Predictive Analytics - Hotel Booking Management

PREPARED BY

Syeda Sarah Mashhood

JANUARY 2024

OBJECTIVE

This case study aims to equip you with practical skills in data science, focusing on predicting customer behaviors and booking cancellations in the hotel industry. You will apply EDA, KNN, Decision Tree algorithms, and learn to handle class imbalances using SMOTE

About Dataset (INN Hotel)

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

At A Glance

(36275, 19)

Shape of Dataset

139 Entries (0.38%)

number of adults is 0 or less

0

Missing values

545 Entries (1.50%)

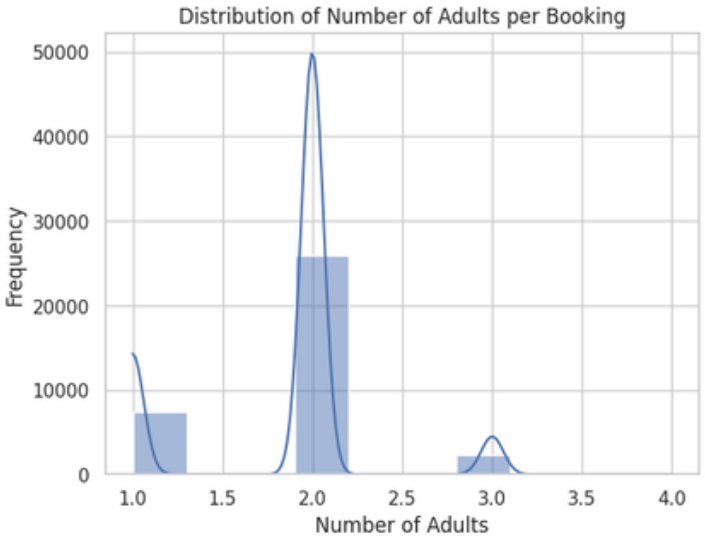
avg_price_per_room is 0 or less

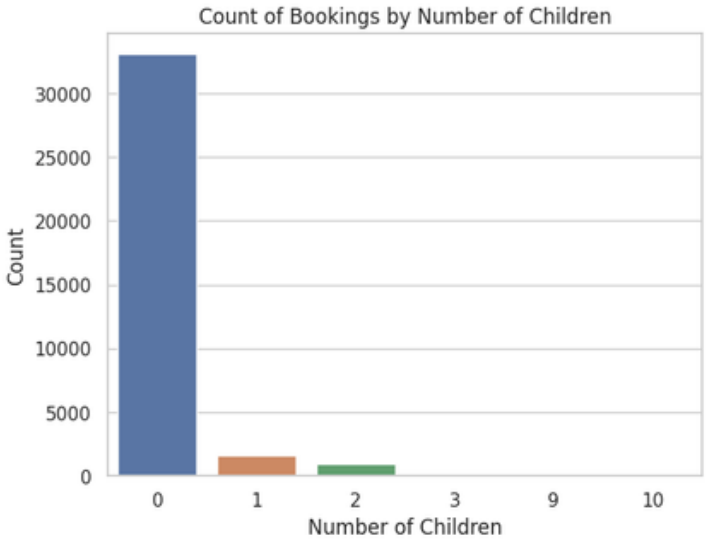
Anomolies are quite less in number so we'll drop the entries

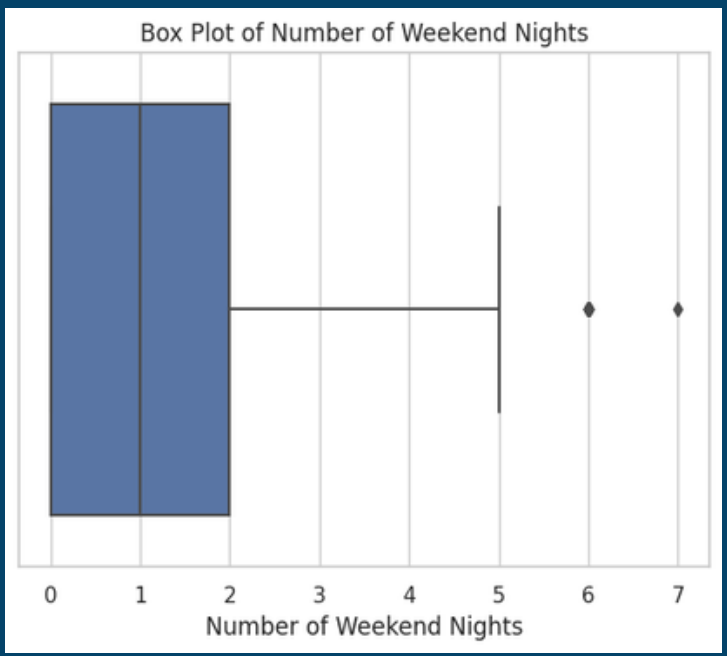
The average number of adults per booking is about 1.86, with a maximum of 4 adults. The number of children per booking averages around 0.1, indicating that most bookings do not include children. The average number of weekend nights is approximately 0.82, and weeknights are around 2.22. The lead time (number of days between the booking and the arrival date) averages 86 days. The average price per room is about €105.08.

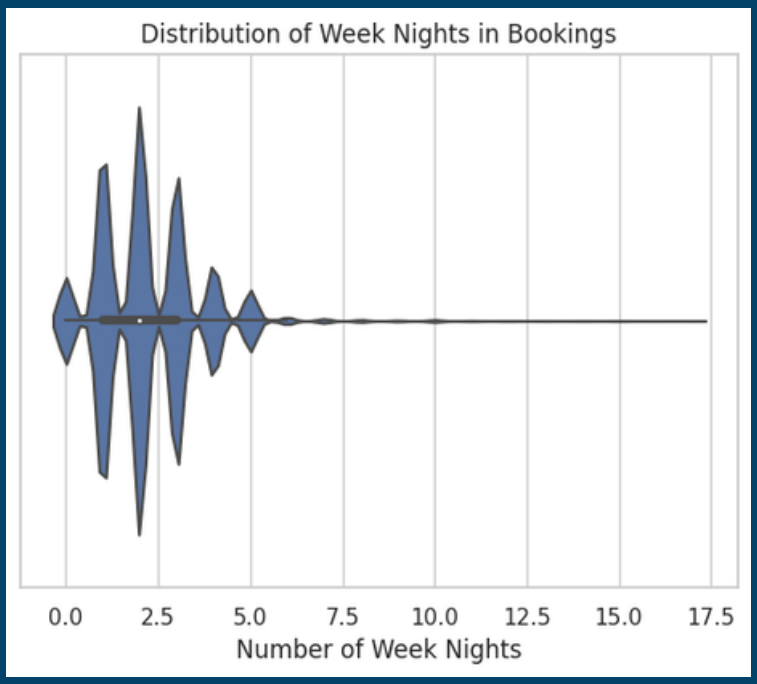
The distribution of the number of adults is skewed towards bookings for two adults. The number of weekend nights stayed is generally low, with most bookings including 0 to 2 weekend nights. The average price per room has a wide range but is mostly concentrated between €50 and €150. The lead time for bookings shows a broad distribution, with many bookings made relatively close to the arrival date and some made well in advance.

Exploratory Data Analysis

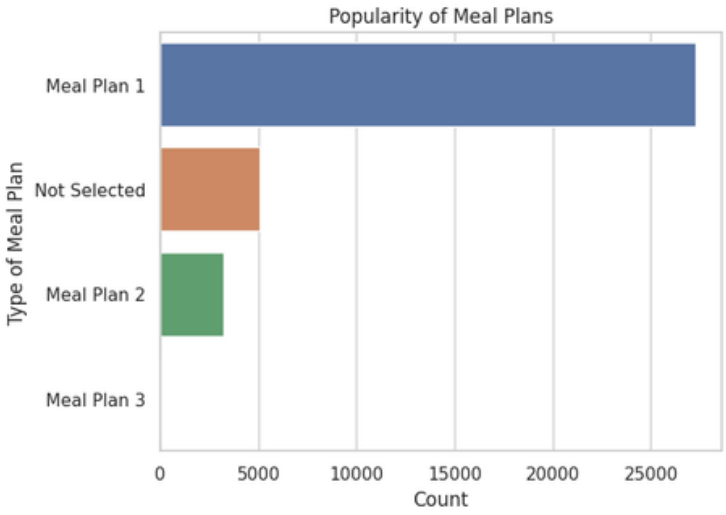


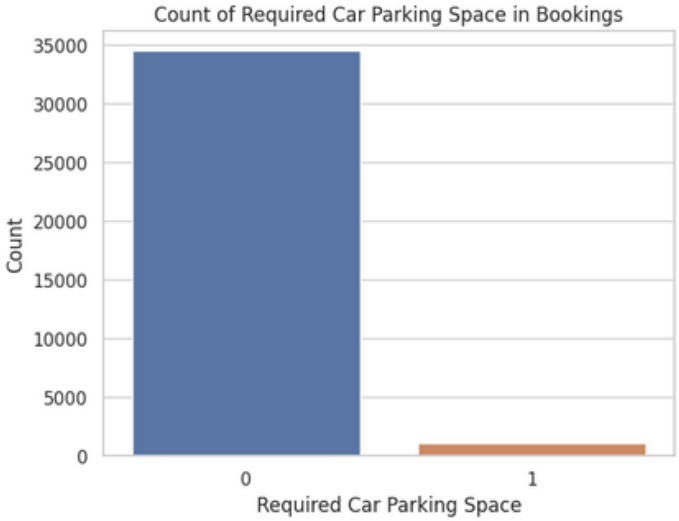


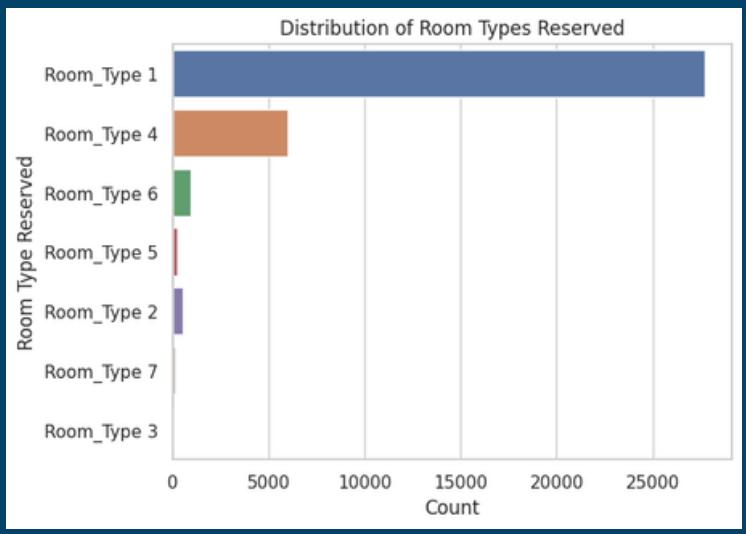


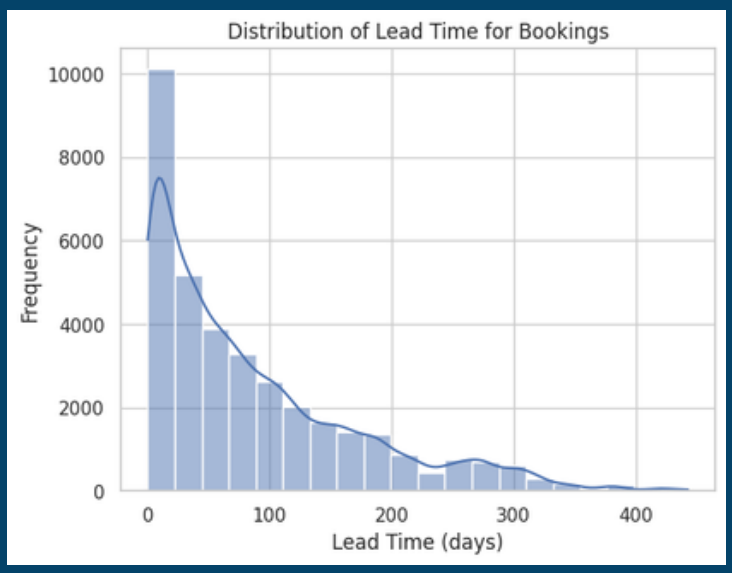


Exploratory Data Analysis

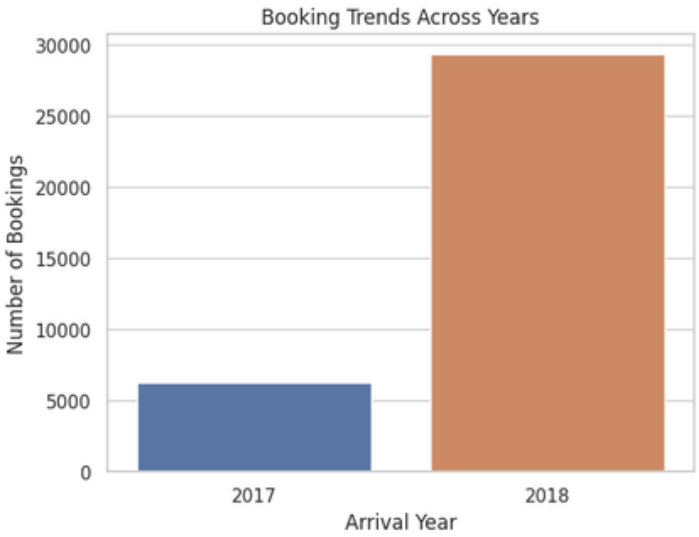


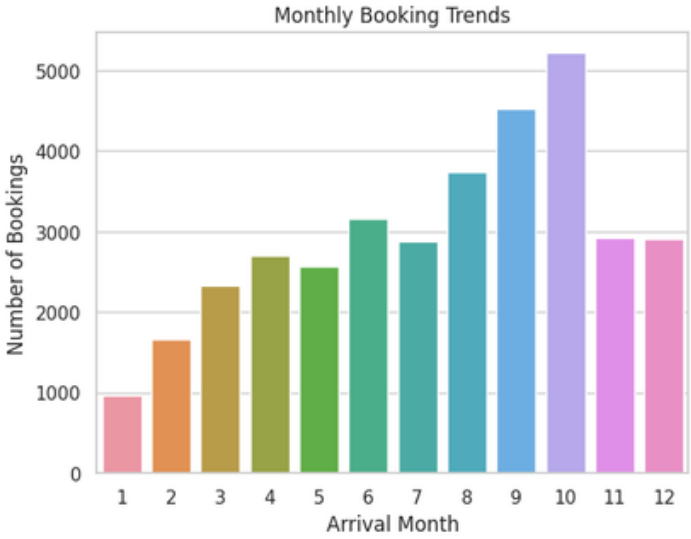


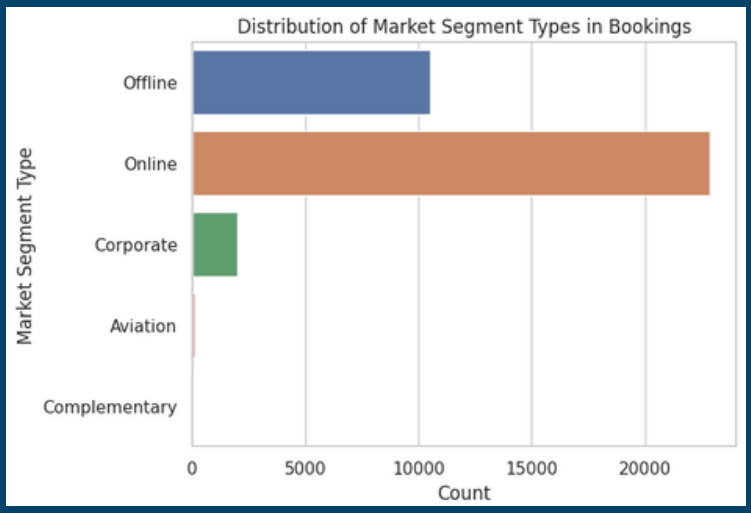


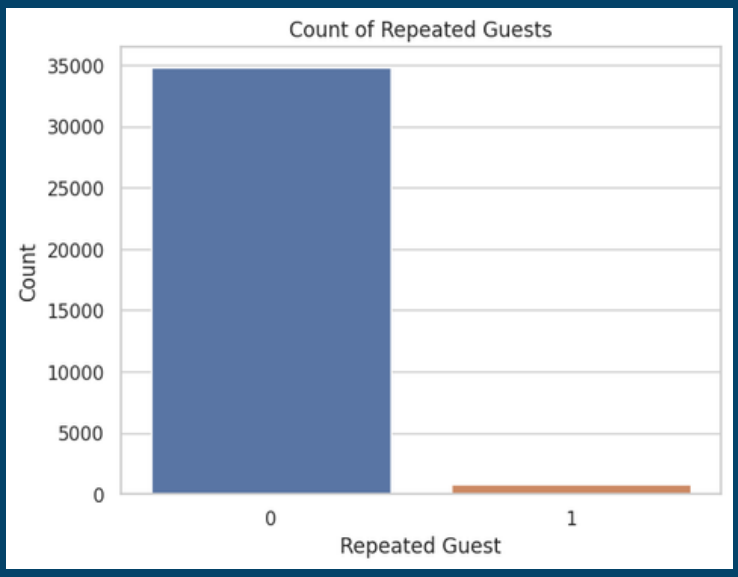


Exploratory Data Analysis

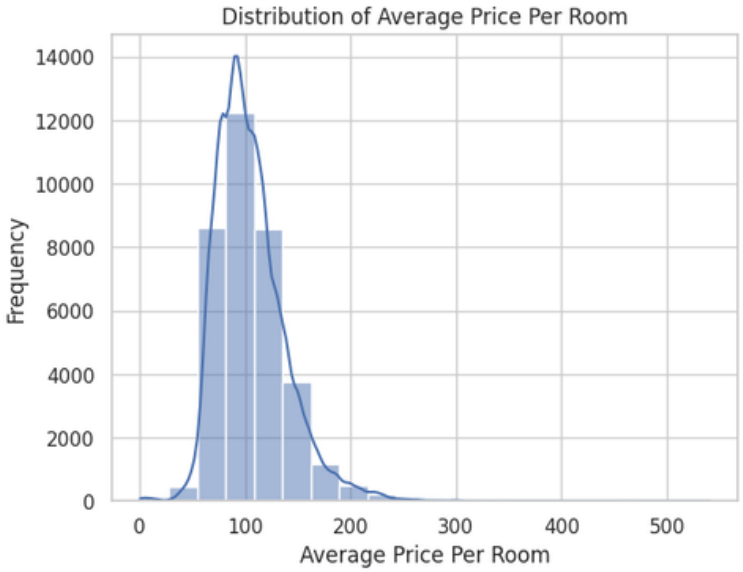


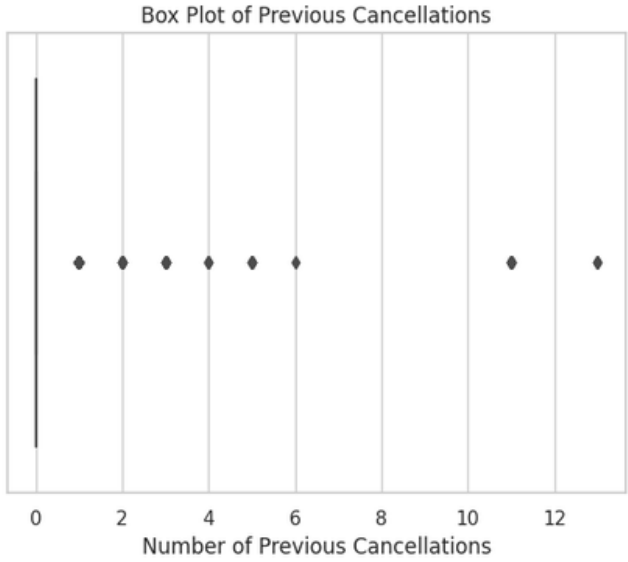


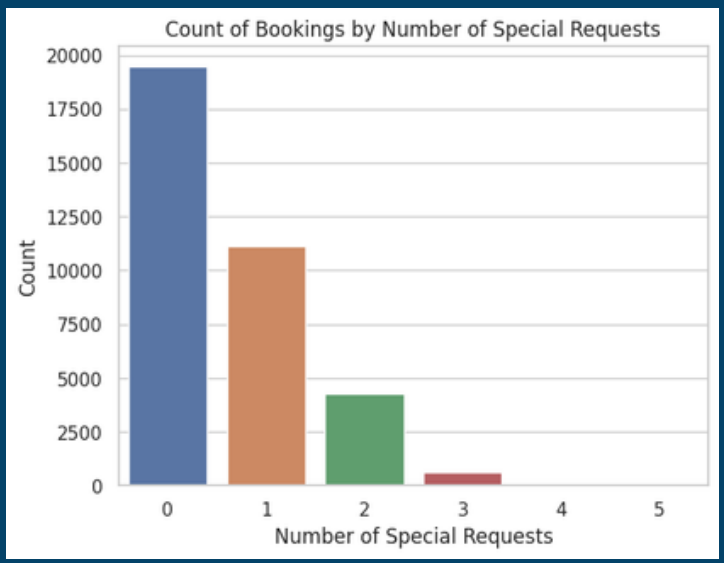


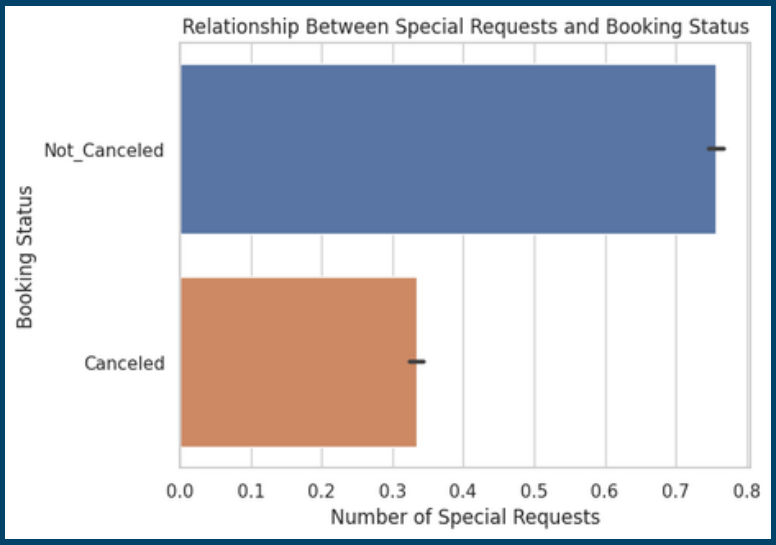


Exploratory Data Analysis

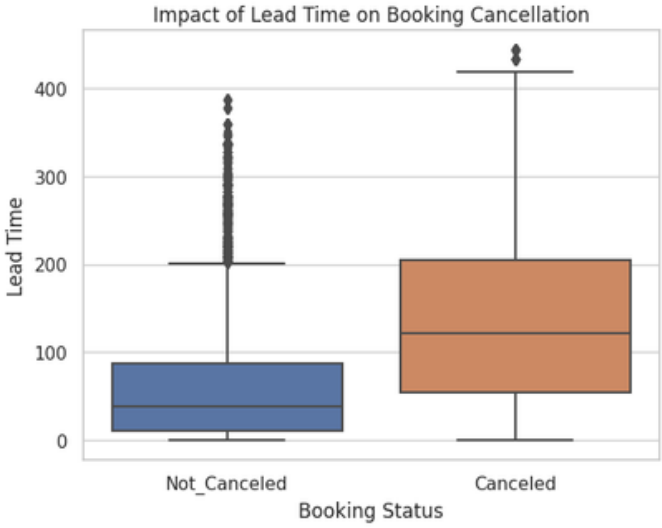


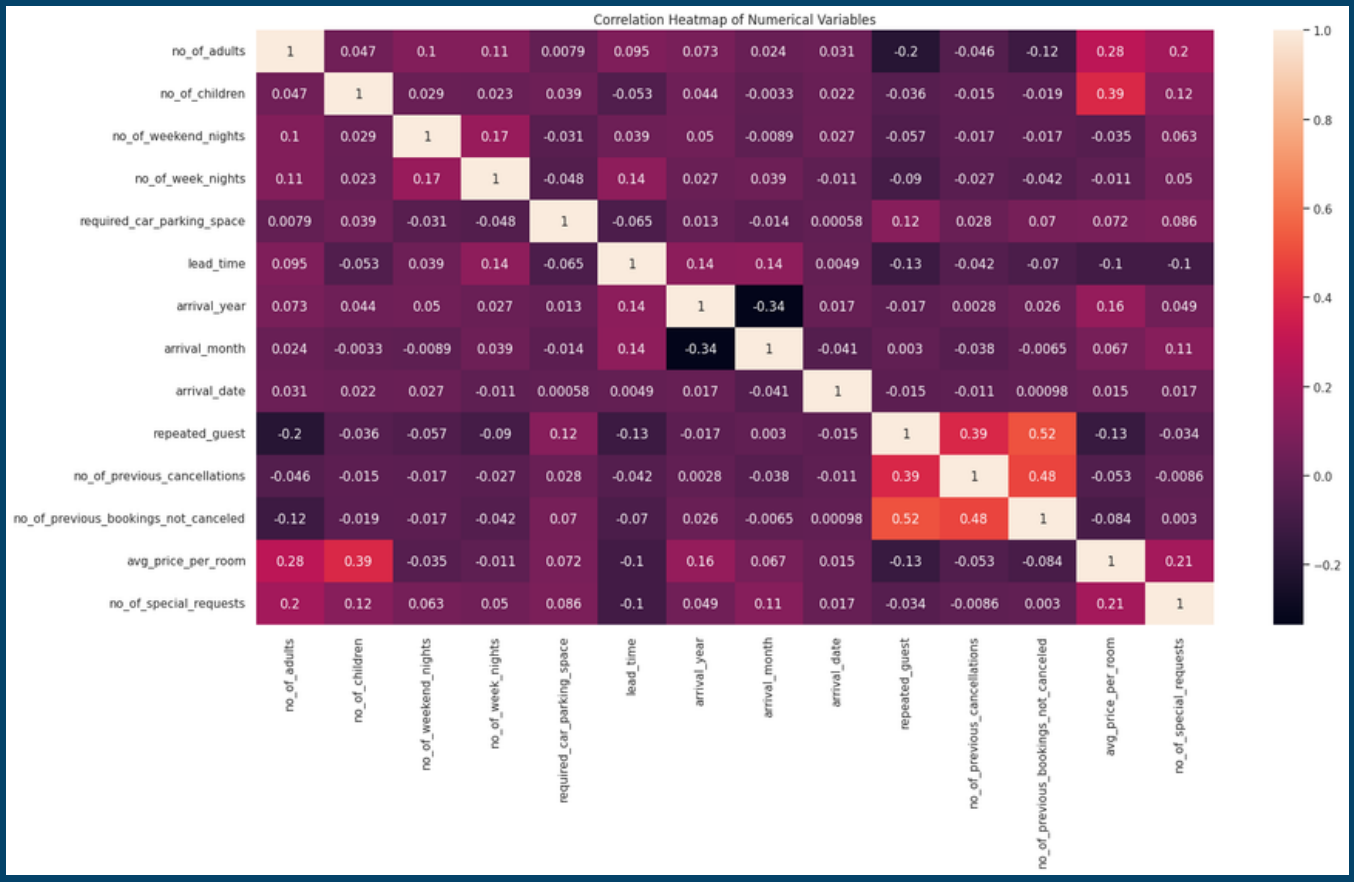






Exploratory Data Analysis





EDA INSIGHTS

-**Adult Bookings:** The majority of bookings typically include 2 adults.
- **Children in Bookings**: Bookings often do not include children, indicating a trend towards adult-only stays.
- **Weekend Nights:** On average, stays include 1 weekend night, though there are instances of stays extending up to 6 or 7 weekend nights.
- **Week Nights:** The average number of weeknights per booking ranges from 2 to 3, with some bookings extending up to a maximum of 17 nights.
- **Car Parking Preference:** Most customers do not opt for car parking spaces with their booking.
- **Meal Plan Preference:** Meal Plan 1 emerges as the most popular meal choice among customers.
- **Room Type Preference:** Room Type 1 is the most frequently chosen room type.
- **Lead Time for Bookings**: Many bookings are made with no significant lead time, although some bookings are made more than 200 days in advance.

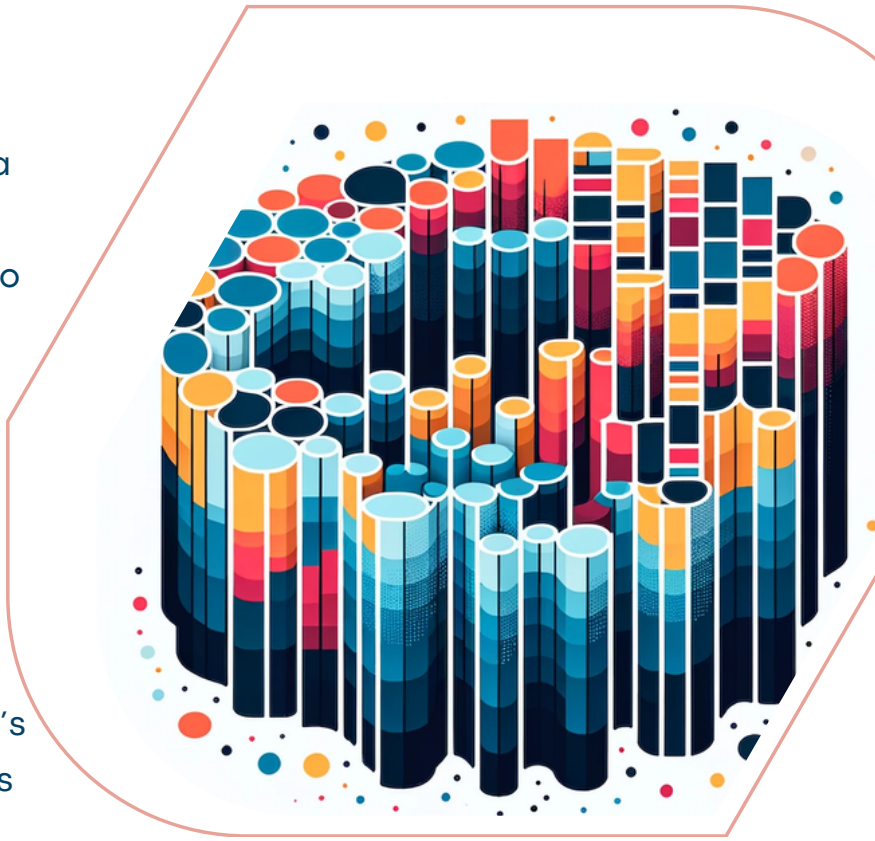
EDA INSIGHTS

- **Booking Year Trends:** The majority of the data comprises bookings from the year 2018.
- **Monthly Booking Trends**: Bookings peak during months 9 and 10, with fewer bookings in months 1 and 2.
- **Booking Method:** Online booking is predominant, with very few bookings coming from the Corporate Market Segment.
- **Repeated Guests:** The dataset indicates an absence of repeated customers, highlighting a notable aspect of customer behavior.
- **Special Requests in Bookings:** Special requests are generally not common in bookings, though some include 1 or 2 special requests.
- **Room Pricing:** The average room price is around 100 Euros.
- **Room Pricing and Children:** The average price per room shows a correlation with the number of children in the booking.
- **Relation of Previous Bookings with Repeated Guests:** The number of previous bookings not canceled shows a relationship with the presence of repeated guests.
- **Lead Time and Cancellations**: There is a noticeable relationship between the lead time of a booking and its likelihood of being canceled.

CLASS IMBALANCE

There is a class imbalance, albeit not an extreme one. The 'Not_Canceled' bookings constitute about two-thirds of the data, while 'Canceled' bookings make up about one-third. In such a scenario, applying SMOTE (Synthetic Minority Over-sampling Technique) can be beneficial for the following reasons:

Balancing Class Distribution:
SMOTE will help in creating synthetic samples of the minority class ('Canceled') to balance the dataset. This ensures that the predictive model has enough data from both classes to learn from, reducing the bias towards the majority class.



Improving Model Sensitivity: With a more balanced dataset, the model is likely to become more sensitive to the characteristics of the minority class, potentially improving its ability to correctly identify cancellations.

Enhancing Evaluation Metrics:
Balanced classes allow for a more accurate assessment of the model's performance, especially for metrics like precision, recall, and F1-score, which are crucial for understanding the model's performance across both classes.

PREDICTIVE MODELING

KNN (BEFORE SMOTE)

	PRECISION	RECALL	F1-SCORE
Canceled	0.58	0.45	0.51
Not_Canceled	0.75	0.84	0.79
Accuracy			0.71

PREDICTIVE MODELING

KNN (AFTER SMOTE)

	PRECISION	RECALL	F1-SCORE
Canceled	0.48	0.61	0.54
Not_Canceled	0.77	0.66	0.71
Accuracy			0.65

PREDICTIVE MODELING

KNN (GRID SEARCH WITH CROSS-VALIDATION)

BEST PARAMETERS: {'METRIC': 'MANHATTAN', 'N_NEIGHBORS': 3, 'WEIGHTS': 'DISTANCE'}

BEST CROSS-VALIDATION SCORE: 0.78

```
KNeighborsClassifier
KNeighborsClassifier(metric='manhattan', n_neighbors=3, weights='distance')
```

	PRECISION	RECALL	F1-SCORE
Canceled	0.49	0.60	0.54
Not_Canceled	0.77	0.68	0.72
Accuracy			0.65

MODEL INTERPRETATION

KNN

The decrease in overall accuracy of the K-Nearest Neighbors (kNN) model after applying SMOTE is analyzed from various perspectives:

1. **Change in Class Balance:** SMOTE creates synthetic samples for the minority class ('Canceled'), potentially making the model more sensitive to this class. This sensitivity may improve recall (from 45% to 61%) for the 'Canceled' class at the cost of increased false positives, where 'Not_Canceled' bookings are incorrectly labeled as 'Canceled'.
2. **Model Characteristics:** kNN's performance is influenced by data distribution and scale. The synthetic samples from SMOTE can alter point distances, affecting kNN's efficacy. The choice of 'k' is crucial; an inappropriate 'k' value might degrade performance in a balanced dataset.
3. **Performance Metrics:** A decrease in overall accuracy is accompanied by an improved f1-score for the 'Canceled' class, suggesting enhanced precision-recall balance. In imbalanced datasets, metrics like recall and f1-score can be more informative than accuracy, especially when identifying the minority class is crucial.
4. **Data Complexity and Overfitting:** The addition of synthetic samples by SMOTE can increase data complexity, risking overfitting. kNN may struggle with this complexity, particularly with high-dimensional data.

TO MAXIMIZE OVERALL ACCURACY, THE MODEL BEFORE APPLYING SMOTE IS THE BEST (ACCURACY OF 0.71).

TO BETTER IDENTIFY 'CANCELED' BOOKINGS (POTENTIALLY AT THE COST OF OVERALL ACCURACY), THE MODEL AFTER SMOTE (RECALL OF 0.61 FOR 'CANCELED') IS PREFERABLE.

FOR A BALANCED PERFORMANCE BETWEEN IDENTIFYING BOTH 'CANCELED' AND 'NOT_CANCELED' BOOKINGS, THE MODEL OBTAINED FROM GRIDSEARCHCV OFFERS A GOOD COMPROMISE, WITH DECENT RECALL RATES FOR BOTH CLASSES AND A REASONABLE OVERALL ACCURACY OF 0.65.

EACH MODEL HAS ITS STRENGTHS DEPENDING ON THE SPECIFIC GOALS OF OUR ANALYSIS (OVERALL ACCURACY VS. BALANCED RECALL ACROSS CLASSES).

PREDICTIVE MODELING

DECISION TREE (BEFORE SMOTE)

	PRECISION	RECALL	F1-SCORE
Canceled	0.78	0.80	0.79
Not_Canceled	0.90	0.89	0.89
Accuracy			0.86

PREDICTIVE MODELING

DECISION TREE (AFTER SMOTE)

	PRECISION	RECALL	F1-SCORE
Canceled	0.75	0.82	0.78
Not_Canceled	0.90	0.86	0.88
Accuracy			0.85

MODEL INTERPRETATION

DECISION TREE

Improvement in Recall for 'Canceled': After applying SMOTE, the recall for the 'Canceled' class increased from 0.80 to 0.82. This means the model is now better at identifying actual cancellations.

Slight Decrease in Precision for 'Canceled': The precision for the 'Canceled' class slightly decreased from 0.78 to 0.75. This indicates a minor increase in false positives (predicting 'Canceled' when it's actually 'Not_Canceled').

Small Change in 'Not_Canceled' Performance: For the 'Not_Canceled' class, there's a slight decrease in both precision and recall, but the changes are minimal.

Overall Accuracy: The overall accuracy slightly decreased from 0.86 to 0.85. This is a trade-off typically observed when addressing class imbalance - we gain better detection of the minority class at a small cost to overall accuracy.

Balanced Performance: The macro and weighted averages show a more balanced performance across both classes after applying SMOTE.

SMOTE IMPROVED MODEL BALANCE, ENHANCING IDENTIFICATION OF BOTH 'CANCELED' AND 'NOT_CANCELED' BOOKINGS WITH A TRADE-OFF IN PRECISION AND SLIGHT ACCURACY DROP.

THE ENHANCED ABILITY OF THE POST-SMOTE MODEL TO ACCURATELY IDENTIFY CANCELLATIONS SUGGESTS ITS BETTER ALIGNMENT WITH OBJECTIVES FOCUSED ON DETECTING SUCH INSTANCES.

THE CHOICE OF MODEL DEPENDS ON PRIORITIZING OVERALL ACCURACY OR BALANCED RECALL.