CMR
GROUP OF INSTITUTIONS
EXPLORE TO INVENT

A PROJECT REPORT

On

HOTEL BOOKING PREDICTION

Submitted to CMREC (UGC Autonomous)

In Partial Fulfillment of the requirements for the Award of Degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

MACHINE LEARNING

Submitted By

MOHAMMED ARSH KHAN **228R1A1294**

Under the guidance of

Mrs. M.JHANSI LAKSHMI

Assistant Professor

Internal Guide
Mrs. M.JHANSI LAKSHMI
Assistant professor,
Department of IT

Head of the Department
Dr. MADHAVIPINGILI
Assoc Professor & HOD,
Department of IT

TABLE OF CONTENTS

ABSTRACT

The hotel industry faces significant challenges due to booking cancellations and unpredictable customer behaviors. To address this, our project performs exploratory data analysis (EDA) on a comprehensive hotel booking dataset using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The dataset, which includes booking information for city and resort hotels, is cleaned and preprocessed to manage missing values and eliminate inconsistent data, such as bookings with no guests. The project replaces missing values intelligently using statistical techniques like mean imputation and mode substitution. Visualizations such as pie charts, bar plots, and heatmaps help uncover trends in booking status, lead time, and booking sources by country. These insights offer valuable indicators for predicting cancellation likelihood and optimizing hotel management strategies. Key findings include dominant countries in bookings, lead time distributions, and booking behaviors associated with cancellations. Though the project focuses primarily on EDA, it sets the stage for future predictive modeling using classification algorithms. By understanding customer behavior and booking trends, this analysis can support better forecasting, enhance revenue management, and reduce operational uncertainties. Ultimately, this project highlights the importance of data-driven decision-making in the hospitality industry through effective data preprocessing and visualization.

INTRODUCTION

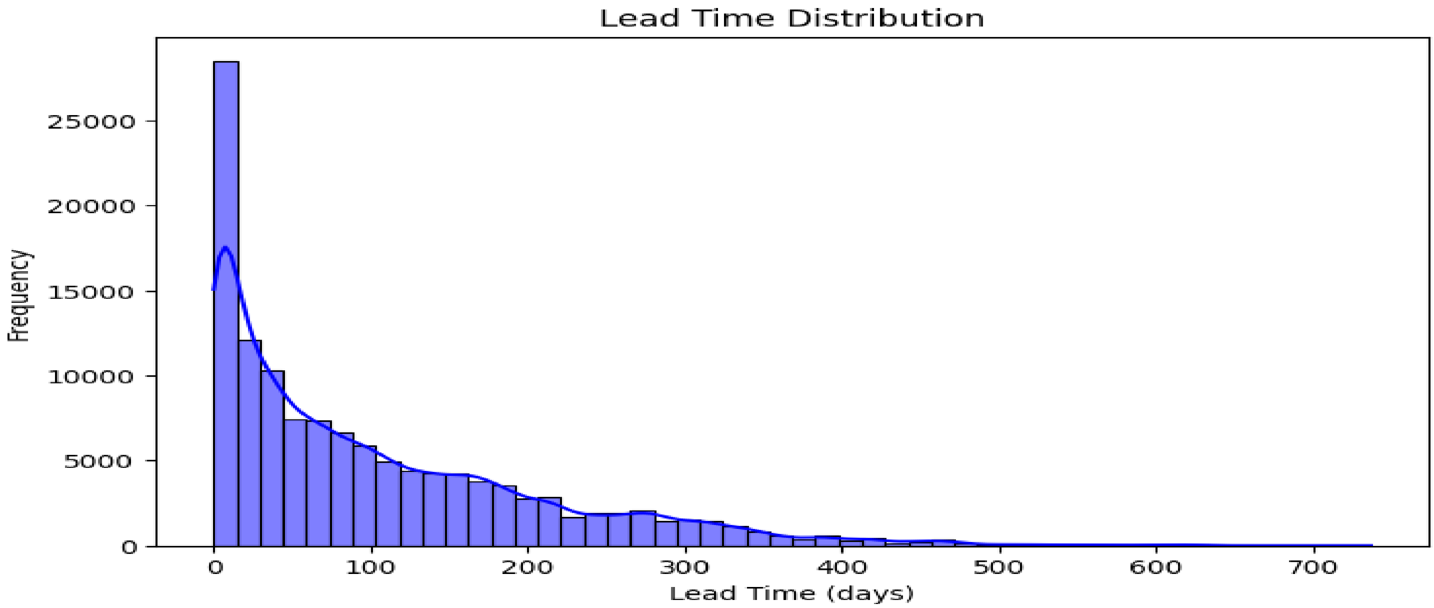
The hotel booking industry operates in a dynamic environment where customer behavior is influenced by various factors such as location, time of booking, and external economic conditions. Predicting and analyzing booking patterns is crucial for optimizing operations, reducing cancellation rates, and enhancing customer satisfaction. In this project, we focus on analyzing hotel booking data to identify trends and gain insights that can support future predictive analytics.The dataset used in this study contains extensive information about bookings made at both city hotels and resort hotels. We begin by loading and preprocessing the data using Python's Pandas library. Data cleaning involves handling missing values—replacing NaN entries in columns such as agent, company, and children with statistically derived values like zero or mean. Entries with no guests (adults, children, or babies) are removed as they do not represent valid bookings.The core of this project involves exploratory data analysis (EDA), a technique for summarizing the main characteristics of a dataset. Algorithms and techniques used include:Descriptive Statistics: For understanding distributions and central tendencies.Data Imputation: Using mean, mode, and conditional replacement for missing data.Data Visualization: Using Matplotlib and Seaborn to visualize patterns like lead time distribution, hotel type distribution, and top booking countries.Correlation Analysis: Using Pearson correlation matrix to evaluate relationships between numerical variables.This foundational analysis not only reveals booking trends but also provides a base for applying machine learning algorithms like logistic regression or decision trees in future predictive models.

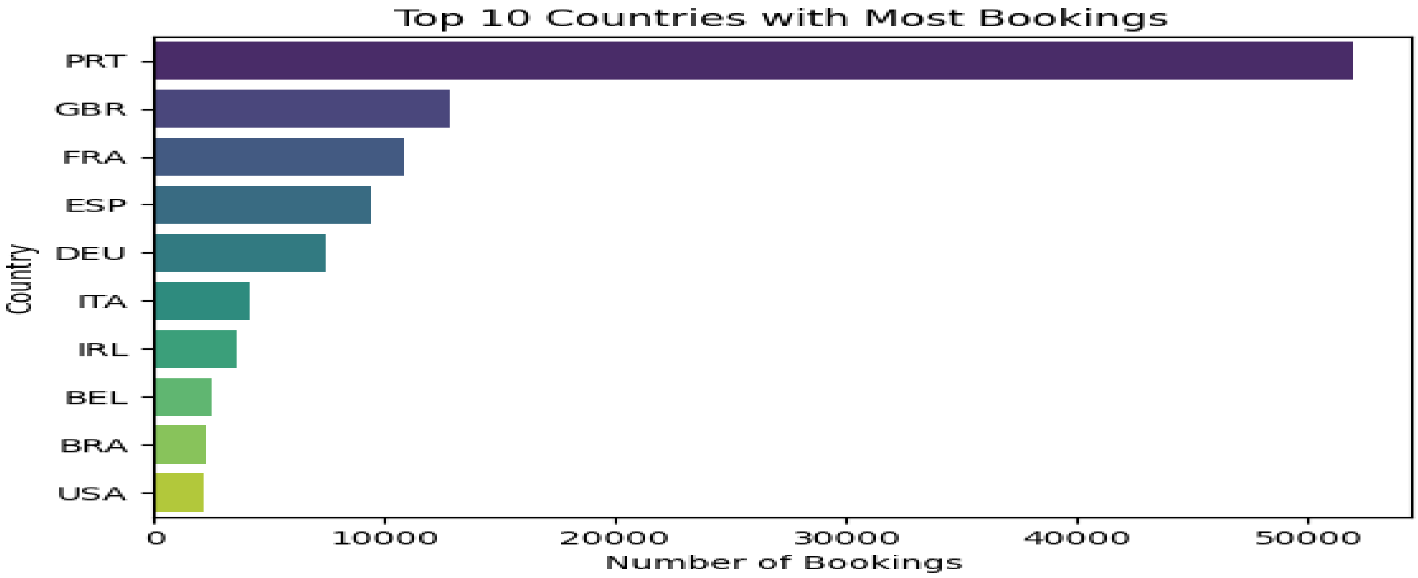
SOURCE CODE:

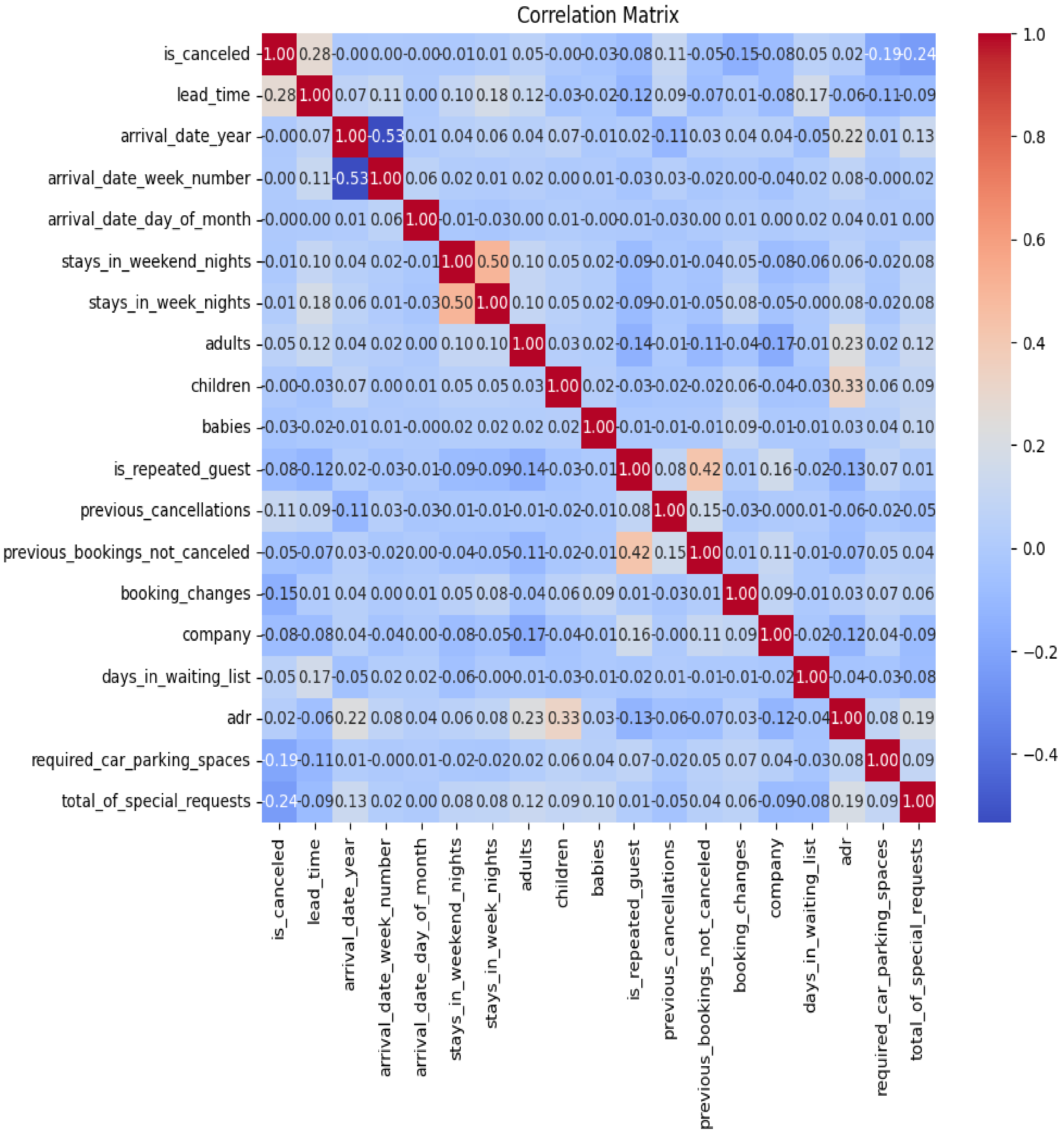
```
!pip install pycountry

Import numpy as np
Import pandas as pd
Import matplotlib.pyplot as plt
Import seaborn as sns
Import pycountry as pc

Pd.options.display.max_columns = None

## Importing Data
Data = pd.read_csv(‘hotel_bookings.csv’)

#!pip install pycountry

Import numpy as np
Import pandas as pd
Import matplotlib.pyplot as plt
Import seaborn as sns
Import pycountry as pc

Pd.options.display.max_columns = None

## Importing Data
Data = pd.read_csv(‘hotel_bookings.csv’)

## Show the first 5 rows of Data
Data.head()

## Copy the dataset
Df = data.copy()

## Find the missing value, show the total null values for each column and sort it in descending order
Df.isnull().sum().sort_values(ascending=False)[:10]

## Drop Rows where there is no adult, baby and child
Df = df.drop(df[(df.adults+df.babies+df.children)==0].index)

## If no id of agent or company is null, just replace it with 0
Df[[‘agent’,’company’]] = df[[‘agent’,’company’]].fillna(0.0)

## For the missing values in the country column, replace it with mode (value that appears most often)
Df[‘country’].fillna(data.country.mode().to_string(), inplace=True)
```

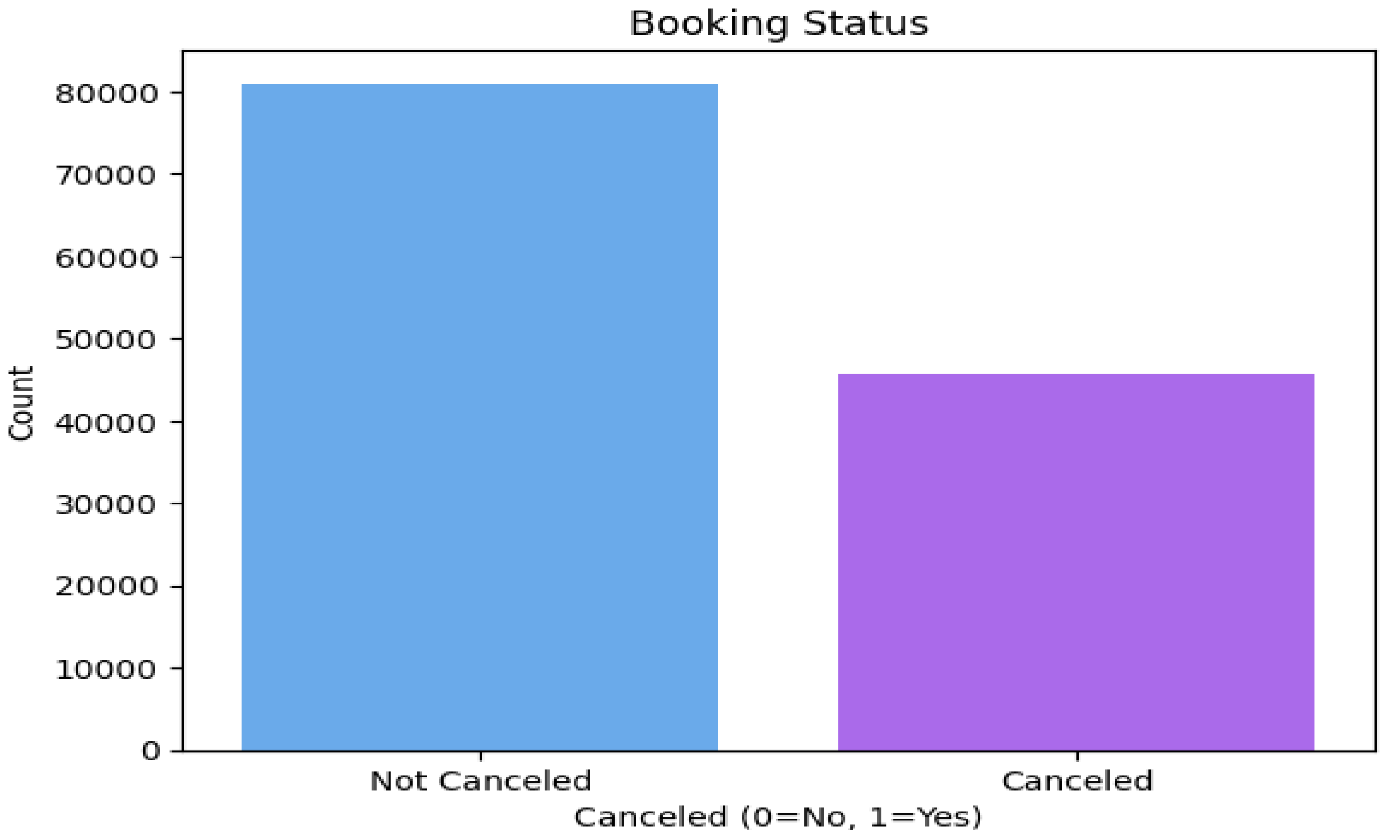
```
## For missing children value, replace it with rounded mean value
Df[‘children’].fillna(round(data.children.mean()), inplace=True)

# Distribution of hotel types
Hotel_counts = df[‘hotel’].value_counts()
Hotel_counts.plot.pie(autopct=’%1.1f%%’, figsize=(6, 6), title=’Distribution of Hotel Types’, ylabel=’’)
Plt.show()

# Top 10 countries with the most bookings
Top_countries = df[‘country’].value_counts().head(10)
Sns.barplot(x=top_countries.values, y=top_countries.index, palette=’viridis’)
Plt.title(‘Top 10 Countries with Most Bookings’)
Plt.xlabel(‘Number of Bookings’)
Plt.ylabel(‘Country’)
Plt.show()

# Lead time distribution
Plt.figure(figsize=(8, 5))
Sns.histplot(df[‘lead_time’], kde=True, bins=50, color=’blue’)
Plt.title(‘Lead Time Distribution’)
Plt.xlabel(‘Lead Time (days)’)
Plt.ylabel(‘Frequency’)
Plt.show()

# Count of reservation status
Sns.countplot(data=df, x=’is_canceled’, palette=’cool’)
Plt.title(‘Booking Status’)
Plt.xlabel(‘Canceled (0=No, 1=Yes)’)
Plt.ylabel(‘Count’)
Plt.xticks([0, 1], [‘Not Canceled’, ‘Canceled’])
Plt.show()
# Correlation heatmap (only numeric columns)
Plt.figure(figsize=(10, 8))

# Select only numeric columns
Numeric_df = df.select_dtypes(include=[‘number’])

# Compute the correlation matrix
Corr_matrix = numeric_df.corr()

# Plot the heatmap
Sns.heatmap(corr_matrix, annot=True, fmt=’.2f’, cmap=’coolwarm’)
Plt.title(‘Correlation Matrix’)
Plt.show()
```

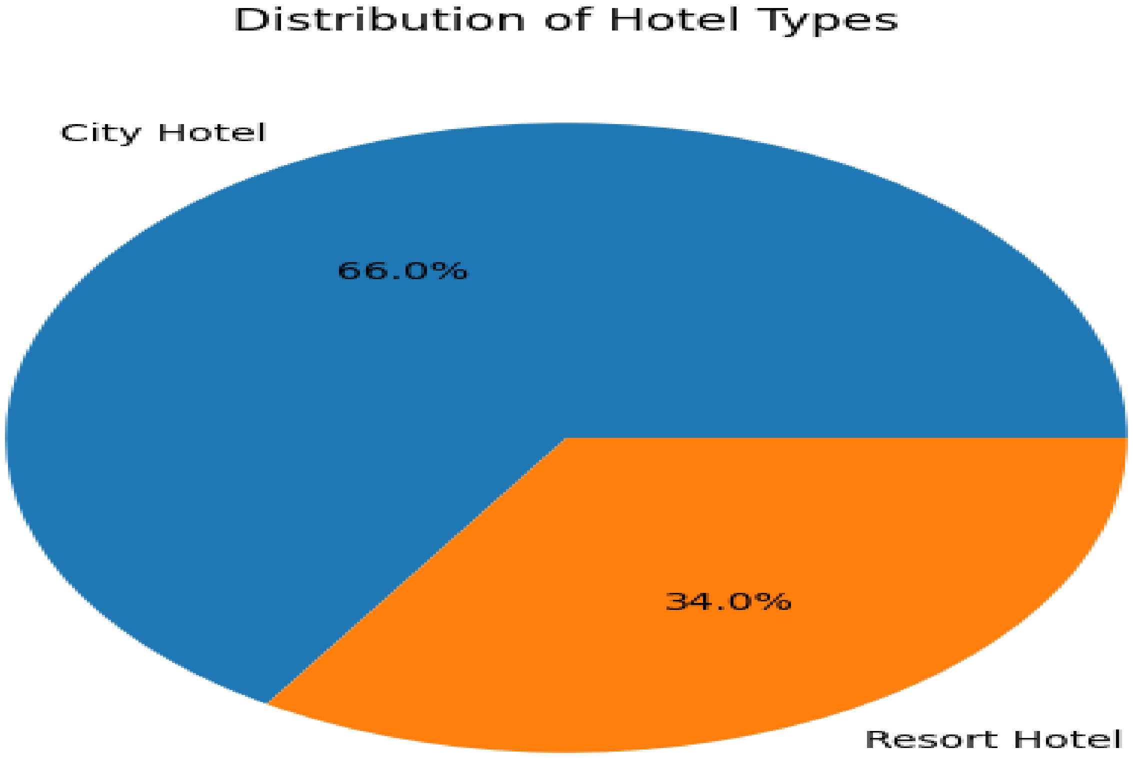
OUTPUT


Lead Time Distribution
Frequency
25000
20000
15000
10000
5000
0
0
100
200
300
400
500
600
700
Lead Time (days)


Top 10 Countries with Most Bookings
Country
PRT
GBR
FRA
ESP
DEU
ITA
IRL
BEL
BRA
USA
0
10000
20000
30000
40000
50000
Number of Bookings


Correlation Matrix
is_canceled 1.00 0.28 -0.00 0.00 -0.00 -0.01 0.01 0.05 -0.00 -0.03 -0.08 0.11 -0.05 -0.15 -0.08 0.05 0.02 -0.19 -0.24
lead_time 0.28 1.00 0.07 0.11 0.00 0.10 0.18 0.12 -0.03 -0.02 -0.12 0.09 -0.07 0.01 -0.08 0.17 -0.06 -0.11 -0.09
arrival_date_year -0.00 0.07 1.00 -0.53 0.01 0.04 0.06 0.04 0.07 -0.01 0.02 -0.11 0.03 0.04 0.04 -0.05 0.22 0.01 0.13
arrival_date_week_number 0.00 0.11 -0.53 1.00 0.06 0.02 0.01 0.02 0.00 0.01 -0.03 0.03 -0.02 0.00 -0.04 0.02 0.08 -0.00 0.02
arrival_date_day_of_month -0.00 0.00 0.01 0.06 1.00 -0.01 -0.03 0.00 0.01 -0.00 -0.01 -0.03 0.00 0.01 0.00 0.02 0.04 0.01 0.00
stays_in_weekend_nights -0.01 0.10 0.04 0.02 -0.01 1.00 0.50 0.10 0.05 0.02 -0.09 -0.01 -0.04 0.05 -0.08 -0.06 0.06 -0.02 0.08
stays_in_week_nights -0.01 0.18 0.06 0.01 -0.03 0.50 1.00 0.10 0.05 0.02 -0.09 -0.01 -0.05 0.08 -0.05 -0.00 0.08 -0.02 0.08
adults 0.05 0.12 0.04 0.02 0.00 0.10 0.10 1.00 0.03 0.02 -0.14 -0.01 -0.11 -0.04 -0.17 -0.01 0.23 0.02 0.12
children -0.00 -0.03 0.07 0.00 0.01 0.05 0.05 0.03 1.00 0.02 -0.03 -0.02 -0.02 0.06 -0.04 -0.03 0.33 0.06 0.09
babies -0.03 -0.02 -0.01 0.01 -0.00 0.02 0.02 0.02 0.02 1.00 -0.01 -0.01 -0.01 0.09 -0.01 -0.01 0.03 0.04 0.10
is_repeated_guest -0.08 -0.12 0.02 -0.03 -0.01 -0.09 -0.09 -0.14 -0.03 -0.01 1.00 0.08 0.42 0.01 0.16 -0.02 -0.13 0.07 0.01
previous_cancellations 0.11 0.09 -0.11 0.03 -0.03 -0.01 -0.01 -0.01 -0.02 -0.01 0.08 1.00 0.15 -0.03 -0.00 0.01 -0.06 -0.02 -0.05
previous_bookings_not_canceled -0.05 -0.07 0.03 -0.02 0.00 -0.04 -0.05 -0.11 -0.02 -0.01 0.42 0.15 1.00 0.01 0.11 -0.01 -0.07 0.05 0.04
booking_changes -0.15 0.01 0.04 0.00 0.01 0.05 0.08 -0.04 0.06 0.09 0.01 -0.03 0.01 1.00 0.09 -0.01 0.03 0.07 0.06
company -0.08 -0.08 0.04 -0.04 0.00 -0.08 -0.05 -0.17 -0.04 -0.01 0.16 -0.00 0.11 0.09 1.00 -0.02 -0.12 0.04 -0.09
days_in_waiting_list 0.05 0.17 -0.05 0.02 0.02 -0.06 -0.00 -0.01 -0.03 -0.01 -0.02 0.01 -0.01 -0.01 -0.02 1.00 -0.04 -0.03 -0.08
adr 0.02 -0.06 0.22 0.08 0.04 0.06 0.08 0.23 0.33 0.03 -0.13 -0.06 -0.07 0.03 -0.12 -0.04 1.00 0.08 0.19
required_car_parking_spaces -0.19 -0.11 0.01 -0.00 0.01 -0.02 -0.02 0.02 0.06 0.04 0.07 -0.02 0.05 0.07 0.04 -0.03 0.08 1.00 0.09
total_of_special_requests -0.24 -0.09 0.13 0.02 0.00 0.08 0.08 0.12 0.09 0.10 0.01 -0.05 0.04 0.06 -0.09 -0.08 0.19 0.09 1.00
is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
company
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Booking Status
Count
80000
70000
60000
50000
40000
30000
20000
10000
0
Not Canceled
Canceled
Canceled (0=No, 1=Yes)


Distribution of Hotel Types
City Hotel
66.0%
34.0%
Resort Hotel

CONCLUSION

In conclusion,The hotel booking prediction analysis conducted in this project reveals several valuable insights into customer behavior, booking patterns, and factors influencing cancellations. By performing data preprocessing and visualization, we gained a clearer understanding of how different features—such as lead time, hotel type, country of origin, and number of guests—impact booking outcomes.Key conclusions include:City hotels receive more bookings than resort hotels.The majority of bookings originate from a small group of countries.Higher lead time is often correlated with a higher likelihood of cancellations.Cancellations and confirmed bookings follow identifiable trends, useful for operational planning.Although no machine learning algorithm was directly implemented in this phase, the groundwork has been laid for classification and regression models that can predict cancellations, booking rates, or revenue forecasts. Algorithms like Logistic Regression, Random Forest, or XGBoost can be integrated later based on the features explored.