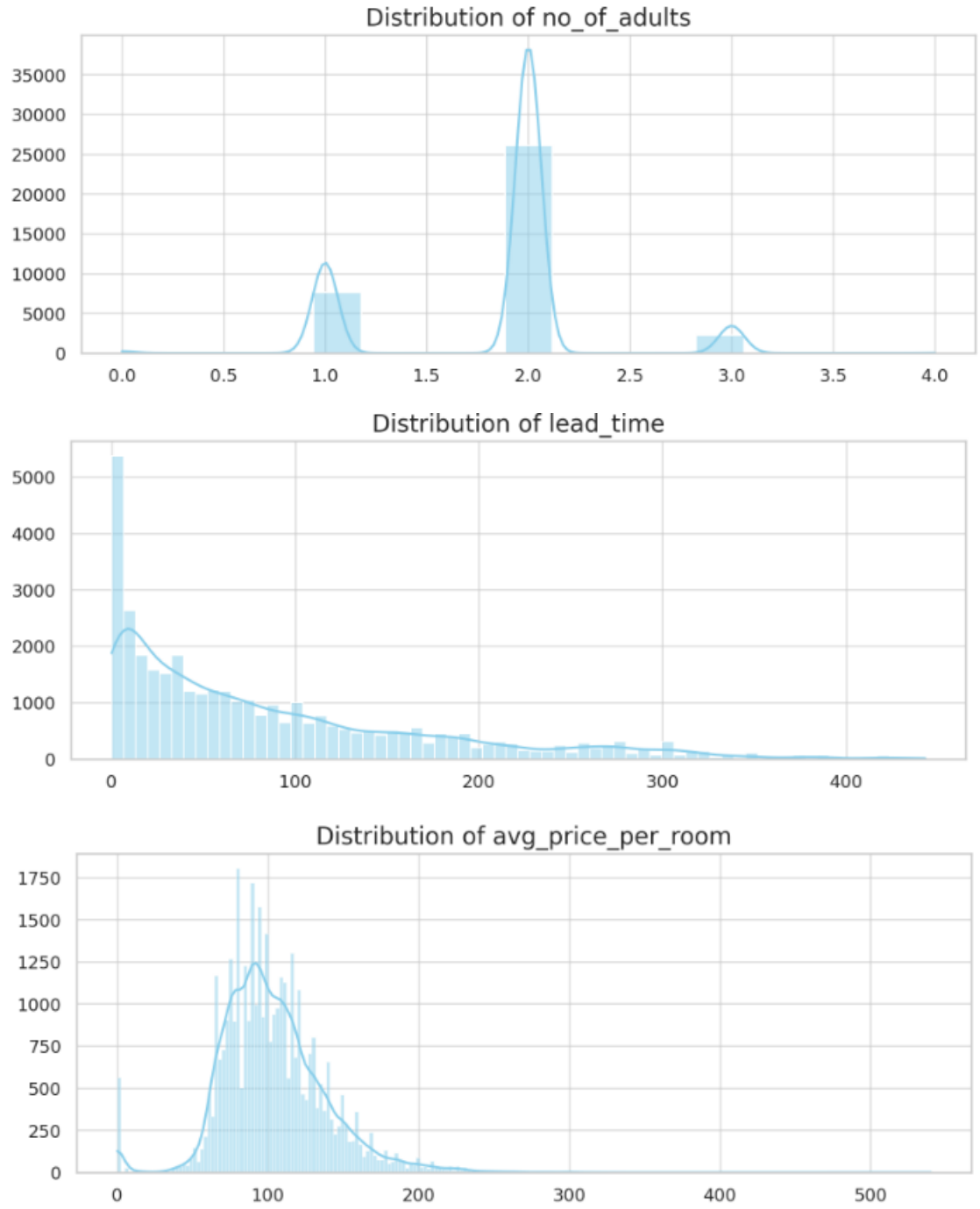
For Examples for “avg_price_per_room” the correlation analysis against other columns gave the results shown below. The moderate positive correlations with the number of adults and children suggest that larger room types or family bookings might have higher prices.

The slight positive correlation with no_of_special_requests could indicate that bookings with more amenities or services tend to be more expensive. The negative correlations, although weak, hint that certain factors like being a repeated guest or having previous bookings might be associated with lower average prices, perhaps due to loyalty discounts or familiarity with cost-effective booking strategies. Given these correlations, it seems that the high prices in avg_price_per_room are associated with specific booking characteristics. This insight suggests that the outliers in room prices are worth keeping for further analysis, especially in predictive modeling

avg_price_per_room	1.000000
no_of_children	0.337728
no_of_adults	0.296886
no_of_special_requests	0.184381
arrival_year	0.178605
required_car_parking_space	0.061304
arrival_month	0.054423
no_of_week_nights	0.022753
lead_time	-0.062596
no_of_previous_cancellations	-0.063340
no_of_previous_bookings_not_canceled	-0.113684
repeated_guest	-0.174900

Data Distribution and Encoding

The Dataset contains Categorical and Nominal data. This data needs to be encoded into numerical form for it to be fit for ML modeling, Categorical data is One hot encoded and Target Variable is label encoded, while rest of the numerical data stays as it is.



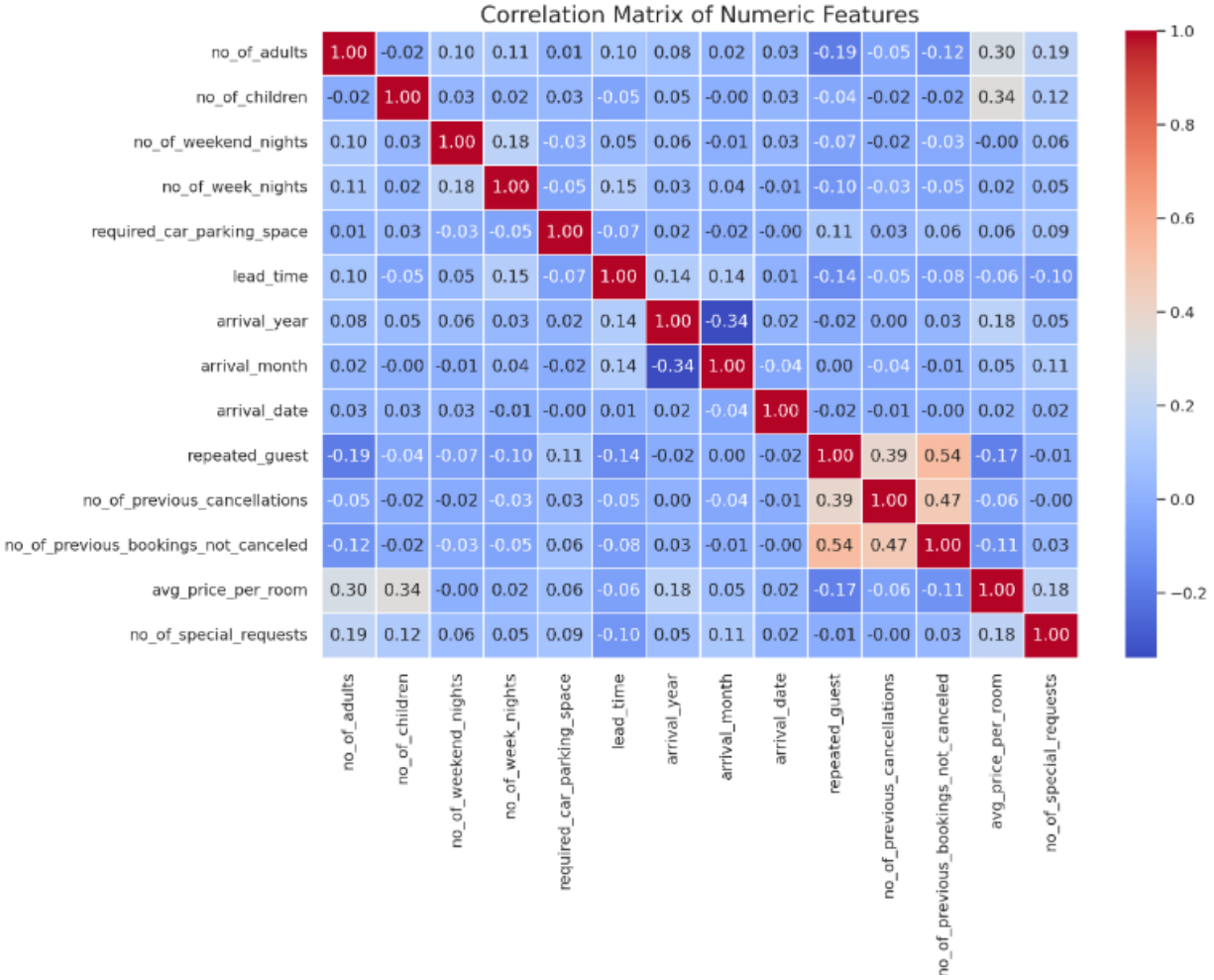
The Features such as 'type_of_meal_plan', 'room_type_reserved', 'market_segment_type' were one hot encoded and target variable "booking_status" was label encoded. With that done we will head on to ML Modeling

Exploratory Data Analysis ■

Exploratory Data Analysis (EDA) is a key step in understanding a dataset. It involves looking for patterns, and finding how variables relate to each other using statistics and graphs. EDA helps in making informed guesses and choosing the right methods for deeper analysis.

Correlation Analysis:

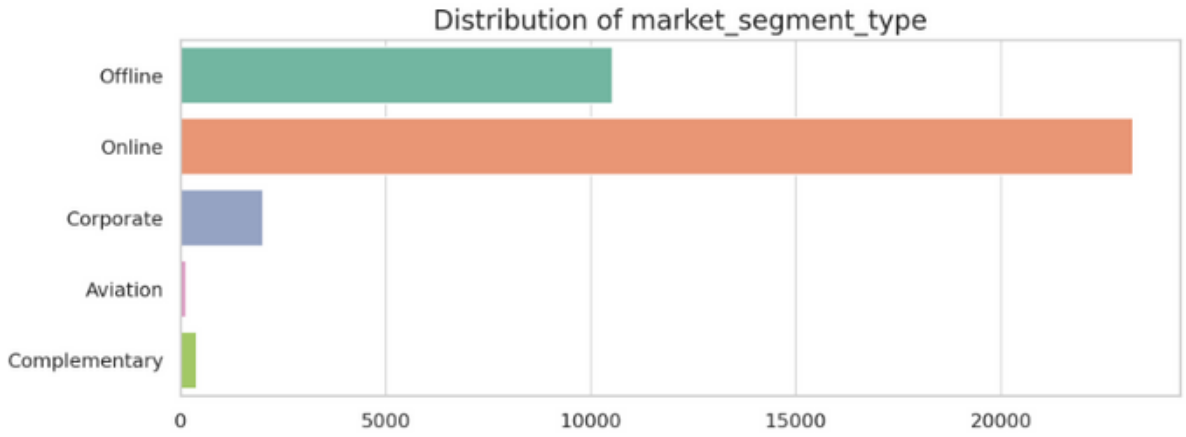
Examining the relationships between numerical variables to identify any strong correlations, which could indicate a significant connection affecting the target variable



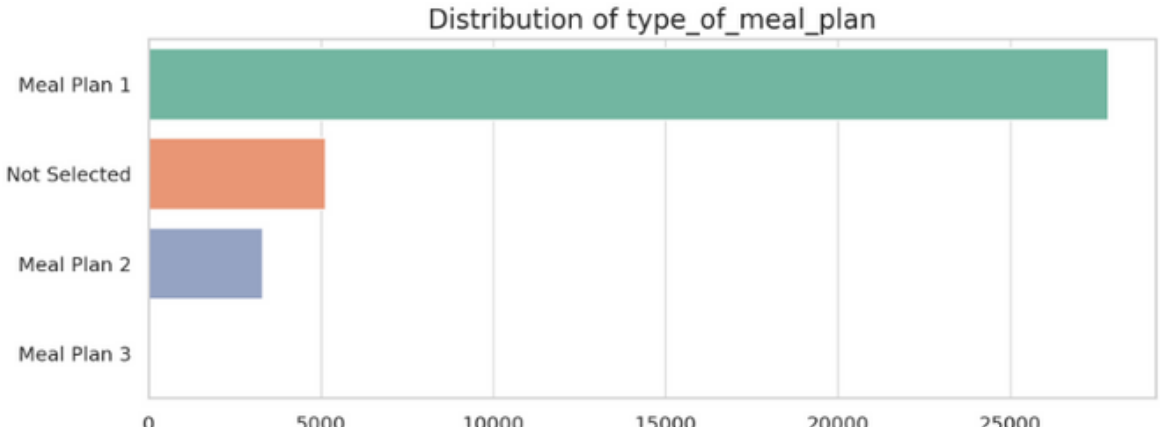
repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled) show low numerical correlation with other features. The impact of being a repeated guest on other factors like pricing might require deeper categorical analysis. no_of_special_requests has very low correlation with other numerical features, so demand for special requests is relatively independent of the other numerical factors considered here.

Categorical Data Analysis

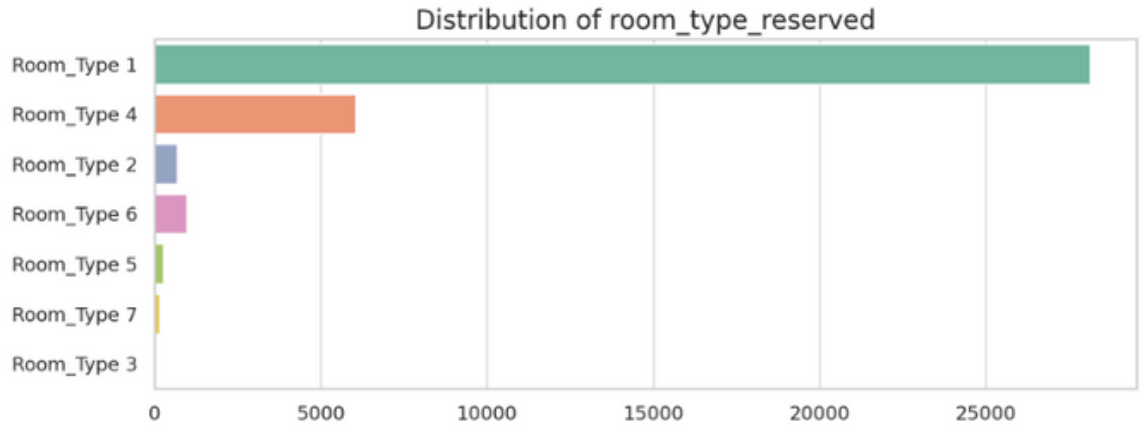
Analyzing the distribution and impact of categorical variables, such as room type or customer type, on the target variable. Here are few shown below:



- The Offline market segment appears to have the highest number of bookings, indicating that a significant portion of the hotel's business comes through offline channels.
- The Online market segment is the second most frequent, which suggests that online bookings are also a major contributor to the hotel's occupancy.
- The Corporate and Aviation segments have noticeably fewer bookings, which might imply that the hotel is not primarily targeting or attracting these segments.
- The Complementary segment has the fewest bookings, which could be expected as these would typically represent promotional or complimentary stays.



- The Not Selected category has the highest frequency, suggesting that many guests opt out of pre-booking meal plans. This could indicate a preference for flexibility, external dining options, or the possibility that guests are not finding the meal plan offerings attractive.
- Meal Plan 1 is the most popular among those who do select a meal plan. The hotel might consider focusing on this meal plan or exploring why it's popular to further improve their offerings.
- Meal Plan 2 and Meal Plan 3 are less commonly selected, which could suggest less value perceived by the guests or a mismatch in the offerings of these plans with guest preferences.



- Room_Type 1 has the highest number of bookings, making it the most popular choice among guests. This could be due to various factors such as cost, amenities, or availability.
- Room_Type 3 is the least popular, which might indicate a lack of demand for this type of room or a higher price point that is not as attractive to guests.
- Room_Type 2, Room_Type 4, Room_Type 5, Room_Type 6, and Room_Type 7 have a moderate number of bookings, with Room_Type 4 being slightly more popular than the others.

Predictive Modelling

In the predictive modeling phase, the focus shifts to using machine learning algorithms to forecast hotel booking cancellations. Following exploratory data analysis (EDA), K-Nearest Neighbors (KNN) and Decision Tree models are applied, selected for their proven effectiveness in classification tasks. This crucial step turns data insights into actionable predictions, aiming to enhance hotel booking management strategies.

Decision Tree

Decision tree model didn't require standardization or normalization so it was implemented first. The testing and training sets were split with the standard ratio. after training the model on training set and then plugging in the testing set we obtained the following classification report:

	precision	recall	f1-score	support
0	0.80	0.74	0.77	3607
1	0.88	0.91	0.89	7276
accuracy			0.85	10883
macro avg	0.84	0.82	0.83	10883
weighted avg	0.85	0.85	0.85	10883

we _ _

KNN

For KNN there was no need for standardization but there is the need for normalization. MinMaxScaler was used for normalization of our dataset. with that done, KNN model was applied onto the dataset and the following classification report was obtained:

	precision	recall	f1-score	support
0	0.80	0.74	0.77	3607
1	0.88	0.91	0.89	7276
accuracy			0.85	10883
macro avg	0.84	0.82	0.83	10883
weighted avg	0.85	0.85	0.85	10883

Handling Class Imbalance with SMOTE

Handling class imbalance is a critical step in ensuring the accuracy of predictive models, especially in scenarios where the outcome classes are not equally represented. The Synthetic Minority Over-sampling Technique (SMOTE) is a popular method used to address this issue. It works by creating synthetic examples of the minority class, effectively balancing the dataset without losing important information. By applying SMOTE, the predictive models are trained on a dataset that mirrors a more realistic scenario, enhancing their ability to generalize and make accurate predictions. This technique is particularly beneficial in the context of hotel booking cancellations, where the proportion of cancellations to non-cancellations can significantly influence model performance.

After Applying SMOTE to the dataset and applying the models to the SMOTE trained dataset, the following classification reports were obtained for the models.

Decision Tree- SMOTE

	precision	recall	f1-score	support
0	0.19	0.08	0.12	3607
1	0.64	0.82	0.72	7276
accuracy			0.58	10883
macro avg	0.42	0.45	0.42	10883
weighted avg	0.49	0.58	0.52	10883

KNN- SMOTE

	precision	recall	f1-score	support
0	0.71	0.83	0.77	3607
1	0.91	0.83	0.87	7276
accuracy			0.83	10883
macro avg	0.81	0.83	0.82	10883
weighted avg	0.84	0.83	0.84	10883

The decrease in accuracy for both Decision Tree and KNN models after applying SMOTE suggests that these models are now giving more attention to the minority class. The Decision Tree's accuracy dropping from 86% to 58% indicates a high sensitivity to class imbalance. This model might have initially overfitted to the majority class. The KNN model's performance decrease from 85% to 83% is relatively smaller, suggesting it might be more robust to class imbalance compared to the Decision Tree.

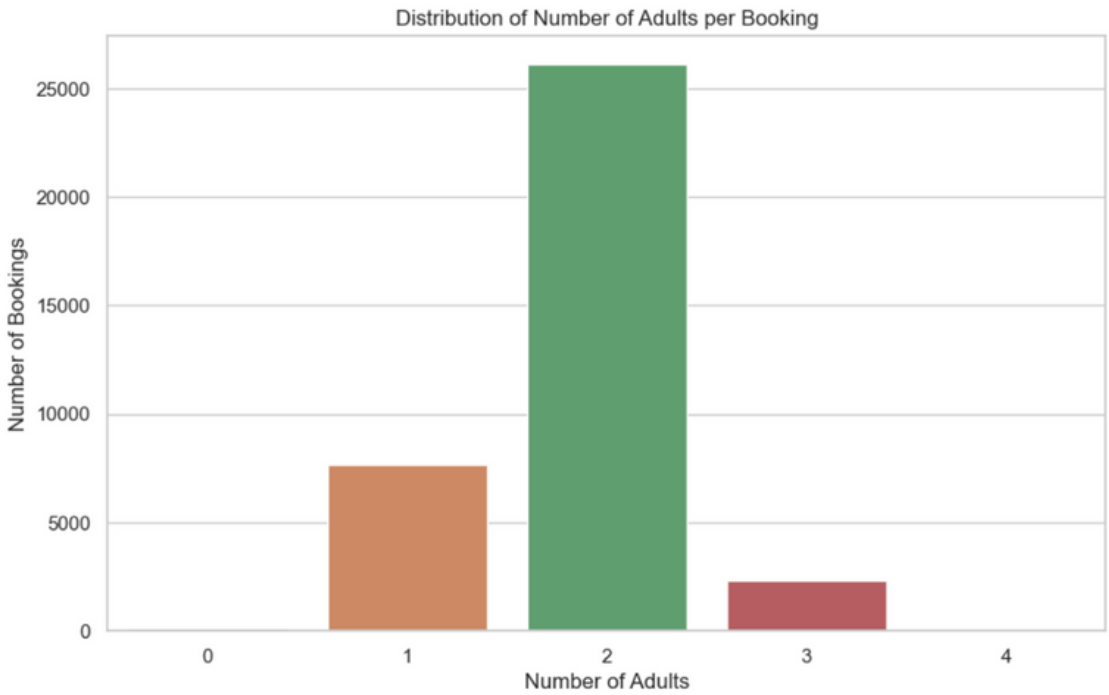
Results and Conclusion

The decrease in accuracy for both Decision Tree and KNN models after applying SMOTE suggests that these models are now giving more attention to the minority class. The Decision Tree's accuracy dropping from 86% to 58% indicates a high sensitivity to class imbalance. This model might have initially overfitted to the majority class.

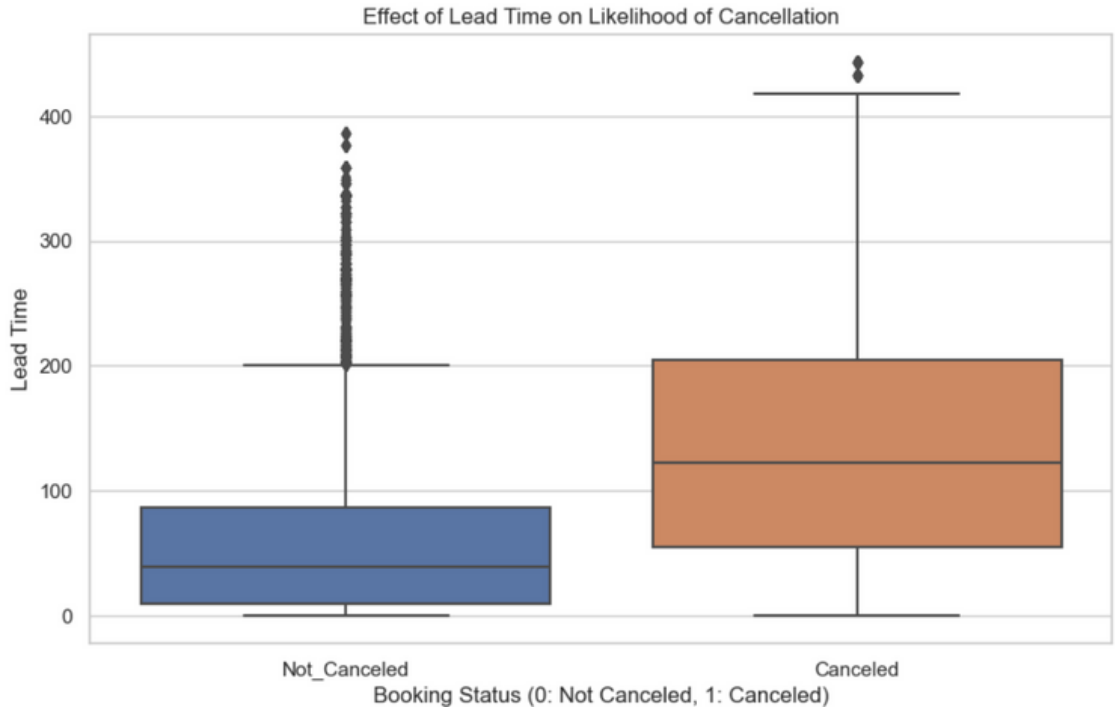
The KNN model's performance decrease from 85% to 83% is relatively smaller, suggesting it might be more robust to class imbalance compared to the Decision Tree.

Questions

Question 1: What is the distribution of the number of adults per booking?

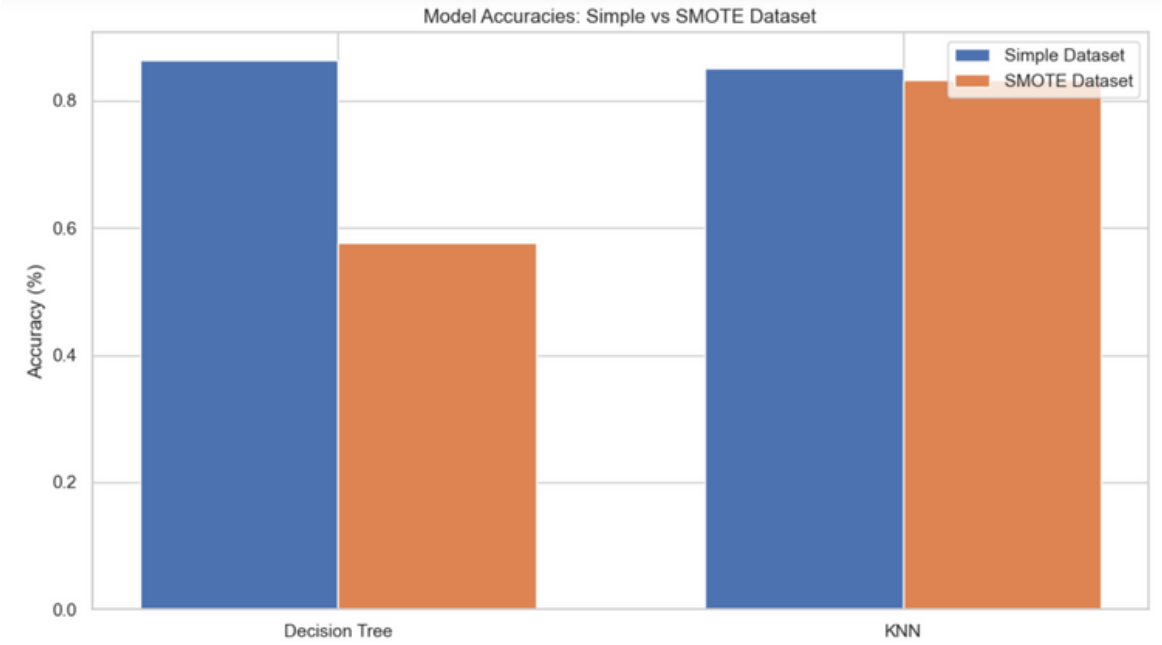


Question 1: What is the distribution of the number of adults per booking?



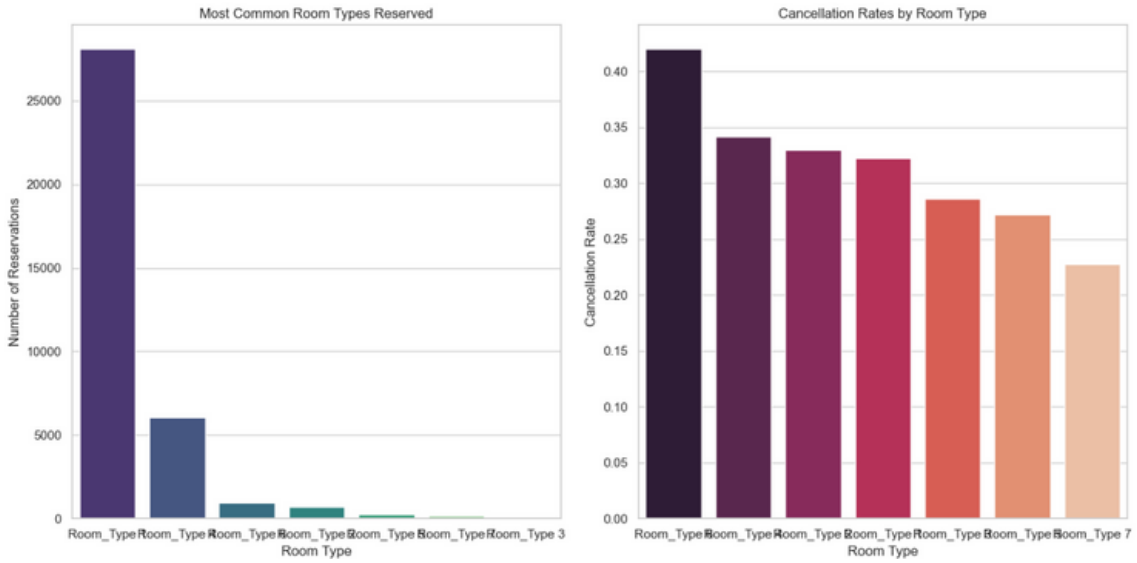
The median lead time for canceled bookings is higher than that for non-canceled bookings. This shows that bookings made well in advance are more likely to being canceled. There are outliers present in both categories, especially for canceled bookings. This shows that there are some bookings with an exceptionally high lead time that still get canceled.

Question 3: Which model has the best accuracy?



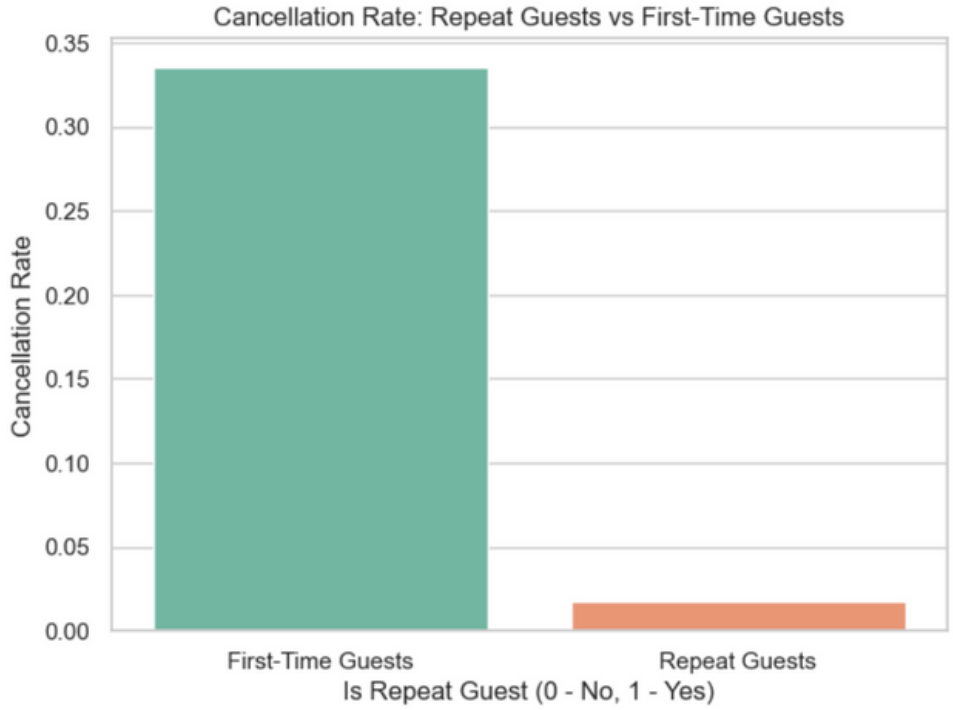
For the simple dataset Decision tree model has better accuracy however the models trained on SMOTE ensure that model doesn't incline towards the majority category. so in that context KNN's model has better accuracy.

Question 4: What are the most common room types reserved according to the dataset, and how does room type selection relate to cancellation rates?



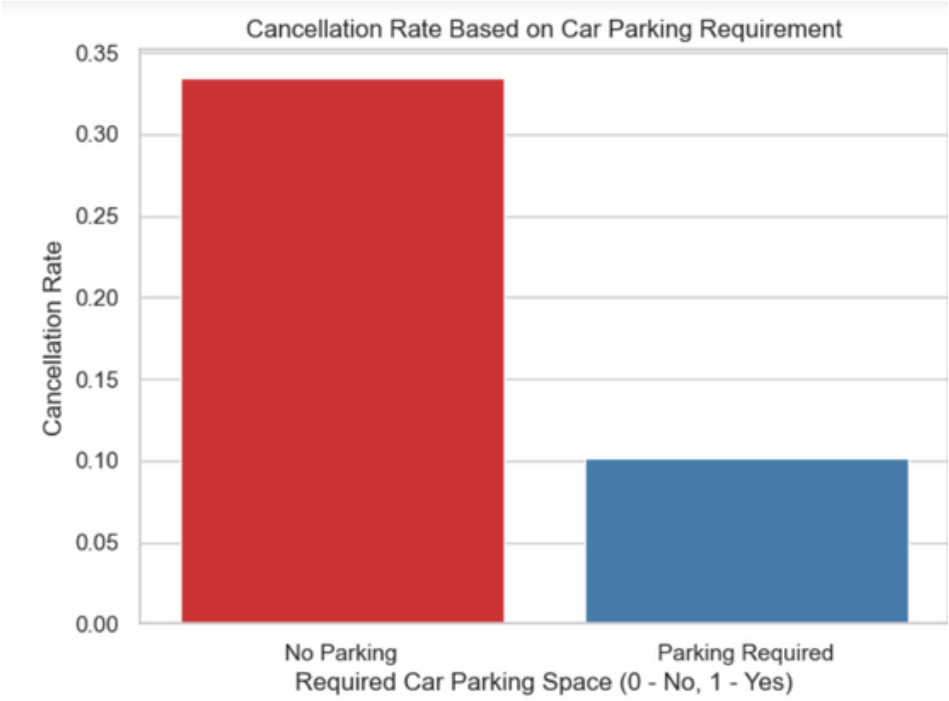
The most common room type reserved is Room_Type 1, with 25,000+ reservations. The least common is Room_Type 3. Room_Type 6 has the highest cancellation rate at approximately 40% above. Room_Type 7 has the lowest cancellation rate. Room_Type 1, despite being the most reserved, has a cancellation rate of around 32.25%, which is not the highest among the room types.

Question 5: Are repeat guests more or less likely to cancel their bookings than first-time guests? answer this. description and visual



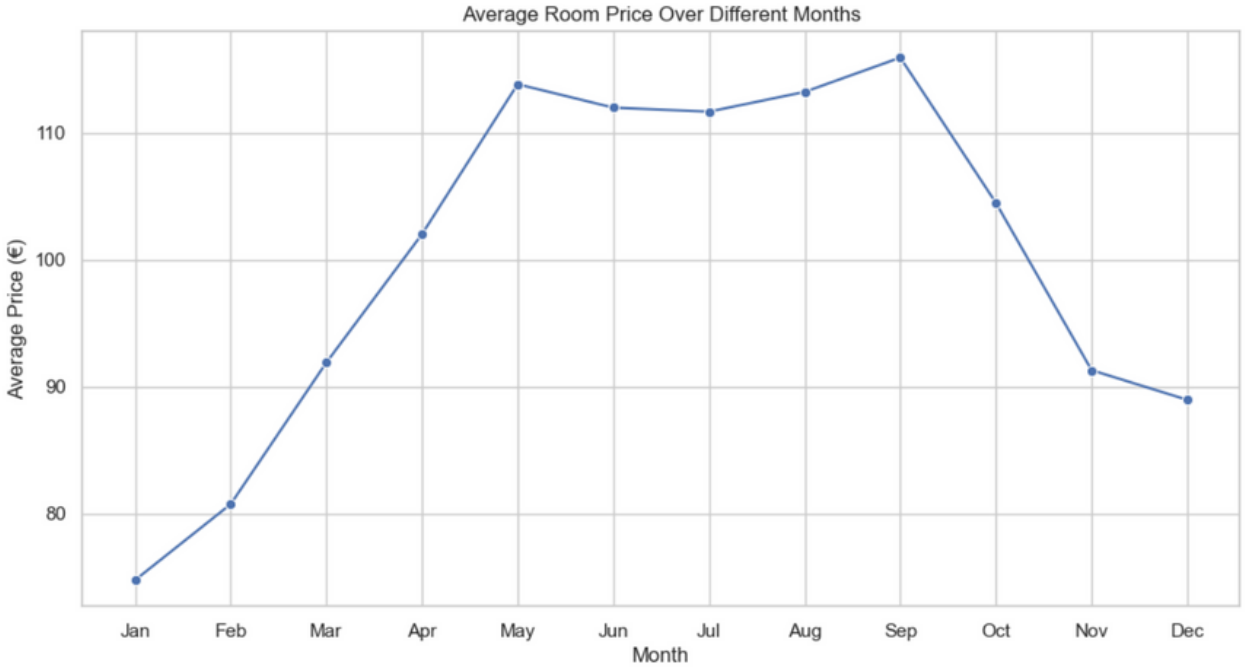
Repeat guests are significantly less likely to cancel their bookings, with a cancellation rate of approximately 1.72%. First-time guests have a much higher cancellation rate of around 33.58%.

Question 6: Can we visually analyze the proportion of bookings that required a car parking space and resulted in a cancellation?



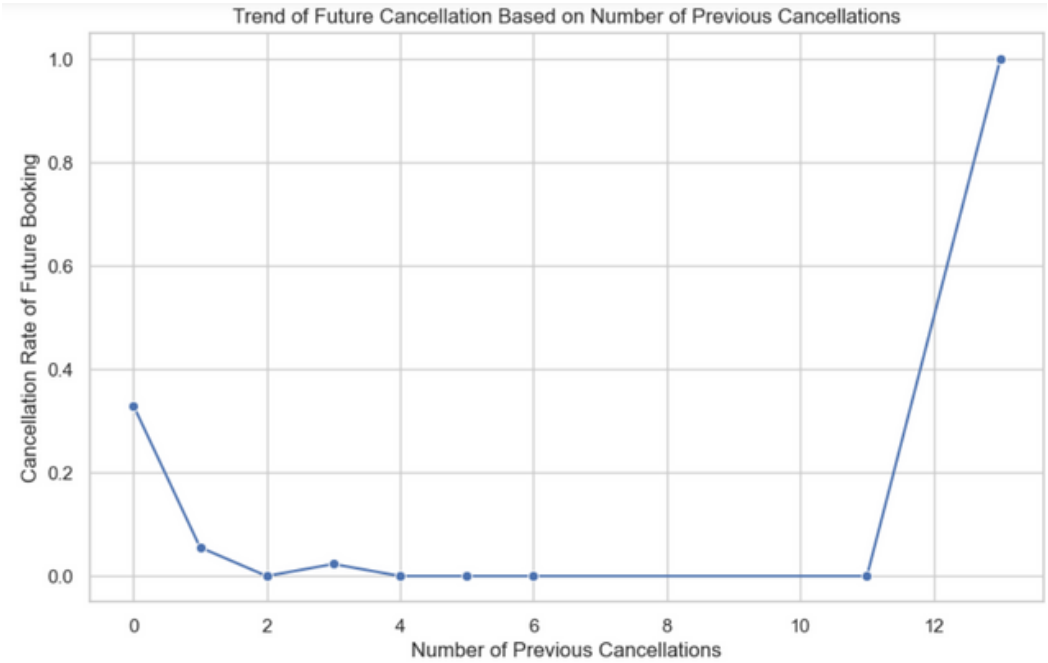
Bookings that did not require a car parking space have a higher cancellation rate. Bookings that required a car parking space have a significantly lower cancellation rate, suggesting that guests who need parking are less likely to cancel their bookings.

Question 7: What are the patterns of average room prices over different months of the year, and how can this be graphically represented?



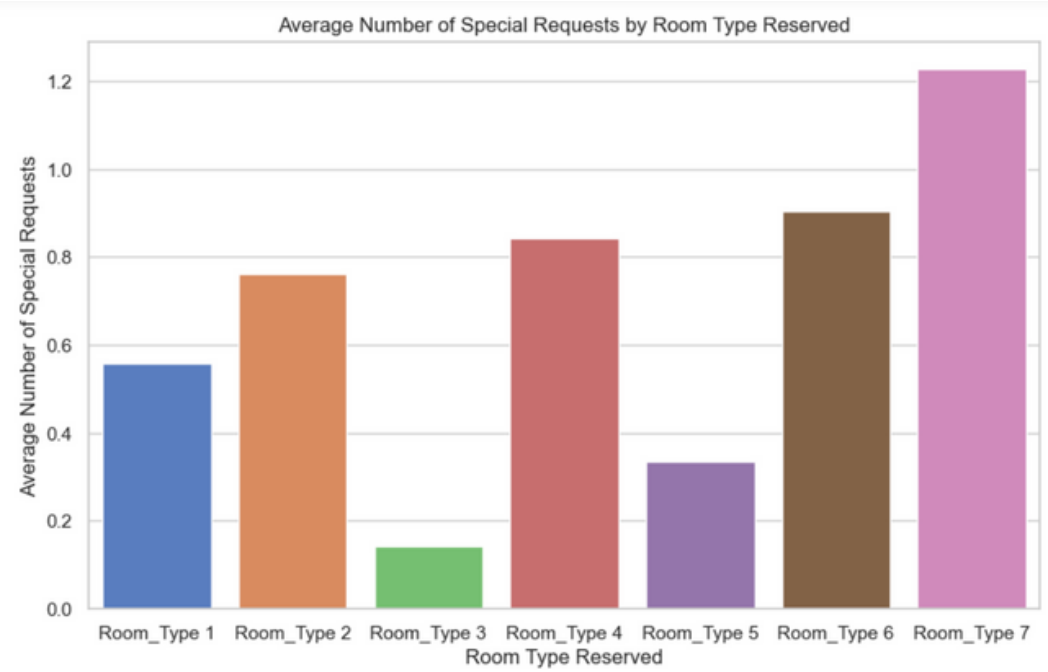
Prices tend to rise starting from January, peaking around September. The highest average price is observed in September. After September, there is a noticeable decrease in average prices towards the end of the year. Room prices are likely influenced by factors such as holidays, events, and overall demand, which tend to vary throughout the year.

Question 8: Are there any noticeable trends in the number of previous cancellations and the likelihood of a future cancellation, and how can we visualize this?



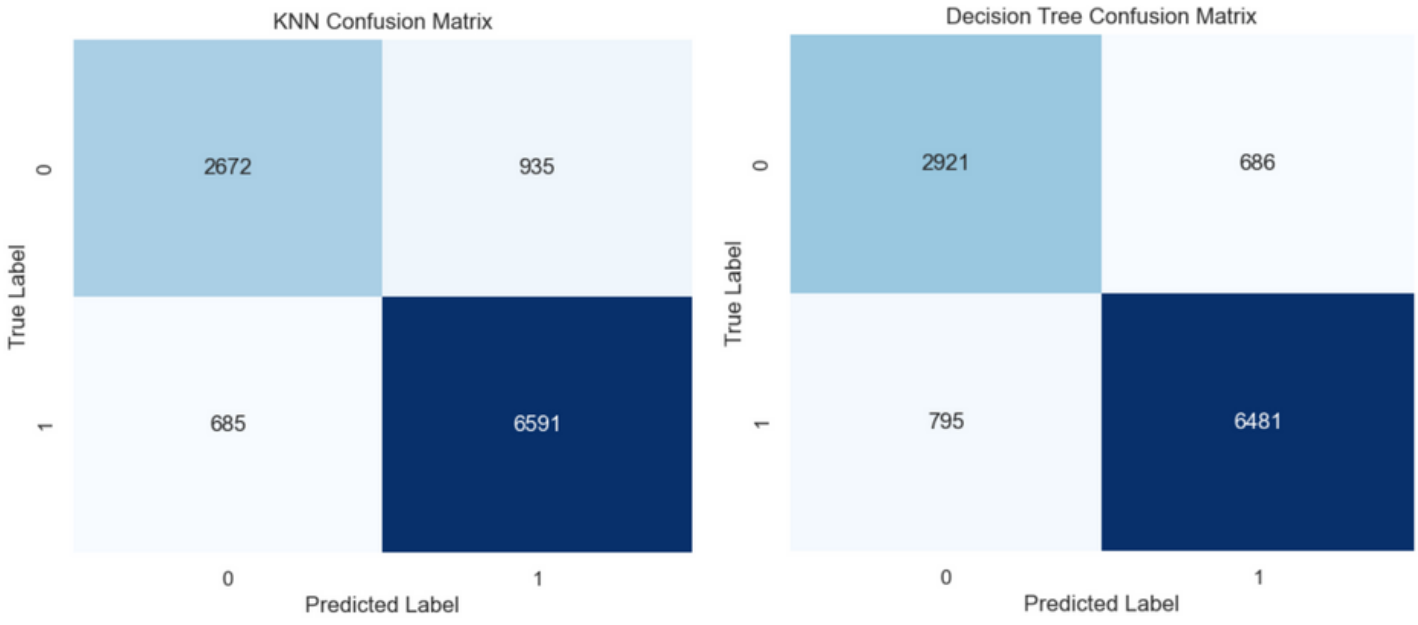
This pattern suggests that room prices are higher during the spring and early summer months, possibly due to increased demand during this period. Prices drop towards the winter months, which may indicate a lower demand or off-season pricing strategies.

Question 9: Can we create a visual comparison of the average number of special requests by room type reserved?



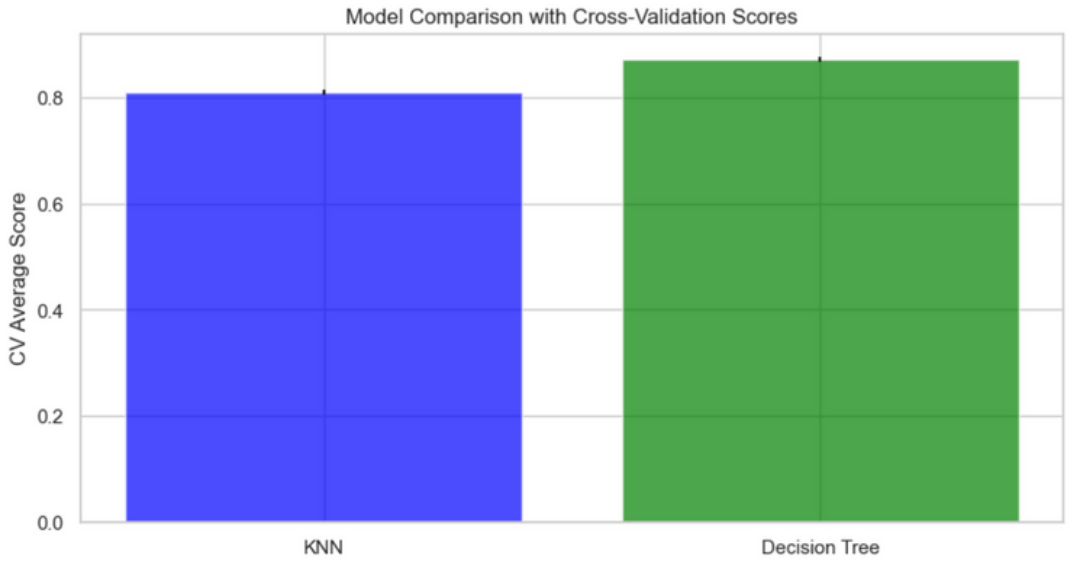
this plot shows that certain room types may be associated with a higher or lower average number of special requests, potentially reflecting the preferences or needs of guests choosing those types.

Question 10: Can we visualize the confusion matrices for both the KNN and Decision Tree models to compare their performance in terms of true positives, false positives, false negatives, and true negatives?



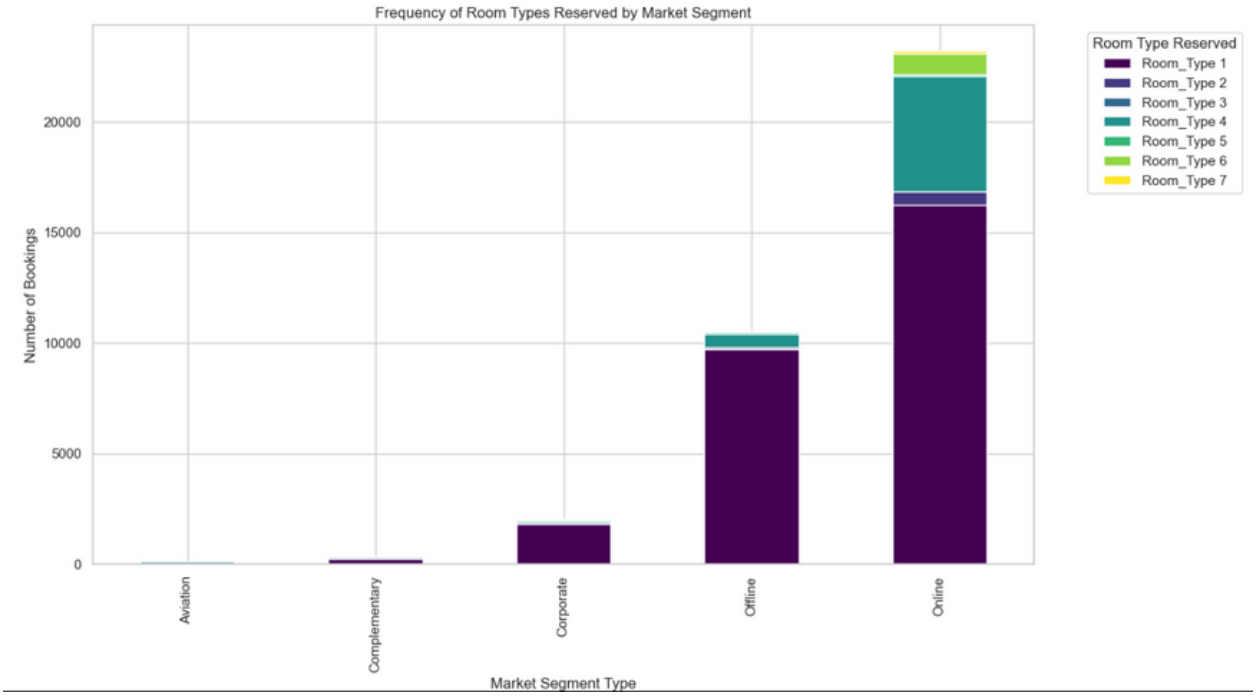
Both models are better at predicting cancellations (TP) than correctly identifying non-cancellations (TN). The Decision Tree model has fewer false positives (FP) than the KNN model, indicating it is less likely to incorrectly predict a booking as canceled. Conversely, the KNN model has fewer false negatives (FN) than the Decision Tree model, suggesting it is better at catching cancellations. Overall, the KNN model has predicted more true positives and true negatives than the Decision Tree model, showing that it has a higher overall accuracy.

Question 11: Can we create a visual comparison of cross-validation scores for both models to assess their generalization ability?



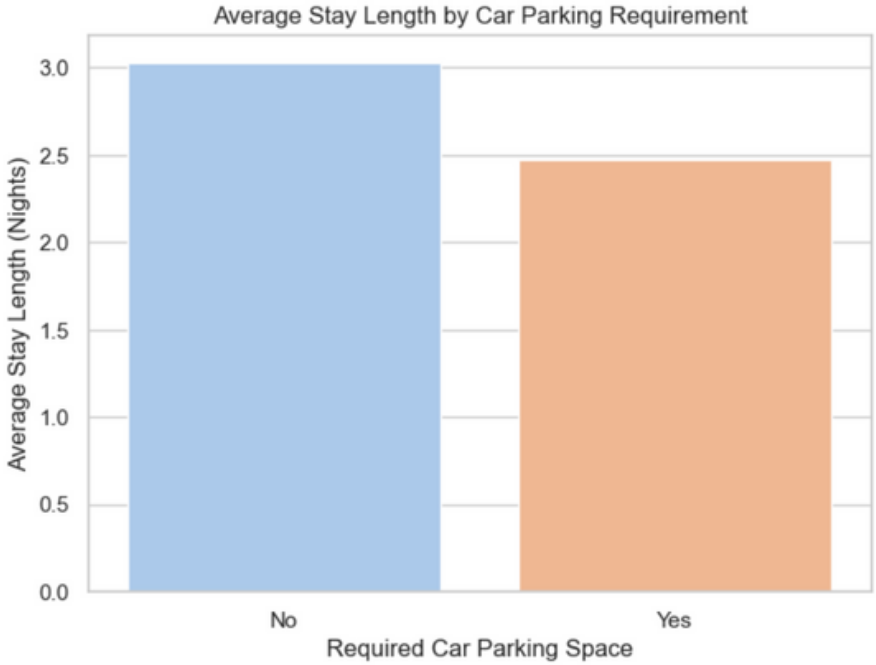
twe can observe that the Decision Tree model has a higher cross-validation score compared to the KNN model. This suggests that the Decision Tree may have better generalization performance on this dataset.

Question 12: What is the frequency of different room types being reserved by market segment, and can this be represented in a visual format?



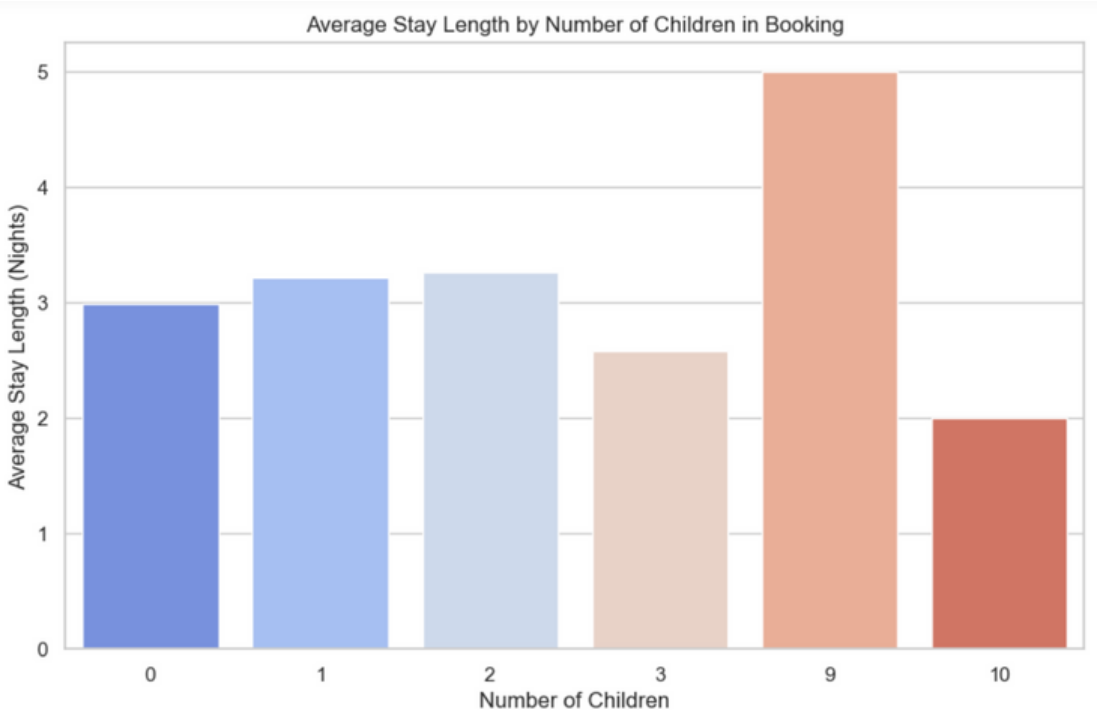
The 'Online' market segment has the highest number of bookings compared to other segments, showing that online channels are the most popular for booking rooms. Within the 'Online' market segment, Room_Type 1 is the most commonly reserved, as indicated by its significant proportion in the stacked bar. This suggests that Room_Type 1 might be a standard or economical option that is preferred by online customers.

Question 13: How does the requirement for car parking space influence the average stay length (weekend plus week nights), and can we depict this association graphically?



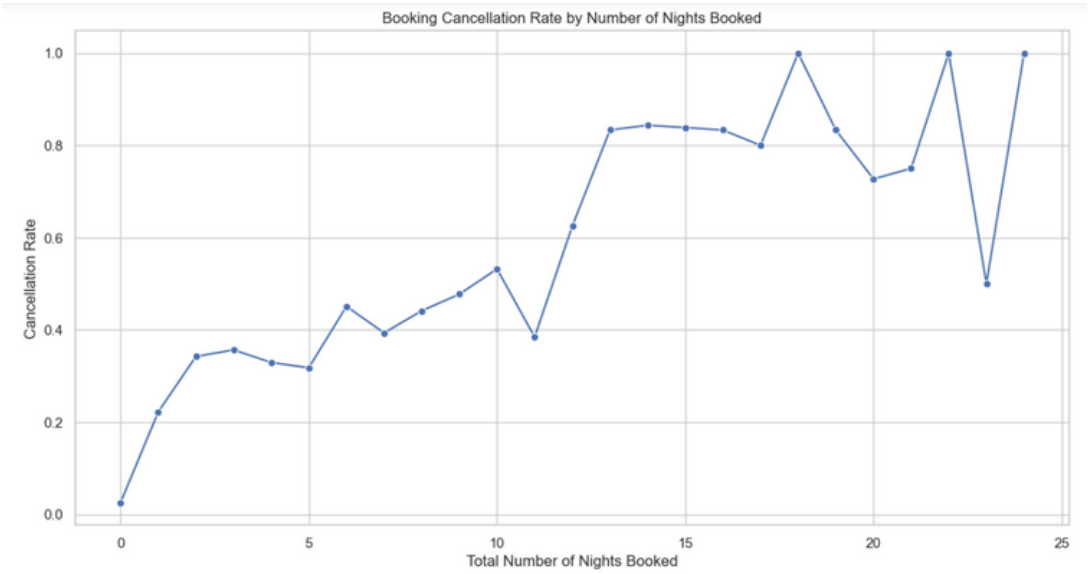
there is an association between requiring a car parking space and a longer hotel stay. Guests who need parking tend to stay longer, which could imply that they may be traveling by car and, as such, might be staying longer due to the nature of their travel

Question 14: Can we explore and visualize if there is a correlation between the average number of children in a booking and the length of stay?



Bookings with a small number of children do not differ in stay length from those with no children. Families with three children might plan longer stays, which could be indicative of extended family vacations. The peak for nine children could suggest a special scenario, such as a larger family gathering or event-related stay, which typically involves longer bookings.The drop for ten children could be an outlier due to the small sample size or specific circumstances affecting those particular bookings.

Question 15: How does the booking cancellation rate vary with the number of nights originally booked, and can we depict this pattern graphically?



tFor shorter stays (up to around 10 nights), the cancellation rate generally increases with the number of nights booked. This shows that as the length of stay increases, the likelihood of plans changing also increases. There is a peak in cancellation rates for stays of around 10 to 15 nights, suggesting a higher uncertainty or change in plans for bookings of this length. Beyond 15 nights, the cancellation rate shows more functuation with sharp increases and decreases. This could be due to a smaller sample size for longer bookings, making the data more likely to fluctuate, or it might reflect specific circumstances affecting these longer stays.