Hotel Haven: Predicting Customer Cancellations

A comprehensive data analysis to understand booking patterns, identify cancellation factors, and develop strategies to improve customer retention,

Presented by Ahmed Olatunji (data scientist)

Project Overview

Objectives

Assist Hotel Haven in understanding challenges related to customer cancellations and provide data-driven insights to improve retention.

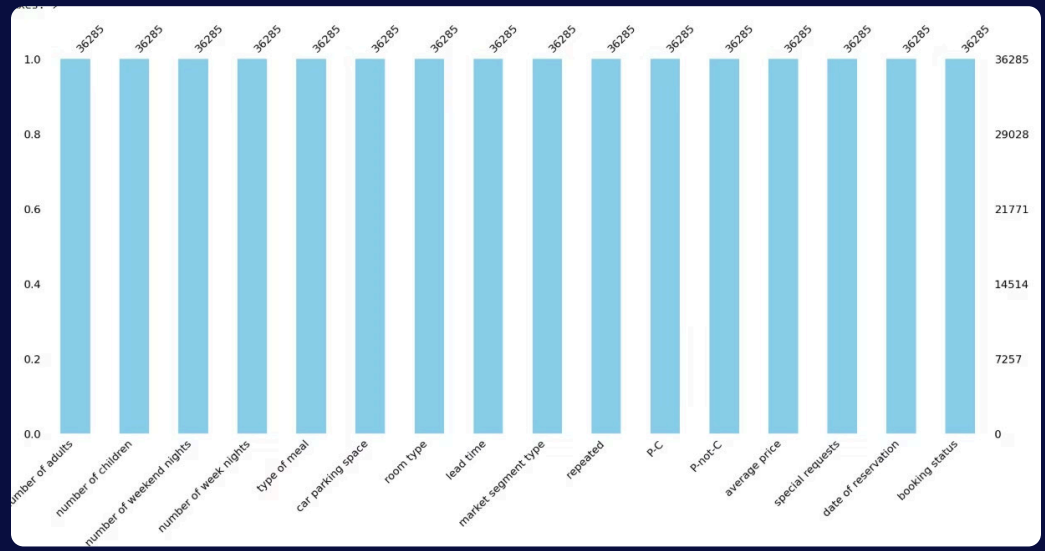
Deliverables

- Data cleaning and preparation
- Exploratory data analysis
- Comprehensive data analysis
- Feature engineering and model development
- Model evaluation and fine-tuning

Data Cleaning Process

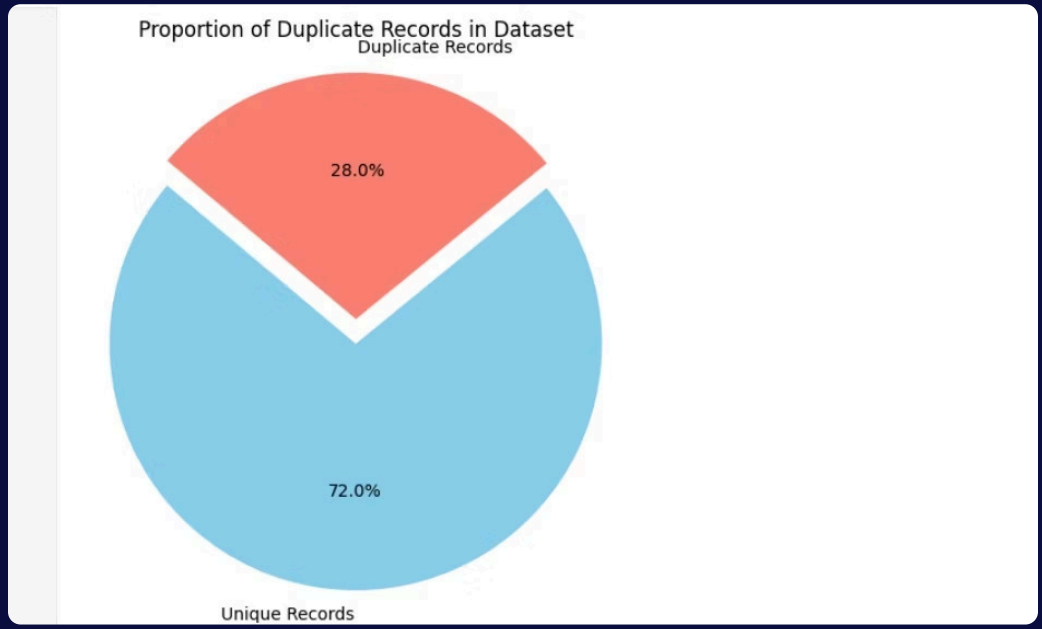
Key Cleaning Steps

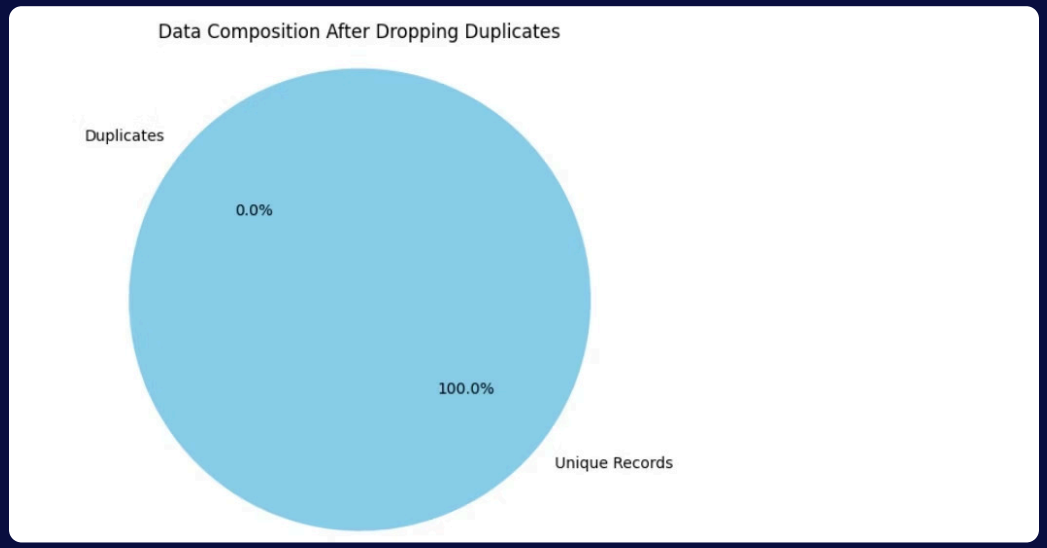
- This Data has no missing value
- Removed irrelevant columns (booking ID, doesn't contribute to prediction)



IDENTIFY AND REMOVE DUPLICATES

- Eliminated 10,276 duplicate values (28% of data)
- Addressed outliers in numerical variables
- Corrected 35 invalid dates in reservation records

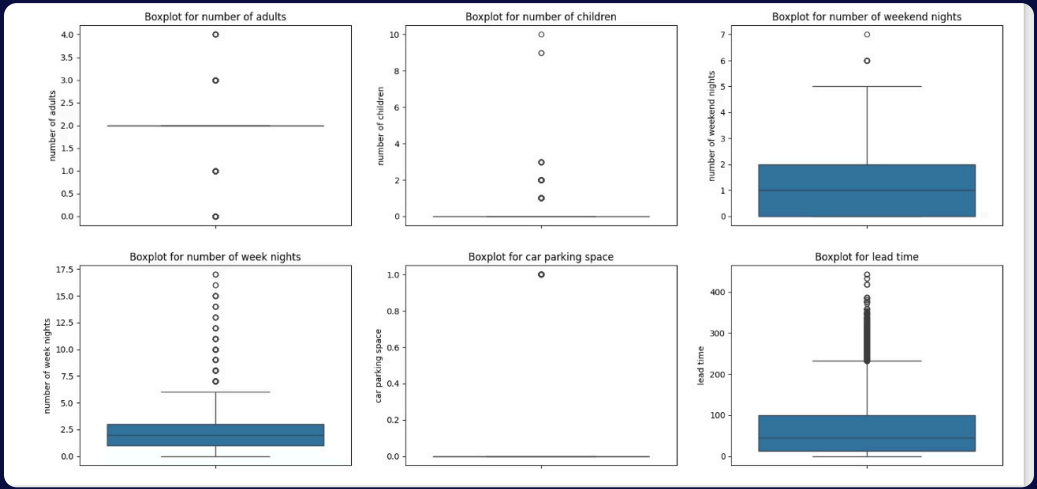




Duplicates cleaned

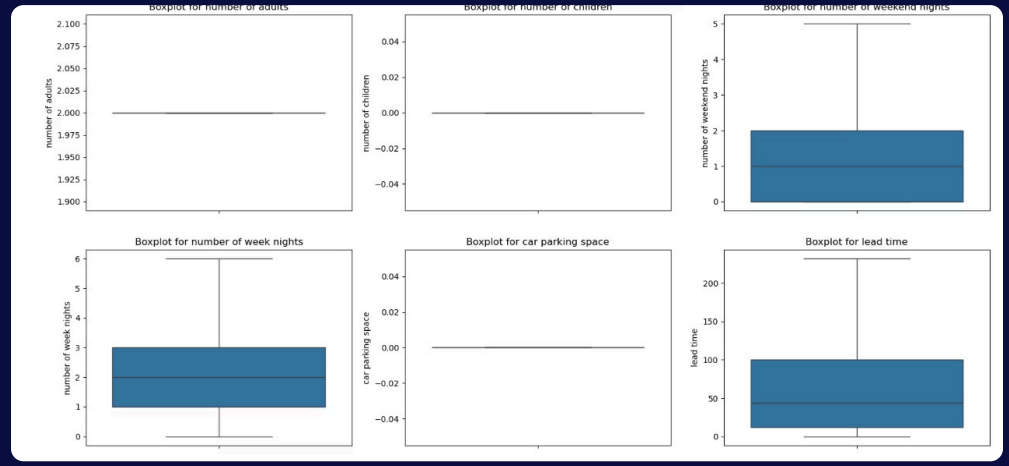
Result after removing duplicate values, our data can now have an accurate analysis , after duplicate has been removed

HANDLED OUTLIERS



Outliers were found in the following key columns:

- **Guest & Booking Details:** number of adults, number of children, number of weekend nights, number of week nights, car parking space, lead time, repeated
- **Booking Status & Financials:** P-C (percentage confirmed), P-not-C (percentage not confirmed), average price, special requests



As part of our data quality improvement initiative, we conducted a thorough analysis of the Hotel Haven booking dataset and identified several outliers that could potentially distort insights and business decisions.

i have documented both the pre- and post-cleaning results to demonstrate the impact of the changes. This enhancement aligns with our goal of leveraging high-quality data to drive operational efficiency , accuracy and customer satisfaction.

SPOTTED AND HANDLED INVALID DATES

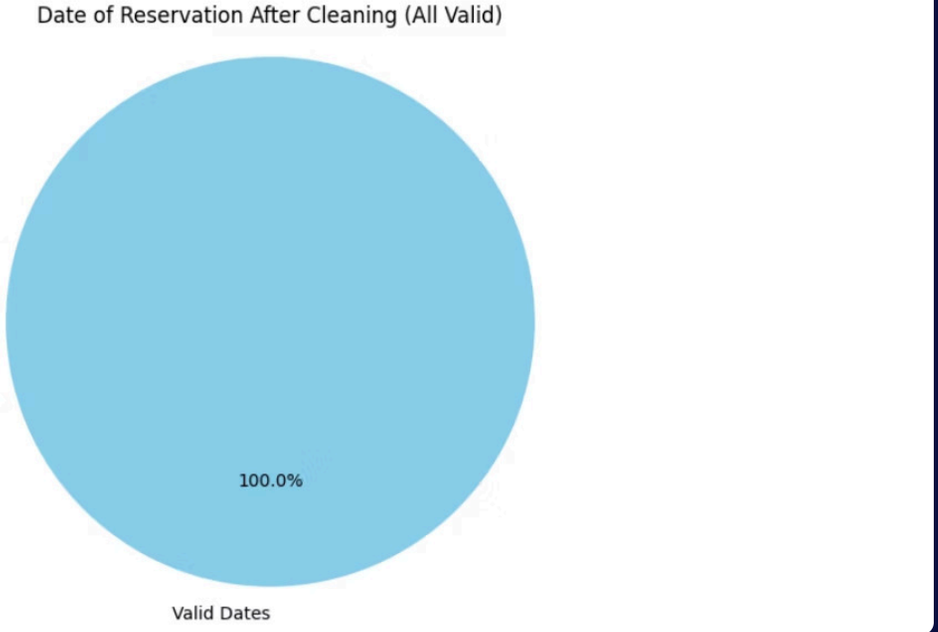
```
[26]: invalid_dates = data[data['date of reservation'].isna()]

[27]: invalid_dates
```

[27]:

	number of adults	number of children	number of weekend nights	number of week nights	type of meal	car parking space	room type	lead time	market segment type	repeated	P-C	P-not-C	average price	special requests	date of reservation	booking status
2322	2	0	1	5	Meal Plan 1	0	Room_Type 1	104	Online	1	1	0	61.43	0	NaT	Canceled
3202	1	0	1	3	Meal Plan 1	0	Room_Type 1	21	Online	0	0	0	102.05	0	NaT	Canceled
4722	2	0	1	3	Meal Plan 1	0	Room_Type 1	24	Offline	0	0	0	45.50	0	NaT	Not_Canceled
5297	1	0	1	1	Meal Plan 1	0	Room_Type 1	117	Offline	0	0	0	76.00	0	NaT	Not_Canceled
6304	2	1	1	5	Meal Plan 1	0	Room_Type 1	35	Online	0	0	0	98.10	1	NaT	Canceled
6590	2	2	1	3	Meal Plan 1	0	Room_Type 6	3	Online	0	0	0	183.00	1	NaT	Not_Canceled
7382	1	0	1	2	Meal Plan 1	0	Room_Type 1	117	Offline	0	0	0	76.00	0	NaT	Not_Canceled
7511	2	2	1	3	Meal Plan 1	0	Room_Type 6	3	Online	0	0	0	189.75	0	NaT	Not_Canceled
7582	2	0	1	3	Meal Plan 1	0	Room_Type 4	15	Online	0	0	0	85.55	1	NaT	Not_Canceled
7913	1	0	1	0	Meal Plan 1	0	Room_Type 4	21	Online	0	0	0	117.00	0	NaT	Not_Canceled
8107	1	0	1	2	Meal Plan 1	0	Room_Type 1	45	Online	0	0	0	76.30	0	NaT	Not_Canceled

Number of invalid dates was 35, out of 26009 data, accounting for 0.13%. of the whole data set

```
plt.show()
```



These entries were removed to ensure the accuracy and consistency of our time-based analyses, such as booking trends, lead time calculations, and seasonality insights.

This minor adjustment enhances data integrity without impacting the overall quality or representatives of the dataset.

EXPLORATORY DATA ANALYSIS

Lead Time Distribution(UNIVARIATE)

Last-Minute Bookings

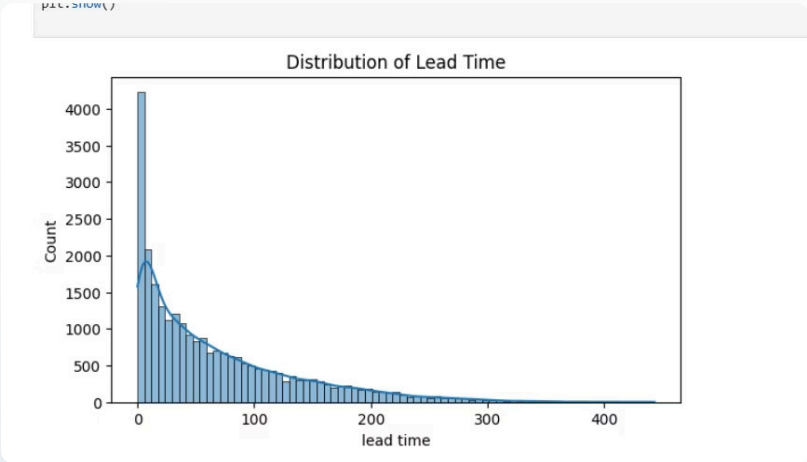
Most customers book just days or weeks before their stay, creating a right-skewed distribution.

Early Planners

A smaller group books months in advance, forming the extended right tail of the distribution.

Business Implications

Last-minute bookings are common while early planners are relatively rare, suggesting opportunities for targeted marketing strategies.

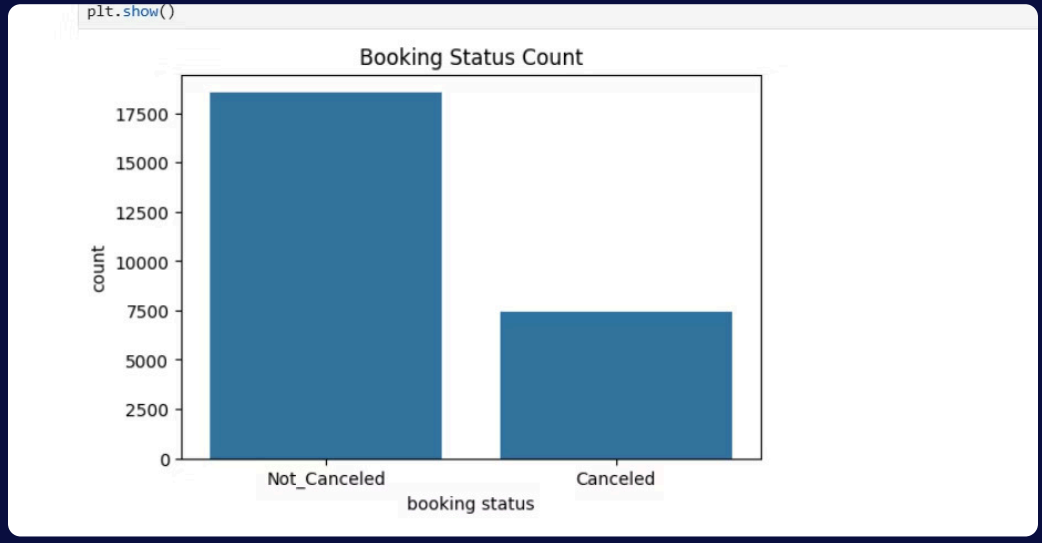


BOOKING STATUS COUNT

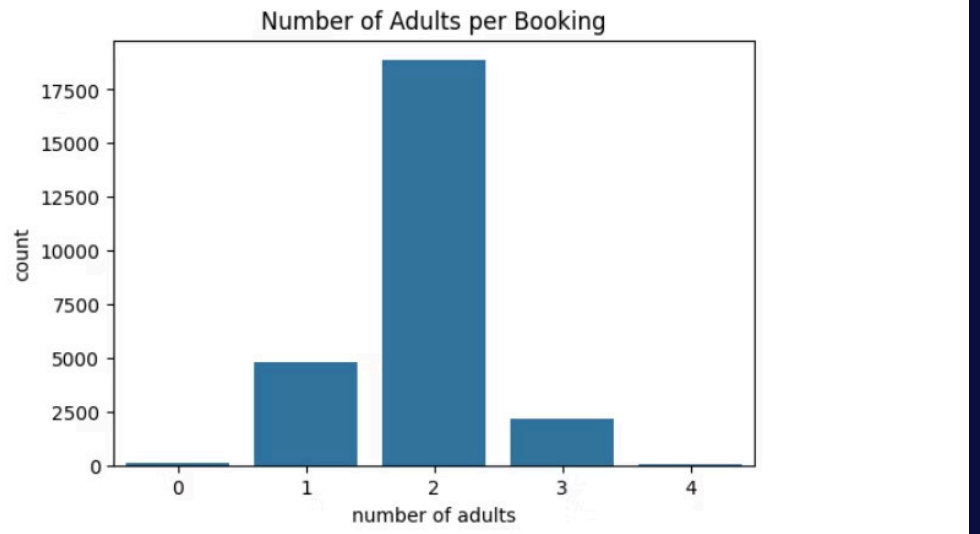
Fewer cancellations reflects a well-optimized user experience

Cancellation Policies Are Effective , and can have non-refundable, penalized, or limited cancellation windows.

Such policies often discourage casual or impulsive cancellation



```
sns.countplot(x='number of adults', data=data)
plt.title('Number of Adults per Booking')
plt.show()
```



This is typical in hospitality. Many hotel rooms are designed for 2 adults.

Hotels may mostly offer rooms meant for 2 adults, aligning with demand.

this Suggests good alignment between supply and market preference.

Less Demand for Solo or Group Bookings

Bookings with 1 adult or 3+ adults are much fewer.

indicates hotels are less attractive or optimized for solo travelers or large groups.

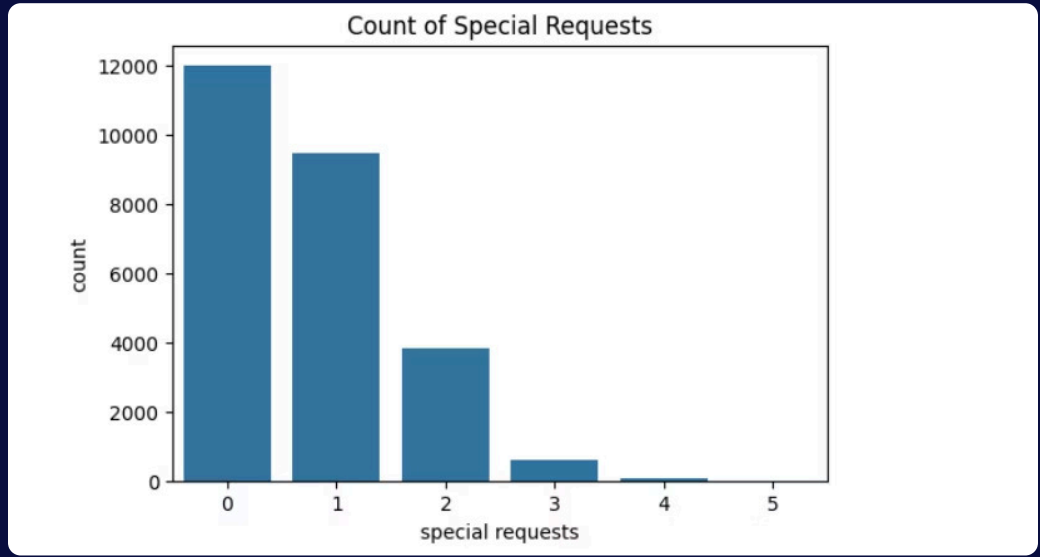
Pricing Strategy May Favor Pairs

Prices or promotions may be more affordable for two, making it the “sweet spot”

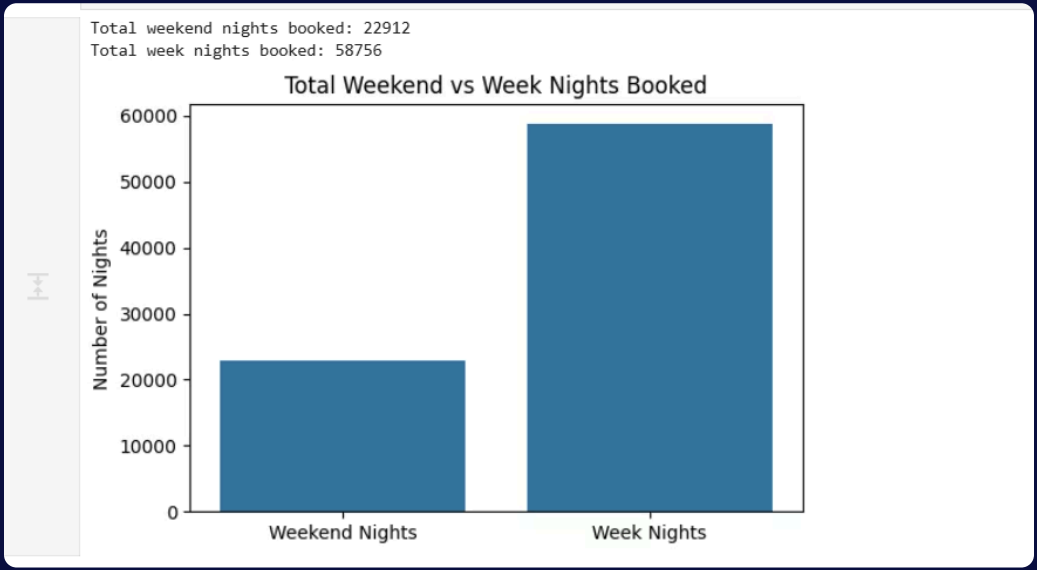
Count of special request

A majority of customers (value 0) did not request anything extra.

Could indicate: Standard services meet most needs., Guests are not aware they can make special requests. ,Booking platforms don’t emphasize or offer this feature clearly.



BI-VARIATE ANALYSIS



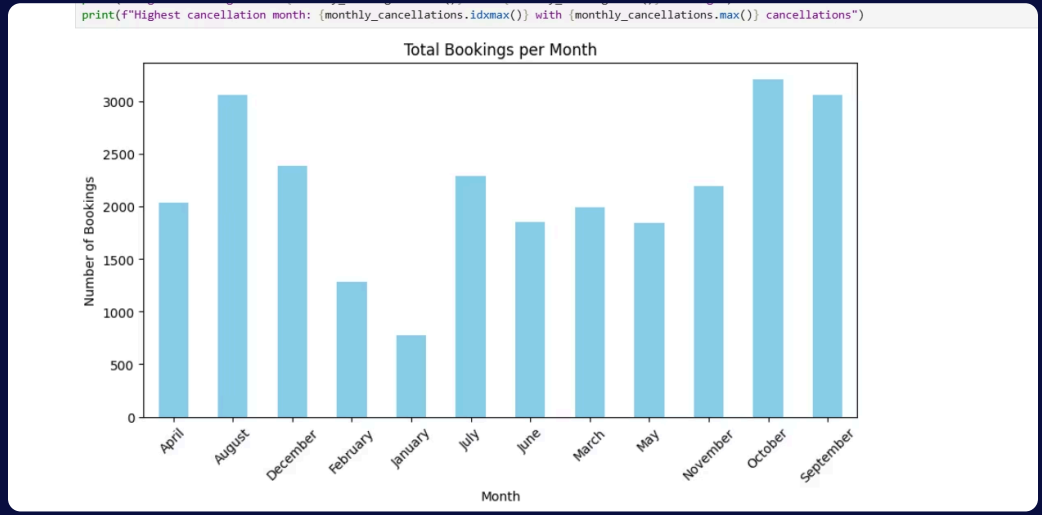
WEEKEND NIGHT VS WEEKDAY NIGHT

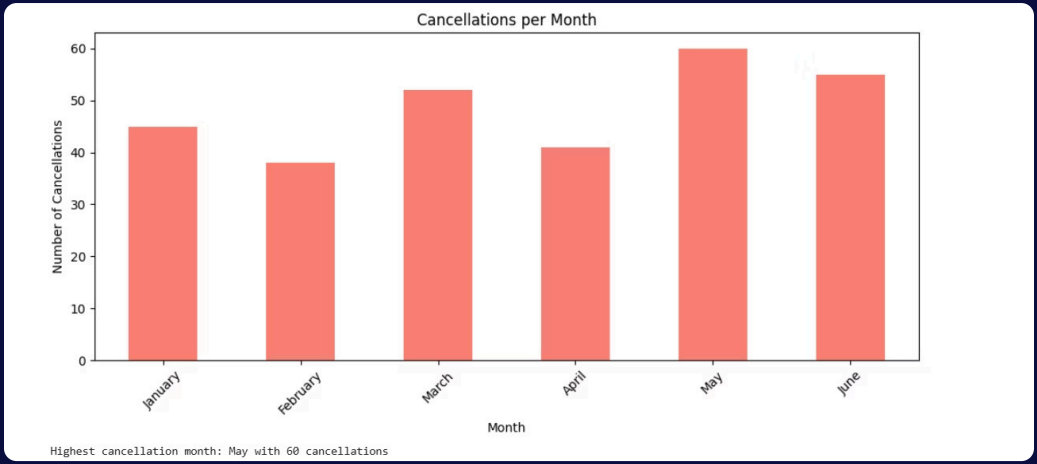
More total nights are booked during the weekdays than weekends.

Guests tend to stay more often on weekdays (Monday to Friday) than on weekends (Saturday and Sunday).

Bookings per month

As seen in the chat, January has the lowest number of booking and October has the highest number of booking,





Cancellations per month

- **Lowest cancellations** were recorded in **February** and **April**, suggesting stronger booking commitments or possibly fewer reservations during these months.
- In contrast, **May**, **June**, and **March** experienced **notably high cancellation volumes**, with **May** having the highest.

These patterns may be influenced by seasonal travel plans, promotional periods, or external factors such as school holidays or weather-related changes.

Understanding these trends allows us to refine our booking policies, adjust overbooking strategies, and target customer engagement efforts more effectively during high-risk months.

ROOM TYPE BY BOOKING STATUS

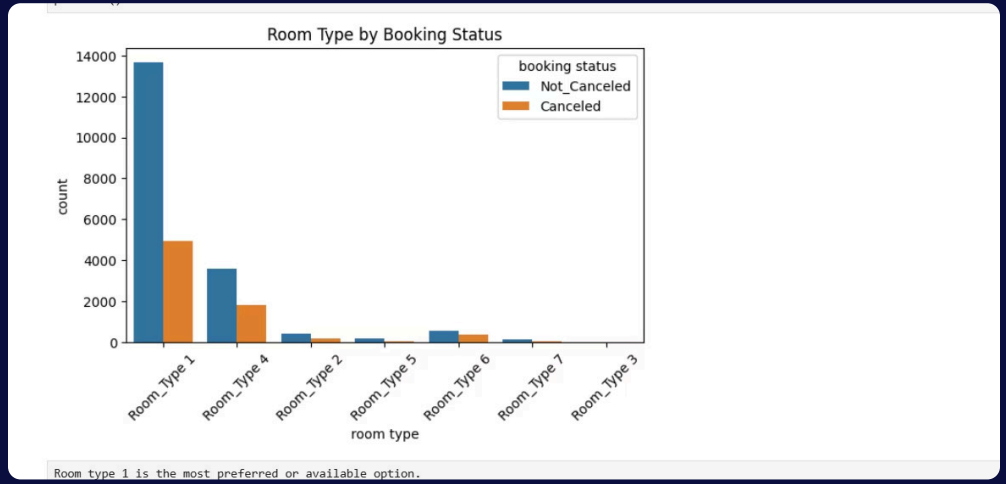
Guests mostly book this room type 1

Probably offers the best value, most comfort, or is the most advertised.

The high count of not canceled bookings for room type 1 suggests that guests who book this room are more likely to follow through with their stay.

Room type 2 and others might be:

Less available,less advertised , Less popular due to price, size, or amenities, Or maybe targeted at niche customers.



Special request by booking status

Guests who made at least 1 special request tend to cancel less. which means they are more engaged or serious about their stay.

Making a request could indicate a higher commitment level.

Guests with no special requests cancel more often. :Some might book casually or be less invested in the stay

Market segment by booking status

Online platforms often allow quick, easy cancellations with fewer penalties.

Guests may book multiple options online and cancel later (“booking shopping”).

Lower Commitment in Online Bookings

Customers booking online may be more price-sensitive or less committed.

Impulsive bookings or uncertain plans can lead to higher cancellations.

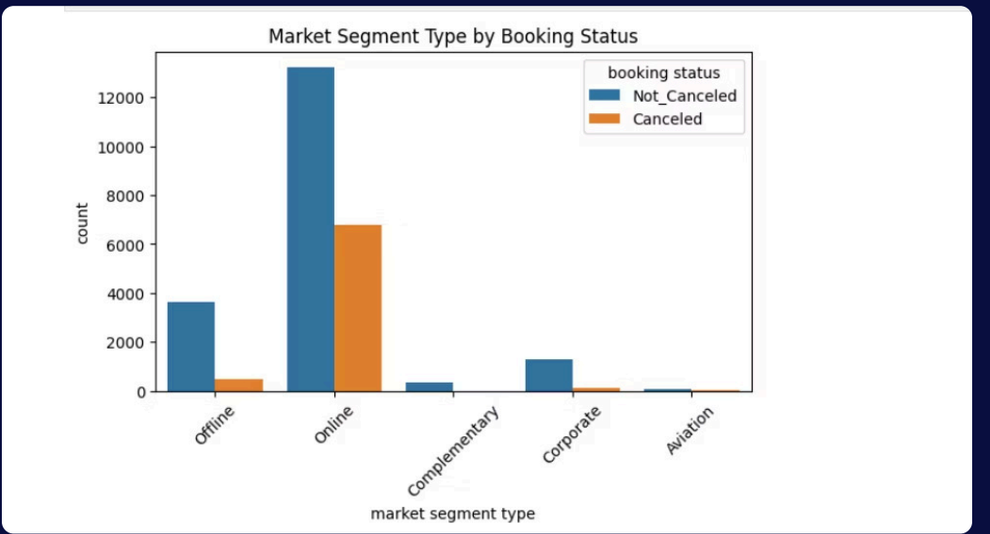
Offline Bookings Often Involve More Personal Contact

More effort invested by guest leads to higher commitment and fewer cancellations.

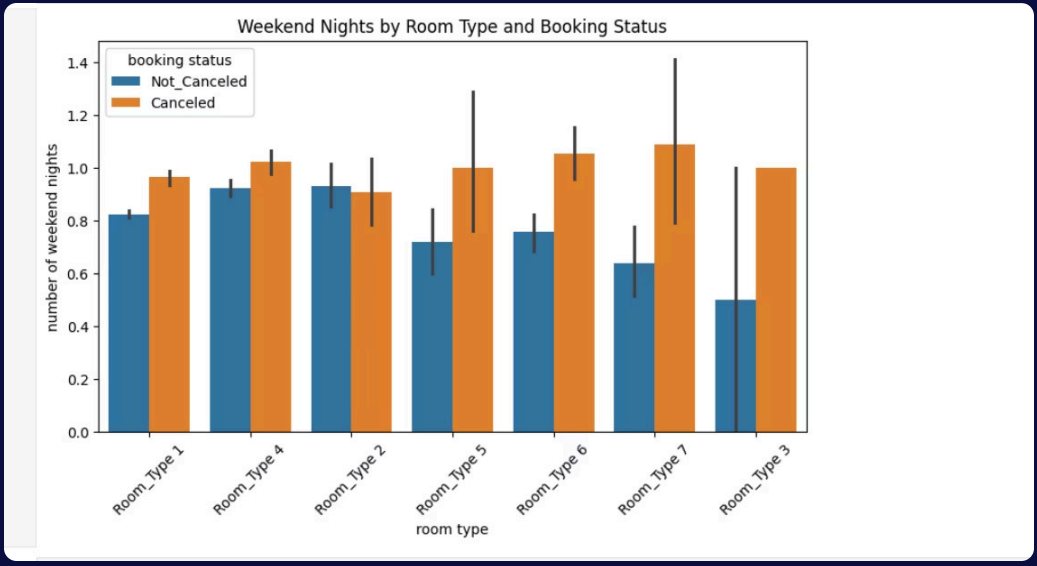
Different Cancellation Policies

Online channels might have more flexible cancellation policies.

Offline bookings may have stricter terms.



MULTIVARIATE ANALYSIS



Guests booking more weekend nights are canceling more often than guests booking fewer weekend nights.

This could suggest that weekend plans are more likely to change or be uncertain.

Room type 1, 4, and 2 dominate bookings: These are your most popular room types, so naturally they show higher booking activity.

Does this mean people cancel more on weekends? Not necessarily the weekends themselves, but rather longer weekend stays are more likely to be canceled.

Weekend trips might be more subject to last-minute changes, like social plans, events, or weather.

Cancellations on weekend stays are due to factors like event changes, travel complications, or shifting social plans .

Feature engineering and model development

Encoded Categorical Variables by Converting categorical variables like type of meal, room type, and Market segment type, booking status, and month into numeric formats.

```
data.head(2)
```

	number of adults	number of children	number of weekend nights	number of week nights	type of meal	car parking space	room type	lead time	market segment type	repeated	P-C	P-not-C	average price	special requests	date of reservation	booking status	Month
0	2	0	2	5	0	0	0	224	3	0	0	0	88.0	0.0	2015-10-02	1	10
5	2	0	0	2	1	0	0	232	3	0	0	0	100.0	1.0	2016-09-13	0	11

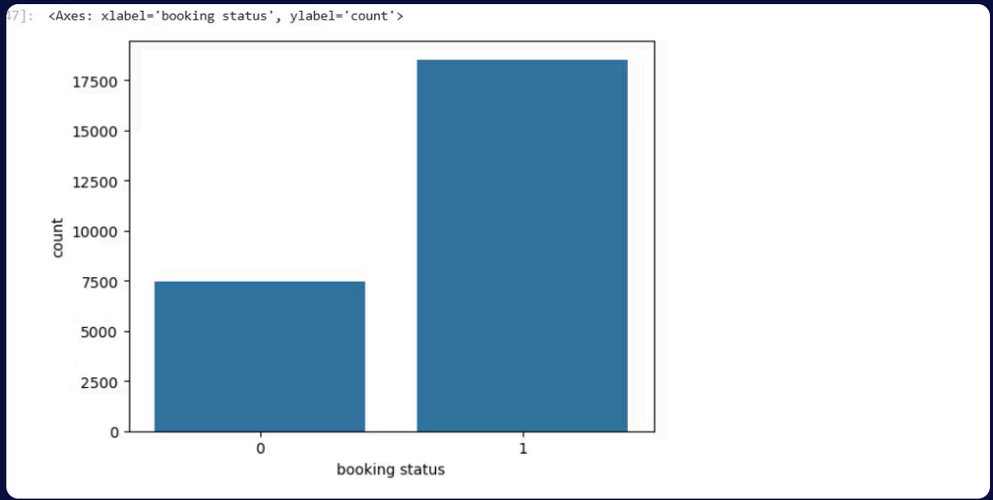
```
data.head()
```

	number of adults	number of children	number of weekend nights	number of week nights	type of meal	car parking space	room type	lead time	market segment type	repeated	P-C	P-not-C	average price	special requests	date of reservation	booking status	Month
0	0.0	0.0	1.266526	2.027412	0	0	0	2.478815	3	0	0	0	-0.489641	-0.942769	2015-10-02	1	10
5	0.0	0.0	-0.997687	-0.166724	1	0	0	2.603571	3	0	0	0	-0.151791	0.355613	2016-09-13	0	11
3	0.0	0.0	-0.997687	-0.166724	0	0	0	2.276086	4	0	0	0	-0.151791	0.355613	2017-05-20	0	8
556	0.0	0.0	-0.997687	-0.166724	1	0	0	2.603571	3	0	0	0	-0.714875	-0.942769	2017-07-01	1	5
656	0.0	0.0	-0.997687	-0.166724	1	0	0	2.603571	4	0	0	0	-0.109559	-0.942769	2017-07-01	0	5

Feature engineering and model development

SCALING OF THE DATA Scaled Numerical Features, Scale numerical columns to bring them to a similar range, which can improve model performance.

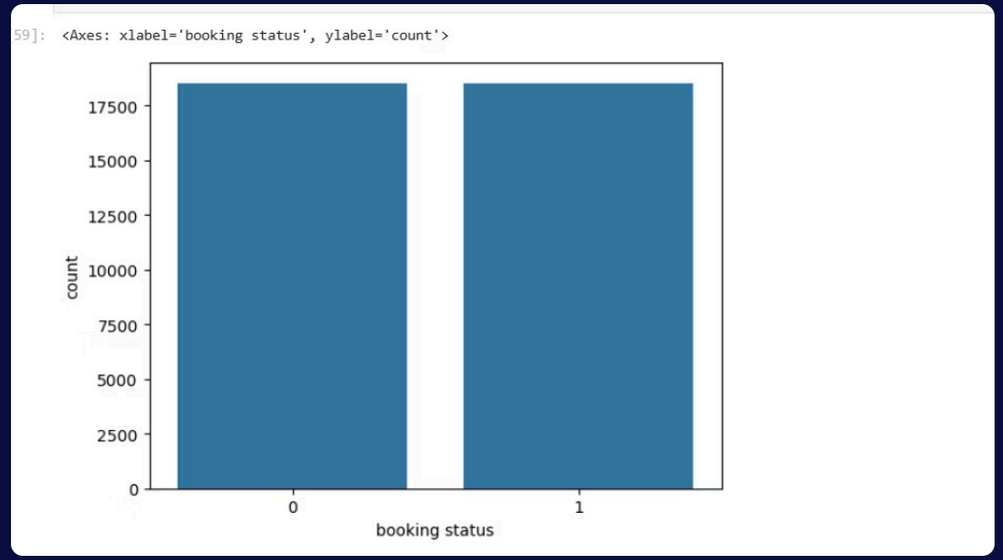
CLASS IMBALANCE AND OVERSAMPLING



After encoding categorical variables in the dataset,. The result revealed a **class imbalance**, where one class significantly outweighed the other.

To prevent the machine learning model from becoming biased toward the majority class, we plan to apply **oversampling techniques** (e.g., SMOTE or RandomOverSampler). This step ensures the model learns equally from both classes and improves its ability to generalize and predict minority outcomes effectively.

.



Balancing the dataset at this stage is essential for achieving **fair, reliable, and unbiased model performance** during training and evaluation

we applied **oversampling** techniques to increase the representation of the minority class. As a result, the number of samples in both classes is now **equal**, ensuring a balanced dataset.

This step is crucial for building a **fair and robust predictive model**, as it prevents the algorithm from being biased toward the majority class.

MODEL SELECTION

We’ll train multiple classification models to find the best one for predicting attrition. Start with a baseline and expand to more advanced models.

Model Selection: Logistic Regression Performance

The Logistic Regression model was evaluated on the balanced dataset, yielding the following classification metrics:

- **Accuracy:** 77%
- **Precision:** Approximately 76-77% for both classes
- **Recall:** Approximately 76-78% for both classes
- **F1-Score:** Balanced at 77% for both classes

The model demonstrates consistent and balanced performance across both classes (0 and 1), indicating it can effectively distinguish between confirmed and cancelled bookings without bias toward either class.

This solid baseline performance supports Logistic Regression as a strong candidate for further tuning and comparison with other models during the model selection phase.

```
print("Logistic Regression Classification Report:\n")
print(classification_report(y_test, y_pred))
```

```
Logistic Regression Classification Report:

              precision    recall  f1-score   support

           0       0.76      0.78      0.77      3707
           1       0.77      0.76      0.77      3707

    accuracy                           0.77      7414
   macro avg       0.77      0.77      0.77      7414
weighted avg       0.77      0.77      0.77      7414
```

```
# Evaluate performance
print("Random Forest Classification Report:\n")
print(classification_report(y_test, y_pred_rf))
```

```
Random Forest Classification Report:

              precision    recall  f1-score   support

           0       0.90      0.89      0.89      3707
           1       0.89      0.90      0.89      3707

    accuracy                           0.89      7414
   macro avg       0.89      0.89      0.89      7414
weighted avg       0.89      0.89      0.89      7414
```

Random Forest Performance

The Random Forest classifier was evaluated on the balanced dataset, achieving strong predictive performance:

- **Accuracy:** 89%
- **Precision:** Approximately 89-90% for both classes
- **Recall:** Approximately 89-90% for both classes
- **F1-Score:** Balanced at 89% for both classes

These metrics indicate that the Random Forest model delivers robust and consistent classification results, effectively distinguishing between confirmed and cancelled bookings with high accuracy and minimal bias.

Compared to the Logistic Regression baseline, the Random Forest shows a significant improvement in all key performance metrics.

Tuned Random Forest: Improved Model Performance

After applying **hyperparameter tuning** to the initial Random Forest model, we observed a measurable improvement in performance across all key metrics:

- **Accuracy:** 90%
- **Precision:** 90% (Class 0), 89% (Class 1)
- **Recall:** 89% (Class 0), 90% (Class 1)
- **F1-Score:** 90% for both classes
- **Support:** 3,707 samples per class (balanced dataset)

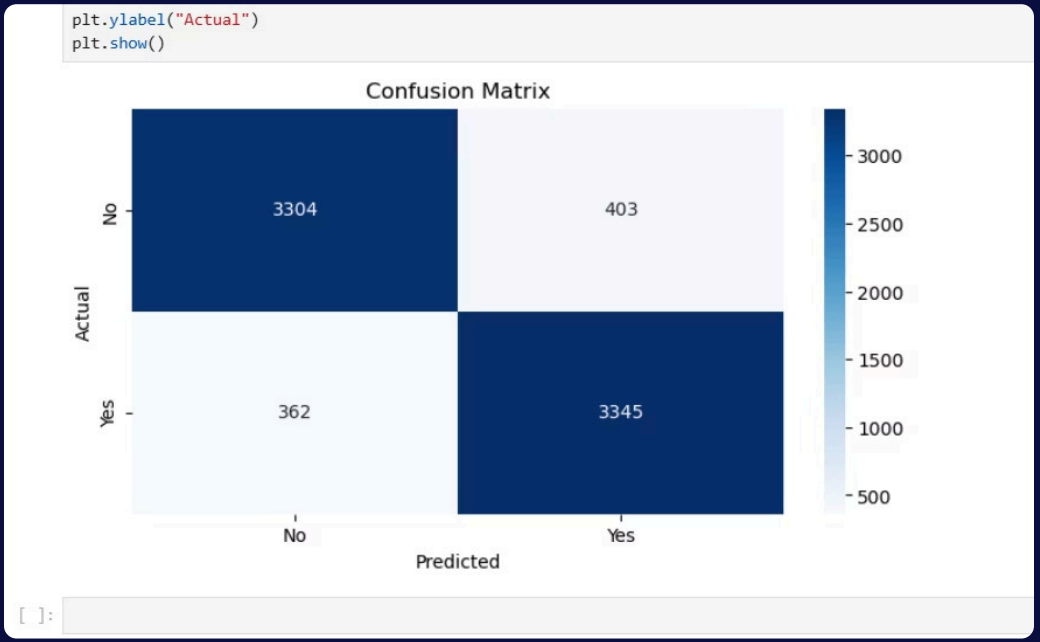
Key Highlights:

- The model shows **high precision and recall** for both booking outcomes — confirmed and cancelled.
- **Balanced F1-scores** reflect strong, consistent predictive capability.
- Tuning enhanced the model's ability to generalize, minimizing bias and overfitting.

```
print(classification_report(y_test, y_pred_best_rf))
```

```
Tuned Random Forest Classification Report:

              precision    recall  f1-score   support

           0       0.90      0.89      0.90      3707
           1       0.89      0.90      0.90      3707

    accuracy                           0.90      7414
   macro avg       0.90      0.90      0.90      7414
weighted avg       0.90      0.90      0.90      7414
```

CONFUSION METRICS



- **True Positives (3,345):** Model correctly predicted *cancellations*.
- **True Negatives (3,304):** Model correctly predicted *non-cancellations*.
- **False Positives (402):** Model predicted *cancellation* when it was *not*.
- **False Negatives (362):** Model predicted *no cancellation* when it was actually *cancelled*.

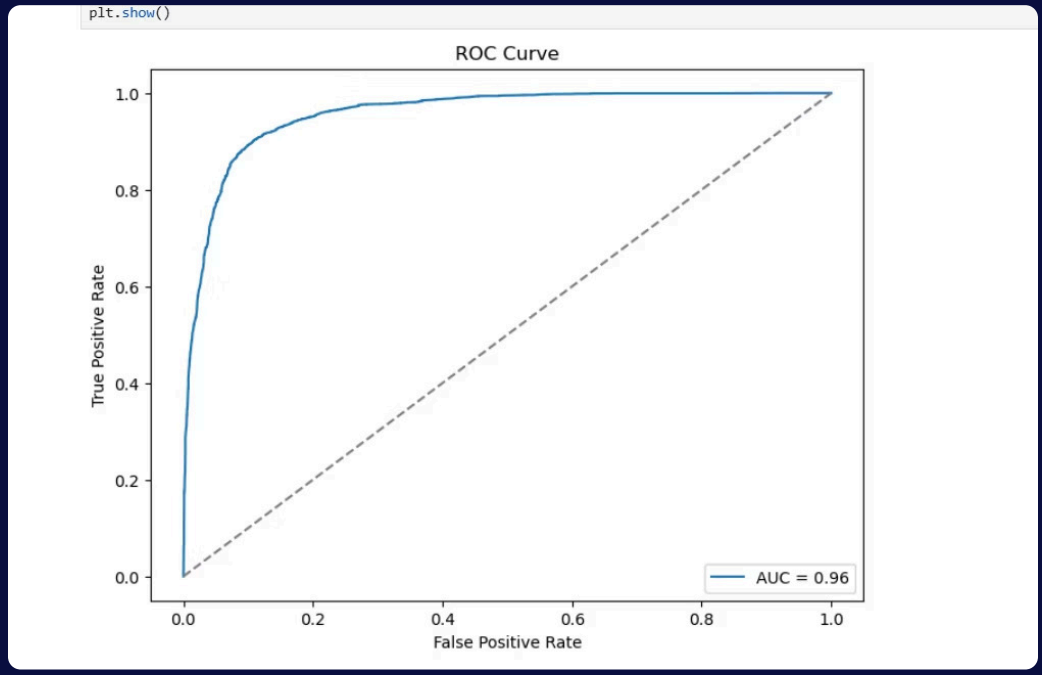
Insights:

- The model correctly classified **89% of bookings**.
- **False positive and false negative rates** are relatively low, indicating a well-balanced model.
- Strong performance in both classes confirms that the model generalizes well across booking statuses.

ROC Curve and AUC Score

As part of evaluating our model’s predictive power, we calculated the **ROC AUC score**, which measures the model's ability to distinguish between bookings that are likely to be **cancelled** and those that are **not**.

Our model achieved a **ROC AUC score of 0.96 (out of 1.00)**. this means the model can correctly differentiate between cancelled and non-cancelled bookings **96% of the time** and This indicates **very high performance** in identifying risk of cancellations, even before they happen.



INSIGHTS AND RECOMMENDED STRATEGIES

High Cancellation Months Identified (May, June, April)

Insight: Cancellations spike during specific months.
Solution: Introduce stricter cancellation policies or flexible pricing options during these peak months to reduce risk.

Lead Time Correlation

Insight: Bookings with **very long lead times** (i.e., booked far in advance) show higher cancellation rates.
Solution: Add incentives for early bookings to stay (e.g., non-refundable discounts or perks for holding the reservation).

Frequent Cancellers Can Be Profiled

Insight: Some guests, especially repeat users, have a higher tendency to cancel.
Solution: Use this model to flag high-risk bookings early and consider confirmation follow-ups or deposits for repeat cancellers.

High Number of Special Requests

Insight: Bookings with many special requests often result in cancellations possibly due to unmet expectations.
Solution: Improve communication and confirmations on special requests to manage guest expectations better.

Room Booking Characteristics Matter

Insight: Specific patterns (e.g., low number of adults or children, short stays) are more cancellation-prone.
Solution: Use predictive modeling at booking time to flag potentially risky reservations and offer personalized incentives.

Balanced Data Helps Predict Better

Insight: After oversampling and balancing your data, the model achieved **90% accuracy and a 0.96 ROC AUC**, meaning it's highly effective at detecting likely cancellations.
Solution: Integrate the model into the booking system to score each reservation and trigger interventions when risk is high.

Most Cancellations Are Predictable

Insight: With precision, recall, and F1-scores all at **90%**, most cancellations are not random — they follow patterns.
Solution: Move from reactive to **proactive cancellation management** — use the model's prediction to:

- Send automated reminders
- Offer loyalty points for honoring the booking
- Require partial prepayment based on risk level

THANK YOU