A

Course End Project Report on

**Analysis of Hotel Bookings**

Submitted in partial fulfillment of the requirements for the award of Degree of

**BACHELOR OF ENGINEERING**

**In**

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**By**

**B. Yuva Satya Kunaal (160122771031)**

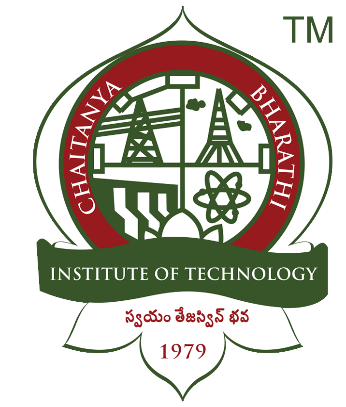
**Sasi Vakul Rithwik (160122771059)**

Under the Guidance of

**Dr K Radhika.**

Designation

Department of Artificial Intelligence and Data Science



**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY**

(AUTONOMOUS)

(Affiliated to Osmania University, Approved by AICTE and Accredited by NAAC)

Hyderabad – 500075, Telangana, India

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to Dr. K. Radhika, Professor in the AI&DS Department at CBIT (Chaitanya Bharathi Institute of Technology), for her invaluable guidance, mentorship, and support throughout the duration of this project. Dr. Radhika's expertise, encouragement, and constructive feedback have been instrumental in shaping our understanding of data analysis and visualization techniques, and her unwavering commitment to our academic growth has been truly inspiring. We are deeply grateful for her dedication and generosity in sharing her knowledge and expertise, which have contributed significantly to the success of this endeavour.

Additionally, we extend our heartfelt thanks to all our peers, colleagues, and friends who have provided support, encouragement, and assistance in various aspects of this project. Their collaboration, insights, and camaraderie have enriched our learning experience and made this journey all the more rewarding.

Thank you to everyone who has contributed to this project in any capacity. Your support and encouragement have been invaluable, and we are truly grateful for the opportunity to undertake this endeavour.

Sincerely,

B. Yuva Satya Kunaal, Sasi Vakul Rithwik  
Students of AI&DS – 1 (SEM 4).  
11 May 2024

**ABSTRACT**

In this project, we delve into the analysis and visualization of a comprehensive hotel booking dataset to uncover valuable insights and trends in booking behavior. Our exploration is structured around eight key questions addressing various aspects of hotel bookings, such as peak booking months, booking patterns based on day of the week, average booking lead time, correlation between stay length and room types, distribution of room types across different countries, cancellation rates on weekdays versus weekends, and reasons for cancellations across room types.

We begin by identifying the peak booking months and examine how this varies across different countries, shedding light on regional booking trends. Next, we investigate booking patterns based on the day of the week, exploring any noticeable trends that may emerge. Subsequently, we analyze the average booking lead time for different room types, providing insights into guests' planning behaviors.

Furthermore, we explore the correlation between stay length, number of guests, and room types, uncovering any notable associations. We then examine the distribution of room types across different countries, offering insights into accommodation preferences in various regions.

Moreover, we investigate cancellation rates between weekdays and weekends, discerning any disparities. Additionally, we analyze the distribution of booking lead time across different room types, providing insights into guests' booking habits for each type of accommodation.

Finally, we delve into the most common reasons for cancellations and assess whether they vary across different room types, providing hoteliers with actionable insights to optimize their operations and improve guest satisfaction.

Through our comprehensive analysis and visualizations, this project offers valuable insights into hotel booking behavior, enabling stakeholders to make informed decisions and enhance the guest experience.

**Table of CONTENTS**

**Page No’s**

|  |  |
| --- | --- |
| **Title Page** | **I** |
| **Acknowledgement** | **II** |
| **Abstract** | **III** |
| **Table of Contents** | **IV** |
| **Data Collection and Preparation** | **1** |
| **Questionnaire** | **3** |
| **Code and Output** | **4** |
| **Conclusion** | **27** |
| **References** | **28** |

**Dataset Collection and Preparation**

This dataset contains information on hotel bookings, encompassing details about guests, their reservations, and hotel attributes. It's a valuable resource for analysing and predicting trends in the hospitality industry.

* **hotel**: The type of hotel, either "City Hotel" or "Resort Hotel."
* **is\_canceled**: Binary value indicating whether the booking was cancelled (1) or not (0).
* **lead\_time**: Number of days between booking and arrival.
* **arrival\_date\_year**: Year of arrival date.
* **arrival\_date\_month**: Month of arrival date.
* **arrival\_date\_week\_number**: Week number of arrival date.
* **arrival\_date\_day\_of\_month**: Day of the month of arrival date.
* **stays\_in\_weekend\_nights**: Number of weekend nights (Saturday or Sunday) the guest stays.
* **stays\_in\_week\_nights**: Number of weekday nights (Monday to Friday) the guest stays.
* **adults**: Number of adults.
* **children**: Number of children.
* **babies**: Number of babies.
* **meal**: Type of meal booked.
* **country**: Country of origin.
* **market\_segment**: Market segment designation.
* **distribution\_channel**: Booking distribution channel.
* **is\_repeated\_guest**: Binary value indicating whether the guest is a repeated guest (1) or not (0).
* **previous\_cancellations**: Number of previous booking cancellations.
* **previous\_bookings\_not\_canceled**: Number of previous bookings not cancelled.
* **reserved\_room\_type**: Code of room type reserved.
* **assigned\_room\_type**: Code of room type assigned at check-in.
* **booking\_changes**: Number of changes/amendments made to the booking.
* **deposit\_type**: Type of deposit made.
* **agent**: ID of the travel agency.
* **company**: ID of the company.
* **days\_in\_waiting\_list**: Number of days in the waiting list before booking.
* **customer\_type**: Type of booking.
* **adr**: Average daily rate.
* **required\_car\_parking\_spaces**: Number of car parking spaces required.
* **total\_of\_special\_requests**: Number of special requests made.
* **reservation\_status**: Reservation last status.
* **reservation\_status\_date**: Date of the last status.
* **name**: Guest's name. (Not Real)
* **email**: Guest's email address.(Not Real)
* **phone-number**: Guest's phone number. (Not Real)
* **credit\_card**: Guest's credit card details. (Not Real)

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

url = "https://drive.google.com/file/d/1w5Mv9SYY8UT3Onvn1SoGJJP7yq8-8CU8/view?usp=sharing"

file\_id = url.split('/')[-2]

download\_url = f"https://drive.google.com/uc?id={file\_id}"

df = pd.read\_csv(download\_url)

**Questionnaire**

1. What are the peak booking months in the dataset, and how does it vary across different countries?
2. Are there any noticeable trends in booking patterns based on the day of the week?
3. What is the average booking lead time for different room types?
4. Is there a correlation between the length of stay, number of guests (adults, children, babies) and the room type booked?
5. How does the distribution of room types vary between different countries?
6. Are there any differences in cancellation rates between weekdays and weekends?
7. How does the distribution of booking lead time vary between different room types?
8. What are the most common reasons for cancellations, and do they vary across different room types?

**Code and Output**

Q.1

Code:

# Analysis: Calculate the total number of bookings for each month and analyze the distribution by country.

df = pd.read\_csv(download\_url)

#changing name of column = arrival\_date\_month

df.rename(columns={"arrival\_date\_month" : "Month"},inplace=True)

#inplace = True -> makes the column name modified in the same table rather than creating a new column

# calculating total bookings by using groupby

monthly\_booking = df.groupby(["Month","country"],).size().reset\_index(name="Total bookings")

# create a pivot table with columns = country & value = Total bookings

monthly\_booking\_pivot = monthly\_booking.pivot\_table(index=monthly\_booking["Month"],columns=df["country"],values="Total bookings",fill\_value=0)

# fill\_value = 0, because if the empty values are filled with any string(like None,no value,etc) it cannot be visualized.Thats why keeping 0 will have no issue.

#Changing the datatype of values to integer..as the values by default grouped as float

monthly\_booking\_pivot = monthly\_booking\_pivot.astype("int64")

# printing the table/dataframe

monthly\_booking\_pivot

# Visualization: Create a line plot to visualize the booking volume, segmented by country.

i=0

colors = ['red', 'blue', 'green', 'orange', 'purple', 'cyan', 'magenta', 'yellow', 'black', 'gray', 'brown', 'pink', 'lime', 'teal', 'lavender', 'coral', 'tan', 'olive']

plt.figure(figsize=(12,6))

for country in monthly\_booking\_pivot.columns:

  plt.plot(monthly\_booking\_pivot.index,monthly\_booking\_pivot[country],marker="o",ls="--",color=colors[i],label=country)

  i+=1

plt.legend(loc="best") #"best" : puts the legend position in best position inside graph

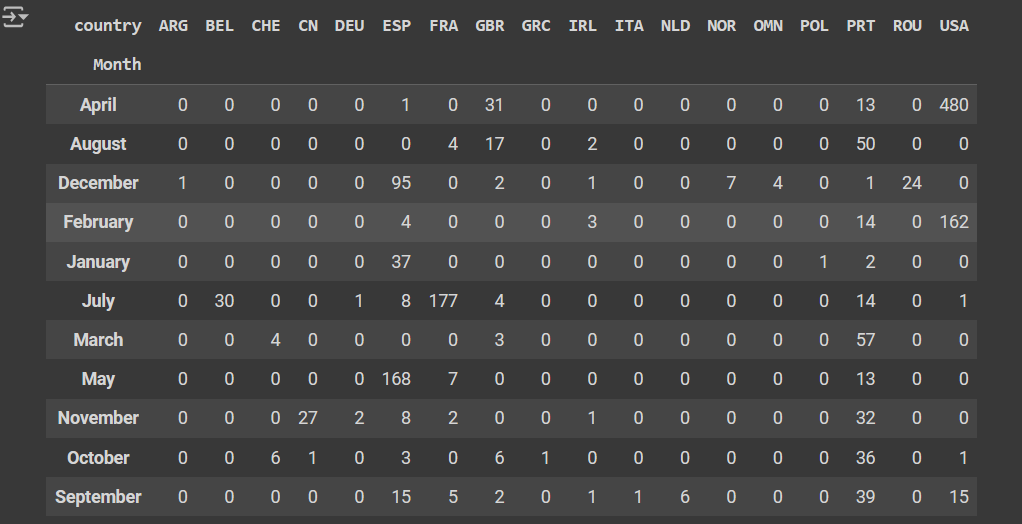
plt.xlabel("Month")

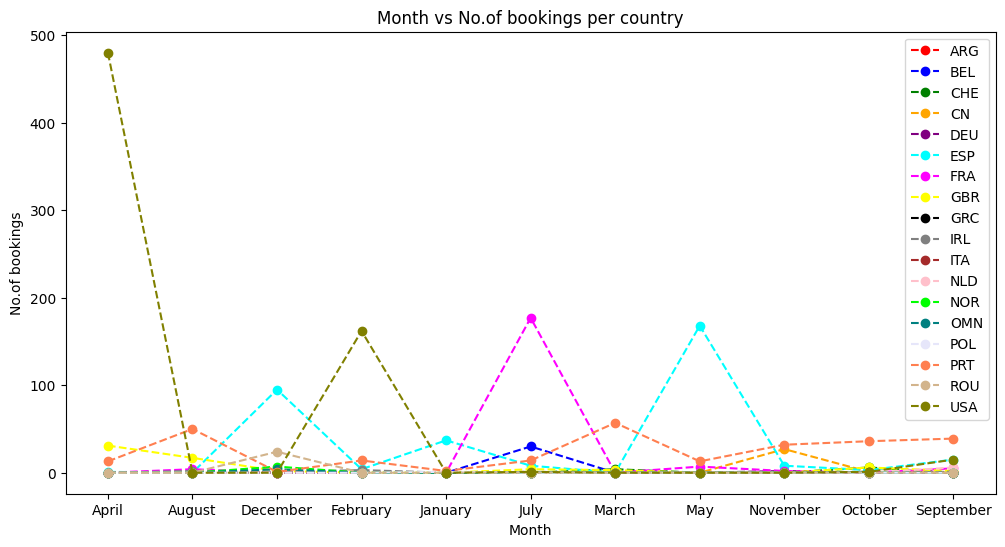
plt.ylabel("No.of bookings")

plt.title("Month vs No.of bookings per country")

plt.show()

Output:





Q2.

Code:

# Analysis: Determine the distribution of bookings by the day of the week and identify any trends or patterns.

df = pd.read\_csv(download\_url)

#convert "arrival\_date\_month" -> datetime format

#format = "%B" : converts string to datetime

df["arrival\_date\_month"] = pd.to\_datetime(df["arrival\_date\_month"], format='%B')

#extracts the day of the week for each date, returning an integer representation of the day of the week (0 for Monday, 1 for Tuesday, ..., 6 for Sunday).

df["day\_of\_week"] = df["arrival\_date\_month"].dt.dayofweek

day\_names = ["Mon","Tue","Wed","Thur","Fri","Sat","Sun"]

#using lambda function we replace each int value for their respective day name

df["day\_of\_week"] = df["day\_of\_week"].map(lambda x: day\_names[x])

# Calculating the number of bookings made on each day of the week

booking\_by\_day = df["day\_of\_week"].value\_counts().reindex(day\_names)

#creates a seperate dataframe/table for only day\_of\_week & No\_of\_bookings for better understanding and also for easy visualization

df1 = pd.DataFrame({"Day\_of\_week" : booking\_by\_day.index, "No\_of\_bookings" : booking\_by\_day.values})

#print(df1)

#since there is no booking in Friday, therefore it became null value,so replace it with 0

df1.fillna(0,inplace=True)

df1

# Visualization: Create a bar chart showing the number of bookings made on each day of the week.

x = df1["Day\_of\_week"]

y = df1["No\_of\_bookings"]

plt.figure(figsize=(8,4))

plt.bar(x,y,color="darkorange",edgecolor="black")

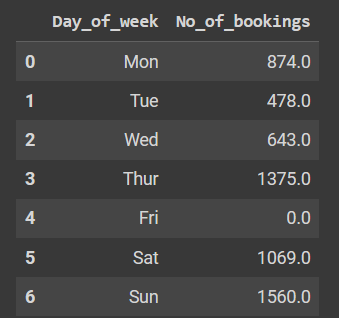
plt.title("Day vs Booking")

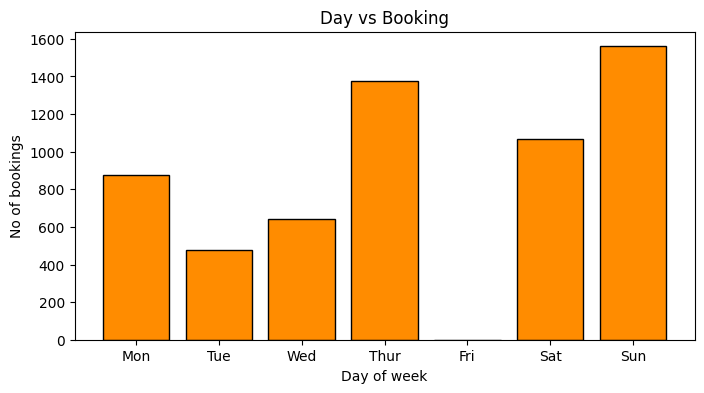
plt.xlabel("Day of week")

plt.ylabel("No of bookings")

plt.show()

Output:





Q3.

Code:

# Analysis: Calculate the average lead time for bookings of each room type and compare them.

df = pd.read\_csv(download\_url)

# Converts 'arrival\_date\_month' and 'arrival\_date\_year' columns to datetime and combine them to create a new 'arrival\_date' column.

df["arrival\_date"] = pd.to\_datetime(df["arrival\_date\_month"].astype(str)+" "+df["arrival\_date\_year"].astype(str))

# Converts 'reservation\_status\_date' column to datetime format

# format = "%d/%m/%y" -> d=day, m=month, y=year

df["reservation\_status\_date"] = pd.to\_datetime(df["reservation\_status\_date"],format="%d/%m/%y")

# Calculate lead time interms of days

lead\_time = (df['arrival\_date'] - df['reservation\_status\_date']).dt.days

# Replace negative lead times with 0 using a lambda function

lead\_time = lead\_time.apply(lambda x: max(x, 0))

# Create a new column 'lead\_time\_days' to store lead time in days

df["lead\_time\_days"] = lead\_time

# Calculate the average lead time for each reserved room type

average\_lead\_time = df.groupby("reserved\_room\_type")["lead\_time\_days"].mean().astype(int).reset\_index()

average\_lead\_time

# Visualization: Create a bar chart to visualize the average lead time for each room type.

x = average\_lead\_time["reserved\_room\_type"]

y = average\_lead\_time["lead\_time\_days"]

plt.figure(figsize=(8,4))

plt.bar(x,y,color="lightgreen",edgecolor="black")

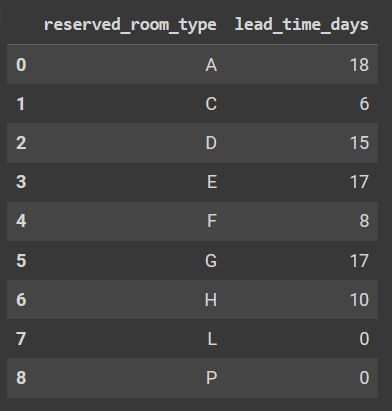
plt.title("Room type vs Lead time")

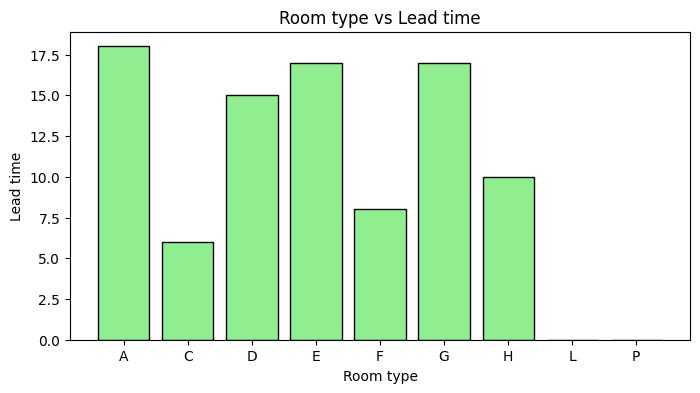
plt.xlabel("Room type")

plt.ylabel("Lead time")

plt.show()

Output:





Q4.

Code:

# Analysis: Calculate the correlation coefficient between the length of stay number of guests and room type.

# hint : Do i) Table for correletion b/w Number of guests and Length of stay.

#          ii) Table for correlation b/w Number of guests and Room type booked.

# And then join those two tables with common column as Number of guests

df = pd.read\_csv(download\_url)

correlation\_length\_stay\_guests = df[['adults', 'children', 'babies']].corrwith(df['adults'] + df['children'] + df['babies'])

# Step 2: Calculate correlation between number of guests and room type booked

room\_type\_dummy = pd.get\_dummies(df['reserved\_room\_type'])

correlation\_room\_type\_guests = pd.concat([df[['adults', 'children', 'babies']], room\_type\_dummy], axis=1).corr().loc[['adults', 'children', 'babies'], room\_type\_dummy.columns]

# Step 3: Join the two correlation tables based on the common column (number of guests)

correlation\_combined = correlation\_length\_stay\_guests.to\_frame(name='length\_of\_stay').join(correlation\_room\_type\_guests)

# Display the combined correlation table

print("Note : A negative correlation coefficient between the number of guests (adults, children, or babies) and a specific room type suggests \nthat as the number of guests in that category increases, the likelihood of booking that particular room type decreases.\n")

print("Correlation between Number of Guests, Length of Stay, and Room Type Booked :-\n")

correlation\_combined

# Visualization: Create a grouped bar chart to visualize the relationship between the number of guests and room type.

# Plotting

plt.figure(figsize=(10, 6))

# Number of bars

num\_bars = len(correlation\_combined.index)

bar\_width = 0.2

index = range(num\_bars)

# Plot bars for each type of guest

for i, (column\_name, column\_data) in enumerate(correlation\_combined.items()):

    plt.bar([x + i \* bar\_width for x in index], column\_data, bar\_width, label=column\_name)

# Customize the plot

plt.xlabel('Room Type')

plt.ylabel('Correlation Coefficient')

plt.title('Relationship Between Number of Guests and Room Type')

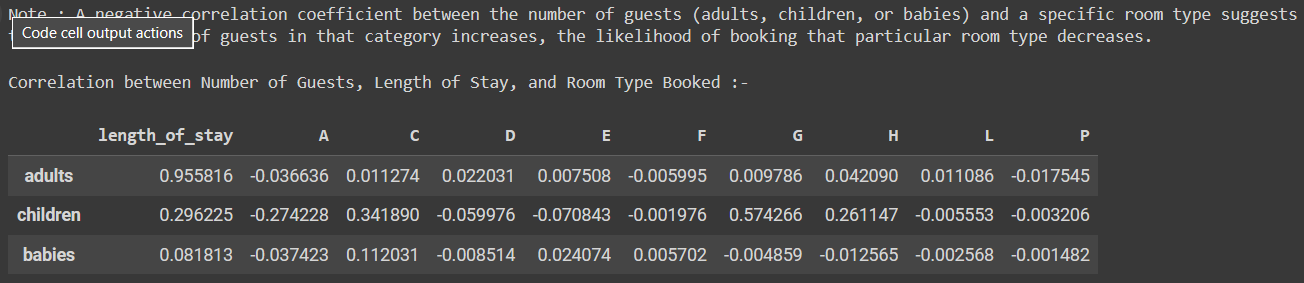
plt.xticks([x + (num\_bars - 1) \* bar\_width / 2 for x in index], correlation\_combined.index)

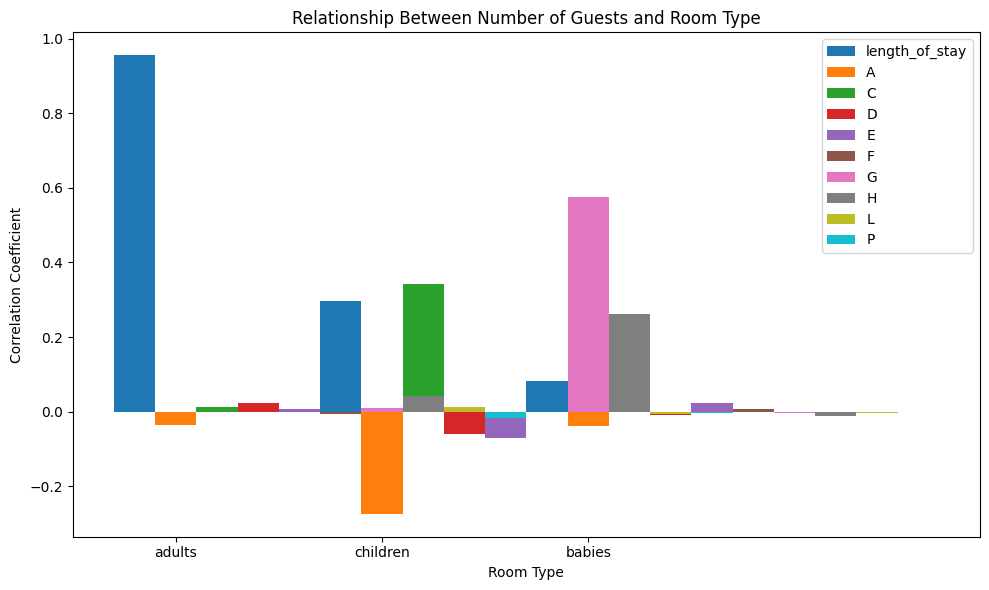
plt.legend()

# Show plot

plt.tight\_layout()

plt.show()





Q5.

Code:

# Analysis: Calculate the proportion of each room type booked from different countries and compare them.

df = pd.read\_csv(download\_url)

# Group the data by country and room type, and count the number of bookings

room\_type\_counts = df.groupby(['country', 'assigned\_room\_type']).size().unstack(fill\_value=0)

# Calculate the proportion of each room type booked from each country

room\_type\_proportions = room\_type\_counts.div(room\_type\_counts.sum(axis=1), axis=0)

print("Proportion of each room type booked from different countries :-\n")

room\_type\_proportions

# Visualization: Create a swarm plot to visualize the distribution of room types by country.

room\_type\_counts\_long = room\_type\_counts.stack().reset\_index()

room\_type\_counts\_long = room\_type\_counts.stack().reset\_index()

room\_type\_counts\_long.columns = ['Country', 'Room Type', 'Count']

# Suppress warnings

warnings.filterwarnings("ignore")

# Plotting the swarm plot

plt.figure(figsize=(8, 4))

sns.swarmplot(data=room\_type\_counts\_long, x='Room Type', y='Count', hue='Country', palette='Set2')

plt.title('Distribution of Room Types by Country')

plt.xlabel('Room Type')

plt.ylabel('Number of Bookings')

# Move the legend below the plot with horizontal labels

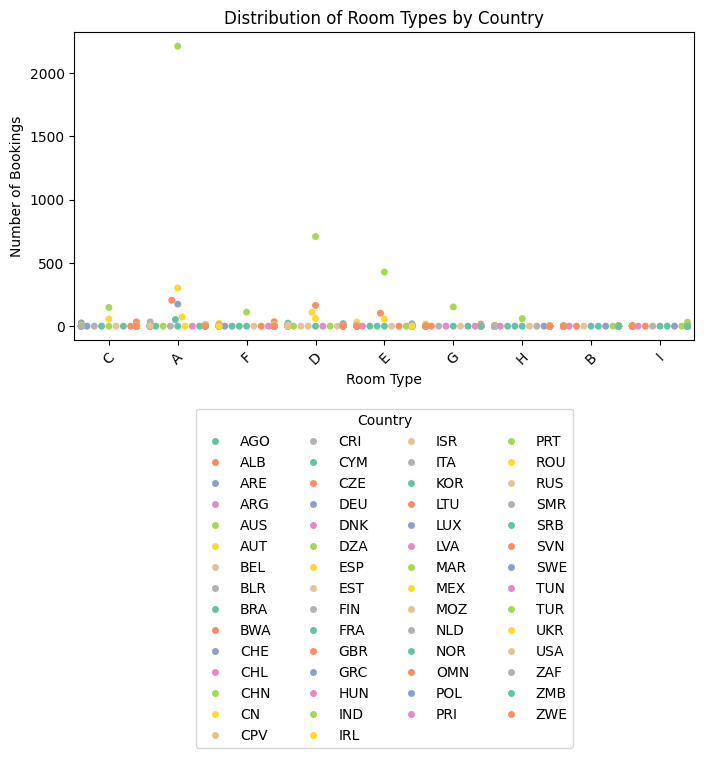
plt.legend(title='Country', bbox\_to\_anchor=(0.5, -0.2), loc='upper center', ncol=4)

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

Output:



Q6.  
Code:

# Analysis: Calculate the cancellation rates for bookings made on weekdays versus weekends and compare them.

df = pd.read\_csv(download\_url)

# Convert the 'reservation\_status\_date' column to datetime

df['reservation\_status\_date'] = pd.to\_datetime(df['reservation\_status\_date'])

# Extract the day of the week (0=Monday, 1=Tuesday, ..., 6=Sunday)

df['day\_of\_week'] = df['reservation\_status\_date'].dt.dayofweek

# Define a function to determine if the day is a weekday (Monday to Friday) or weekend (Saturday or Sunday)

def is\_weekend(day):

    return day >= 5  # Saturday and Sunday are weekend days (5 and 6)

# Create a new column to indicate if the booking was made on a weekday or weekend

df['weekday\_or\_weekend'] = df['day\_of\_week'].apply(lambda x: 'Weekend' if is\_weekend(x) else 'Weekday')

# Group the data by weekday\_or\_weekend and calculate cancellation rates

cancellation\_rates\_df = df.groupby('weekday\_or\_weekend')['is\_canceled'].mean()

print("Cancellation Rates :-\n")

cancellation\_rates\_df

# Visualization: Create separate pie charts for weekdays and weekends to visualize cancellation rates.

cancellation\_rates = pd.Series(cancellation\_rates\_df)

# Plotting the pie charts for cancellation rates

fig = plt.figure(figsize=(8, 4))

# Pie chart for weekdays

ax1 = fig.add\_subplot(121)

ax1.pie([1 - cancellation\_rates['Weekday'], cancellation\_rates['Weekday']], labels=['Not Canceled', 'Canceled'], colors=['skyblue', 'salmon'], autopct='%1.1f%%', startangle=140, shadow=True)

ax1.set\_title('Cancellation Rates - Weekdays')

# Pie chart for weekends

ax2 = fig.add\_subplot(122)

ax2.pie([1 - cancellation\_rates['Weekend'], cancellation\_rates['Weekend']], labels=['Not Canceled', 'Canceled'], colors=['skyblue', 'salmon'], autopct='%1.1f%%', startangle=140, shadow=True)

ax2.set\_title('Cancellation Rates - Weekends')

plt.tight\_layout()

plt.show()

Output:

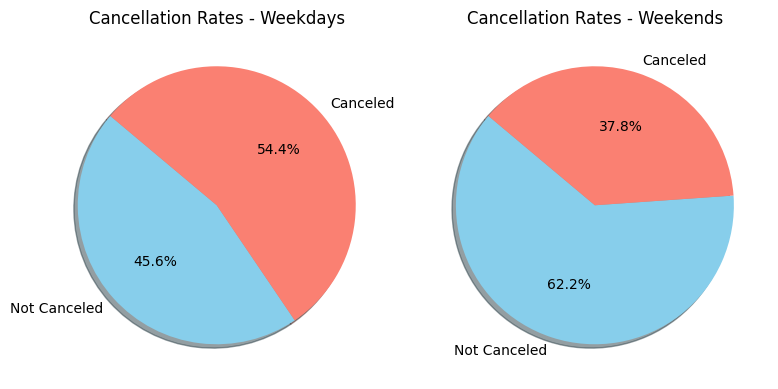
Cancellation Rates :-

weekday\_or\_weekend

Weekday 0.544091

Weekend 0.377736

Name: is\_canceled, dtype: float64



Q7.

Code:

# Analysis: Calculate the average lead time for bookings of each room type and compare them.

df = pd.read\_csv(download\_url)

# Convert the 'reservation\_status\_date' column to datetime

df['reservation\_status\_date'] = pd.to\_datetime(df['reservation\_status\_date'])

# Convert 'arrival\_date\_year' and 'arrival\_date\_month' to a single datetime column

df['arrival\_date'] = pd.to\_datetime(df['arrival\_date\_year'].astype(str) + '-' + df['arrival\_date\_month'], format='%Y-%B')

# Calculate lead time (arrival date - reservation date)

df['lead\_time'] = (df['arrival\_date'] - df['reservation\_status\_date']).dt.days

# Filter out rows with negative lead time values

df = df[df['lead\_time'] >= 0]

# Group the data by 'reserved\_room\_type' and calculate the mean lead time for each room type

average\_lead\_time = df.groupby('reserved\_room\_type')['lead\_time'].mean()

# Display the result

print("Average Lead Time for Bookings of Each Room Type:")

print(average\_lead\_time)

# Visualization: Create a violin plot to visualize the distribution of lead time by room type

plt.figure(figsize=(8, 4))

sns.violinplot(data=df, x='reserved\_room\_type', y='lead\_time', palette='muted', inner='quartile')

plt.title('Distribution of Lead Time by Room Type')

plt.xlabel('Room Type')

plt.ylabel('Lead Time (Days)')

plt.xticks(rotation=45)  # Rotate x-axis labels for better readability

plt.tight\_layout()

plt.show()

Output:

Average Lead Time for Bookings of Each Room Type:

reserved\_room\_type

A 81.693168

C 68.489796

D 73.502203

E 82.608844

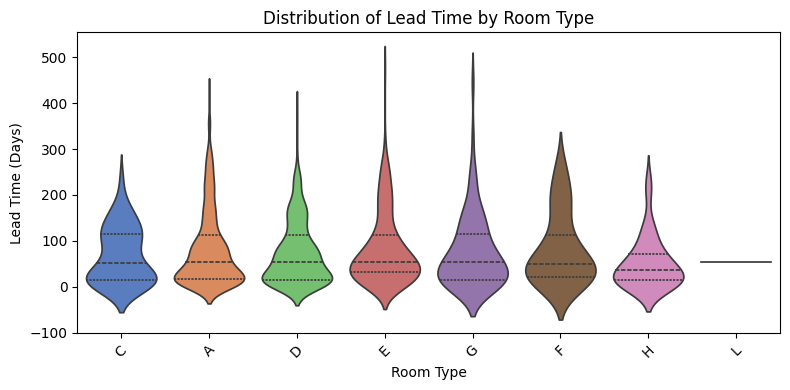
F 83.268293

G 76.630435

H 51.805556

L 54.000000

Name: lead\_time, dtype: float64



Q8.

Code:

# Analysis: Analyze the distribution of cancellation reasons and compare them between different room types.

df = pd.read\_csv(download\_url)

# Filter out canceled bookings

canceled\_bookings = df[df['is\_canceled'] == 1]

# Group the data by room type and cancellation reason

cancellation\_distribution = canceled\_bookings.groupby(['reserved\_room\_type', 'reservation\_status'])['is\_canceled'].count().unstack().fillna(0)

# Display the distribution of cancellation reasons by room type

print("Distribution of Cancellation Reasons by Room Type :-\n")

cancellation\_distribution

# Visualization: Create a pie chart to visualize the distribution of cancellation reasons by room type.

# Filter out canceled bookings

canceled\_bookings = df[df['is\_canceled'] == 1]

# Group the data by room type and cancellation reason

cancellation\_distribution = canceled\_bookings.groupby(['reserved\_room\_type', 'reservation\_status'])['is\_canceled'].count().unstack().fillna(0)

# Plot a pie chart for each room type

plt.figure(figsize=(10, 8))

# Calculate the number of rows and columns dynamically

num\_rows = (cancellation\_distribution.shape[0] + 1) // 2

num\_columns = 2

for i, room\_type in enumerate(cancellation\_distribution.index):

    plt.subplot(num\_rows, num\_columns, i+1)

    plt.pie(cancellation\_distribution.loc[room\_type], labels=cancellation\_distribution.columns, autopct='%1.1f%%', startangle=140)

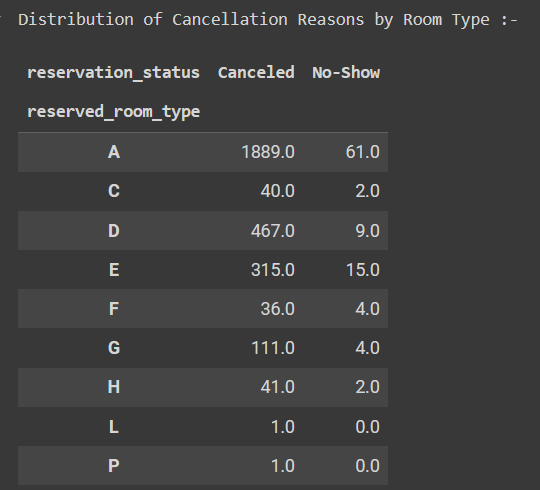
    plt.title(f'Cancellation Reasons - {room\_type}')

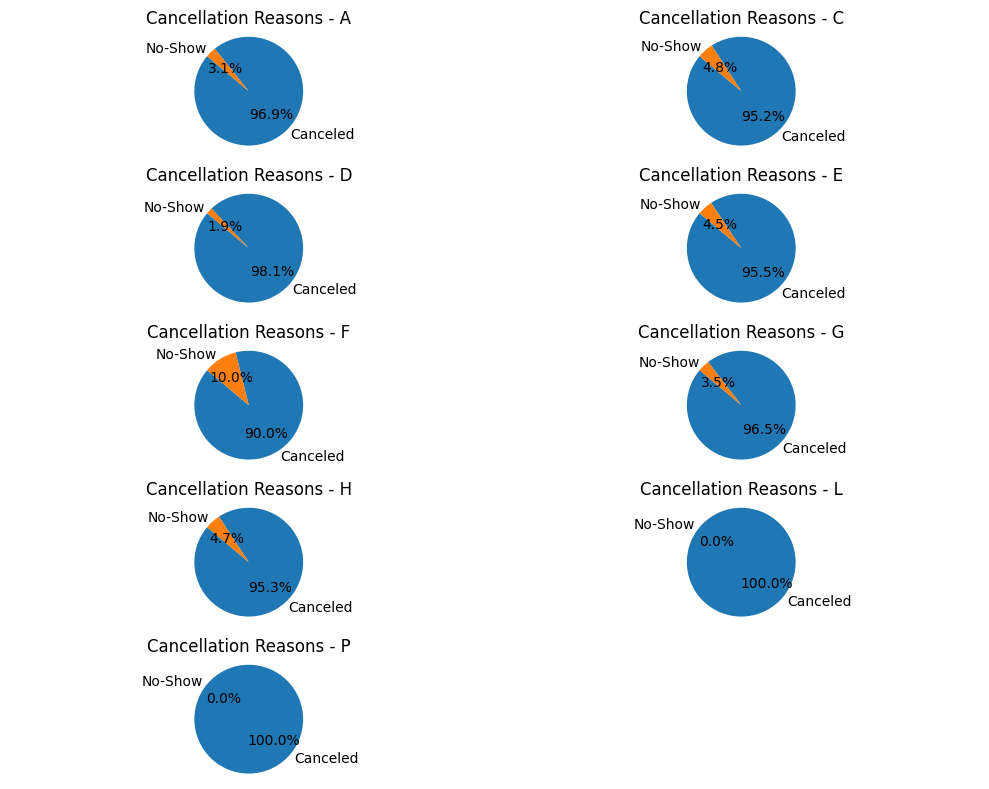
    plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle

plt.tight\_layout()

plt.show()

Output:





**Conclusion**

1. Better understanding of the monthly demand for hotel bookings enabling better resource allocation and pricing strategies tailored to different regions.
2. We have identified specific days that witness higher booking activity, allowing hotels to optimize staffing levels and promotional efforts on those days.
3. Through python coding, we have provided insights into booking behaviours and preferences for different room types, helping hotels anticipate demand and manage inventory effectively.
4. Better understanding of guest preferences for room types based on group size, facilitating room allocation and inventory management.
5. We have Identified popular room types among guests from different countries, guiding decisions on room allocation and customization for different markets.
6. Guides decision-making on cancellation policies and revenue management strategies tailored to weekday and weekend bookings.
7. Provides insights into booking behaviours for different room types, aiding in inventory management and pricing decisions.
8. Better understanding of the underlying reasons for cancellations and tailoring strategies to minimize cancellations, especially for specific room types.

**References**

[**https://www.kaggle.com/datasets/saadharoon27/hotel-booking-dataset?resource=download**](https://www.kaggle.com/datasets/saadharoon27/hotel-booking-dataset?resource=download)

**https://pandas.pydata.org/docs/user\_guide/index.html**

**https://matplotlib.org/stable/plot\_types/index.html**

**https://seaborn.pydata.org/tutorial.html**

**https://numpy.org/doc/stable/user/**