HOTEL BOOKING ANALYSIS

Exploratory Data Analysis

Team Members:

Sunil K

Souvik Karmakar

Mohit Wagish

Project Summary

Hotel booking data analysis provides insights for optimizing pricing, offerings, and strategies. Key factors include lead time, guest count, meal type, origin country, and market segment. Understanding booking patterns, distribution channels, and deposit types informs decision-making for revenue and customer satisfaction. The goal is to adjust pricing, improve offerings, optimize marketing, and enhance customer relationships for increased profitability.

Problem Statement

The main aim of looking at the hotel booking data is to better understand how customers behave. We want to find out trends and patterns in the bookings that can help us make smarter decisions based on data.

By doing this, we hope to improve how well the hotel is doing and make more money. We'll be looking at things like how far in advance people book, how many people are staying, what kind of meals they prefer, where they're from, and what types of bookings they make. We'll also check patterns like how people pay, the status of their reservation, and what kind of customers they are.

The insights we get from all of this will help us set better prices, offer things people want, and improve how we reach customers. Ultimately, our goal is to make more money and make sure customers are happy with their stay at the hotel.

Attributes used in the analysis

This Dataset consists of 119390 rows and 32 columns

- Hotel
- Is_cancelled
- Arrival_date_year
- Arrival_date_month
- Arrival_date_week_number
- Day_of_the_month
- Stays_in_weekend_nights
- Stays_in_week_nights
- Adults
- Children
- Babies
- Meals
- Country
- Market_segment
- Distribution Channel
- Is_repeated_guest
- Previous_cancellations
- Previous_bookings_not_canceled
- Reserved_room_type
- Assigned_room_type
- Booking_changes
- Deposit_type
- Agent
- Lead_time
- Days_in_waiting_list
- Customer_type
- Adr
- Required_car_parking_spaces
- Booking changes
- Reservation_status
- reservation_status_date

Attribute Description

***Categorical Variables* (Ordinal)**:

meal: Represents meal plans. Ordinal because there is likely an order or hierarchy among the meal types (e.g., 'BB' before 'HB' in terms of inclusiveness).

market_segment: Represents market segments. Different categories likely represent distinct groups.

distribution_channel: Represents distribution channels. Different categories represent different channels for booking.

deposit_type: Represents deposit types. Different categories represent different deposit policies.

reservation_status: Represents reservation status. Different categories represent different stages of the reservation process.

hotel: Represents the type of hotel (City Hotel or Resort Hotel). While there are only two categories.

arrival_date_year: Represents the year of arrival. While expressed as numbers, it is categorical as the years do not have a numerical relationship.

market_segment: This variable represents different segments, and the categories.

adults, children, babies: Represents the counts of adults, children, and babies respectively.

agent, company: Categorical.

required_car_parking_spaces, total_of_special_requests: Categorical.

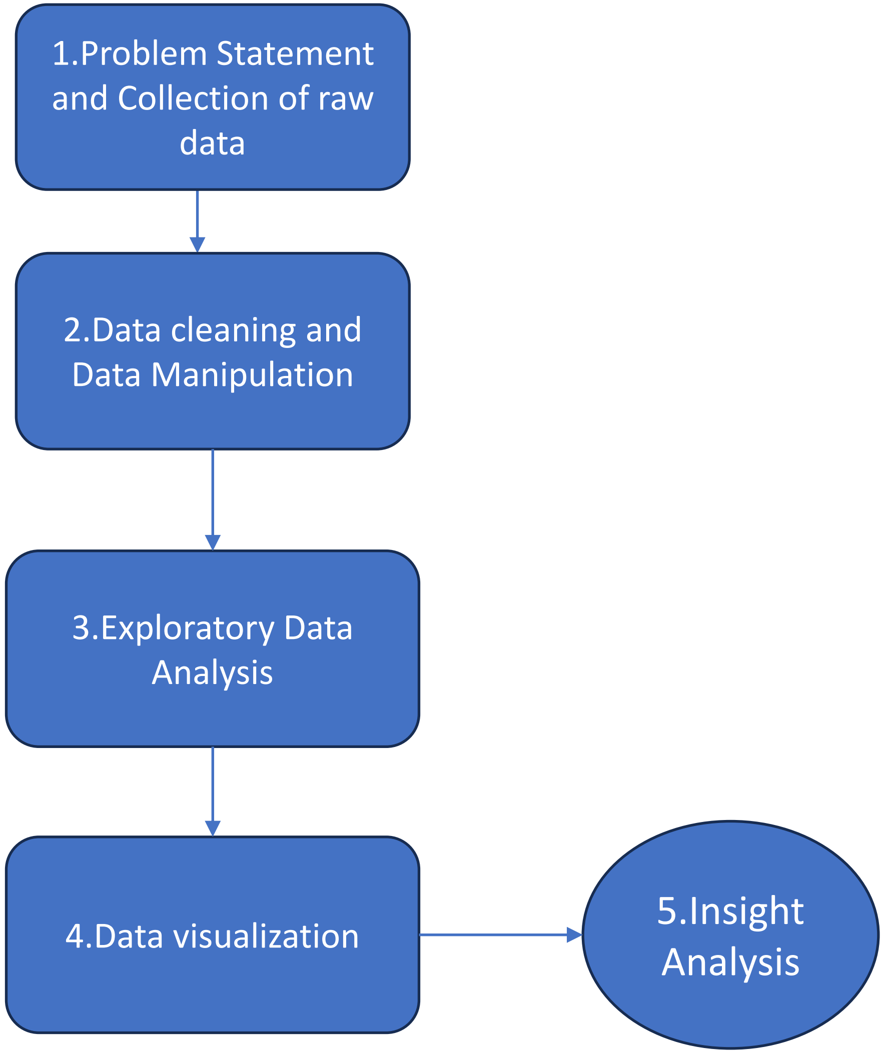
Continuous Variables:

stays_in_weekend_nights: Represents the number of nights stayed over weekends.

days_in_waiting_list: Represents the number of days a booking was in the waiting list.

adr: Represents the average daily rate.

Steps involved



1.Problem Statement and Collection of raw data
2.Data cleaning and Data Manipulation
3.Exploratory Data Analysis
4.Data visualization
5.Insight Analysis

Steps Description :-

1.Firstly identify the problem or the analysis to be done and then collect the data regarding it.

2.Raw data will be having a lot of null values, outliers and out of format data. Thus, it must be cleaned before being worked on.

3.Using the technical skills(python) we need to code to examine and extract results.

4.The result will then be displayed in a visualization style as in graphs for a better understanding.

5. Using the visualization and EDA results, we can wrap up the analysis.

Exploring The Data

Before starting the Exploratory Data Analysis (EDA) process, we typically need to import various libraries in Python to facilitate data manipulation and , visualization, and analysis. Here are some of the essential libraries we used in this EDA project:

```
import numpy as np
```

Import for numerical operations on arrays and matrices.

```
import pandas as pd
```

Import for data manipulation and analysis using dataframes.

```
import matplotlib.pyplot as plt
```

Import for basic data visualization, such as line plots, scatter plots, and histograms.

```
import seaborn as sns
```

Import for statistical data visualization, built on top of Matplotlib.

```
import plotly.express as px
```

Import for creating a variety of interactive visualizations.

Data Wrangling

Data Cleaning:-

- we have dropped "company" column from our further analysis as it has a large number of Null values that will negatively affect our ability to analyze the data further.
- Also we dropped duplicate values , along with that we have dropped ‘adult’,’children’,’babies’ columns because we have added one column as ‘Total_guest’ by adding ‘adult’ , ’children’ and ‘babies’ column.
- Numerical columns with nominal null values have been manipulated by filled them with their mean values.

Data Manipulation:-

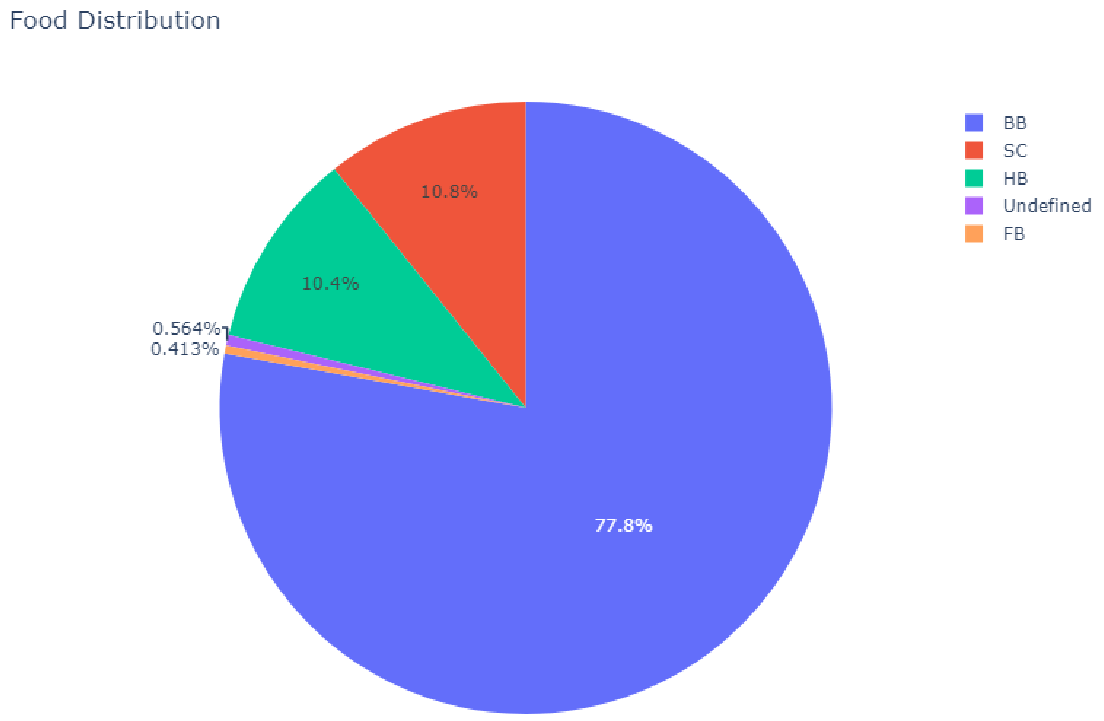
Combining columns for an effective study

➔Total_guest=babies+children+adult

➔total_stays= Stays_in_weekend_nights+Stays_in_week_nights

Data Visualization

Univariate Analysis



Food Distribution

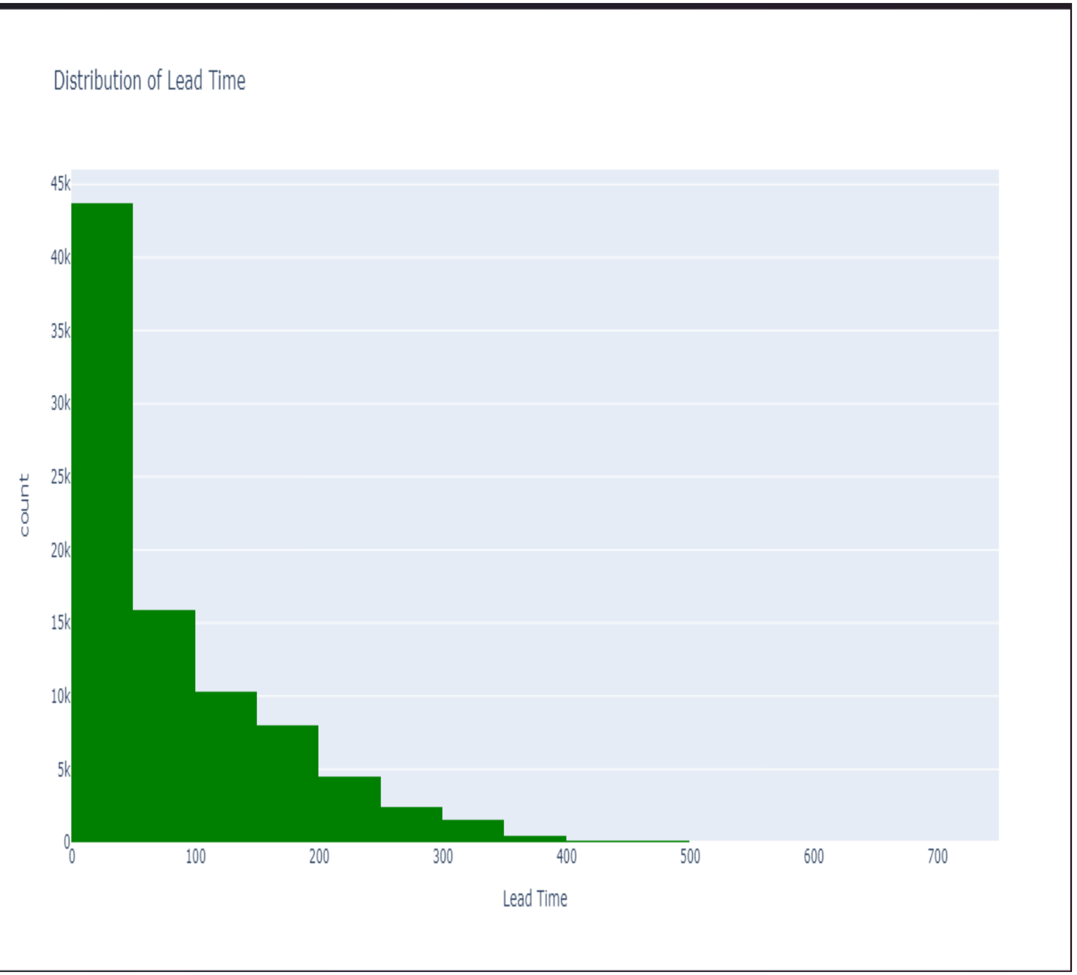
- The pie chart was selected to visually represent the distribution of meal plans in the dataset.

- From the chart, it is evident that the 'BB' (Bed and Breakfast) meal type dominates, constituting the majority of meal plan selections.

- This insight can positively impact business strategies by allowing for resource optimization, targeted marketing efforts, and menu adjustments to align with customer preferences.

- However, it also highlights a potential risk of limited diversity in meal types, signaling the need for ongoing assessment and potential menu enhancements to ensure sustained customer satisfaction and business growth.

Distribution of Lead Time



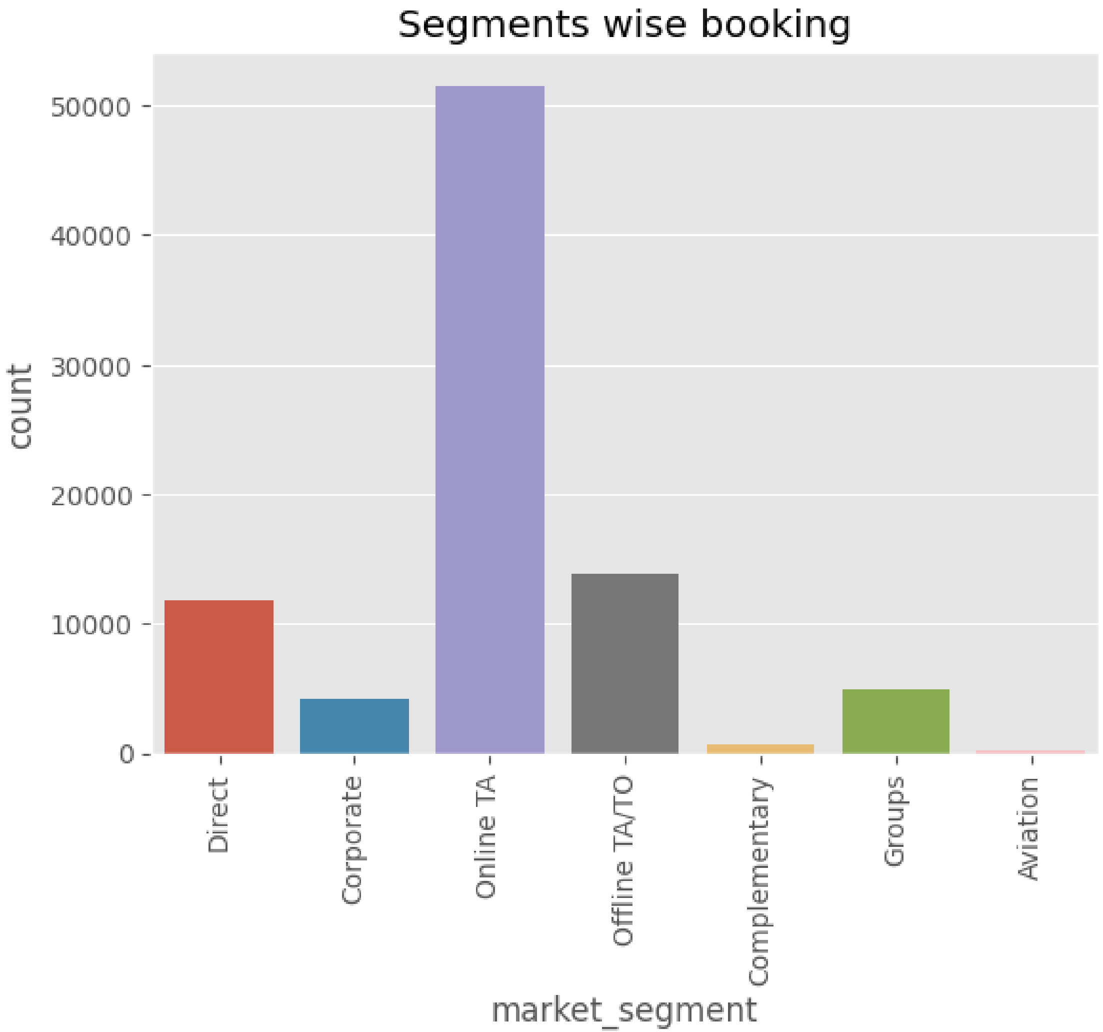
- The histogram was chosen to depict the distribution of lead time in hotel bookings, offering insights into common booking patterns.

- The visualization reveals lead time concentrations, common intervals, and potential outliers, aiding in operational optimization and targeted marketing.

- Positive impacts include improved resource allocation and strategic marketing. Negative growth may occur with a concentration of last-minute bookings, impacting revenue stability and operational efficiency.

- Mitigation strategies, such as adjusted pricing and marketing efforts, can address potential challenges associated with specific lead time patterns.

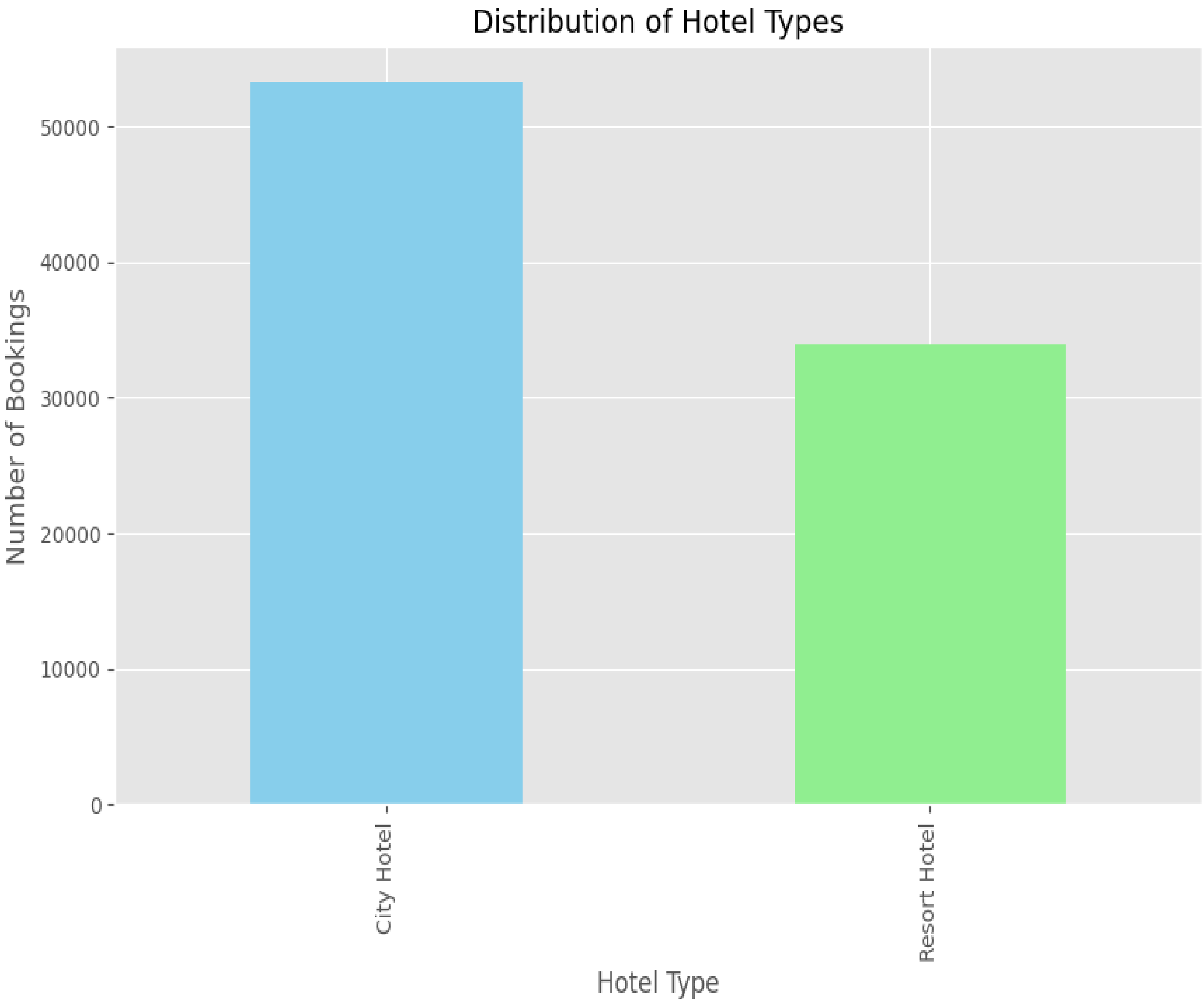
Market Segment wise Booking



Insight:-

The Visualization reveals Indirect bookings through online and offline travel agents are higher compared to direct bookings and same is the case with group bookings which are also high.

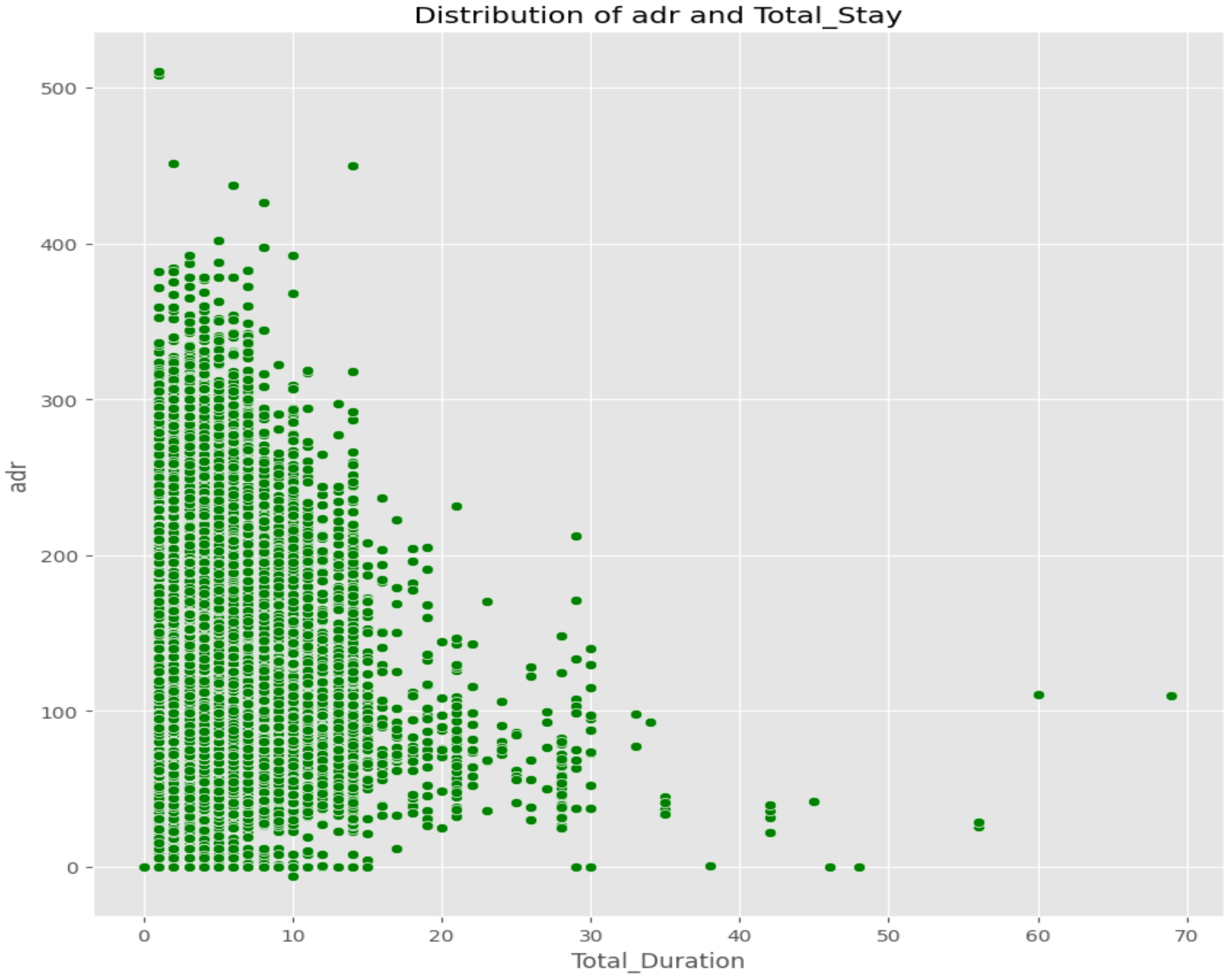
Hotel Type Distribution



- We have chosen bar chart because bar charts are effective for visualizing the distribution of nominal categorical variables.

- The Hotel Type Distribution chart reveals that City Hotels have significantly more bookings compared to Resort Hotels, indicating a higher demand or preference for City Hotels in the dataset.

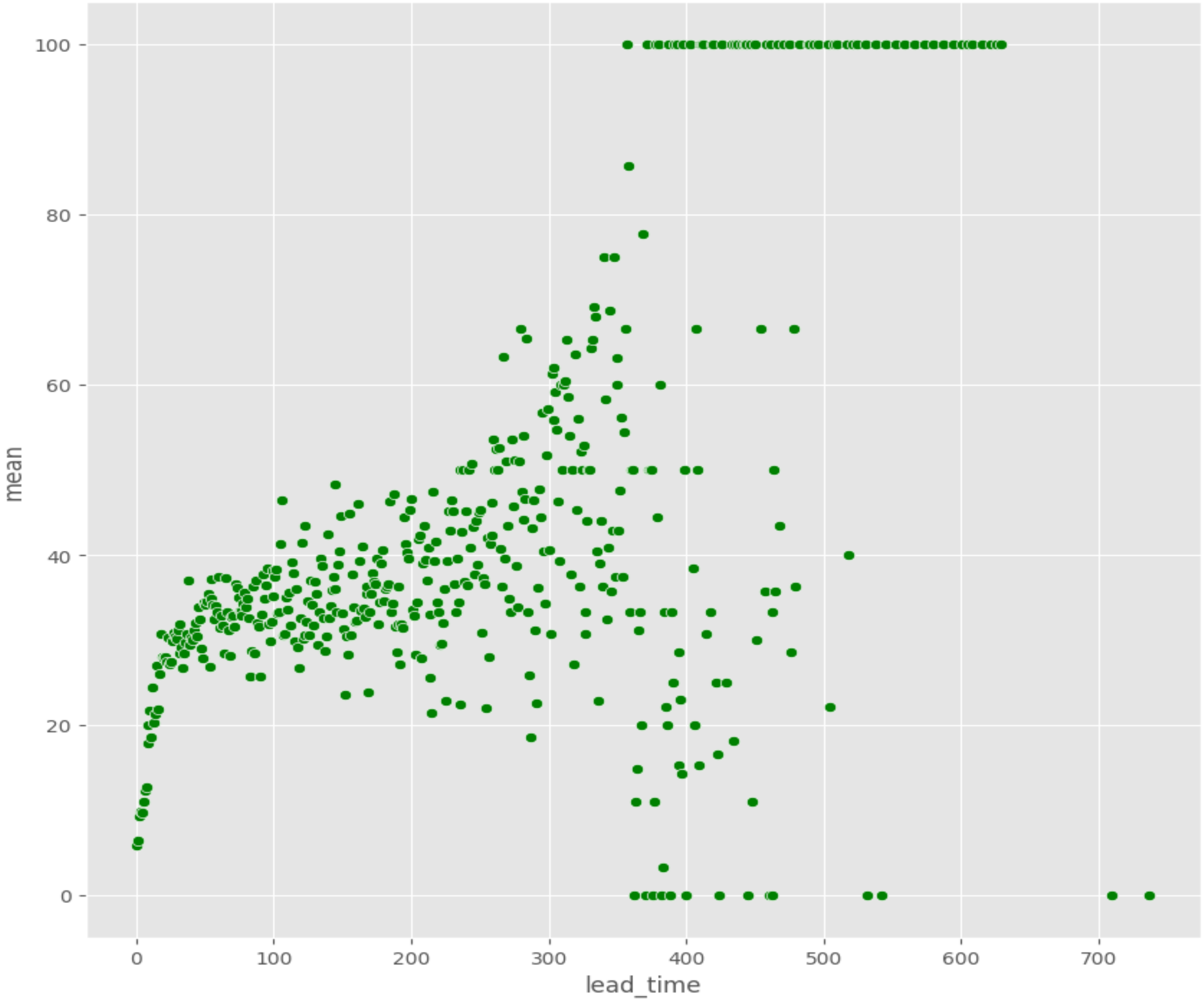
Bi-Variate Analysis:-
Distribution of adr and Total_Duration



- We have used scatter plot to see the correlation between Total_stay and adr,and also we determined how Total_stay impacts adr.

- We can observe from the scatter plot that the adr falls as the length of Total_Duration grows. This implies that a better offer for the customer can be finalized for a longer stay. In practical terms, as the ADR decreases, the Total Duration tends to increase, and vice versa.

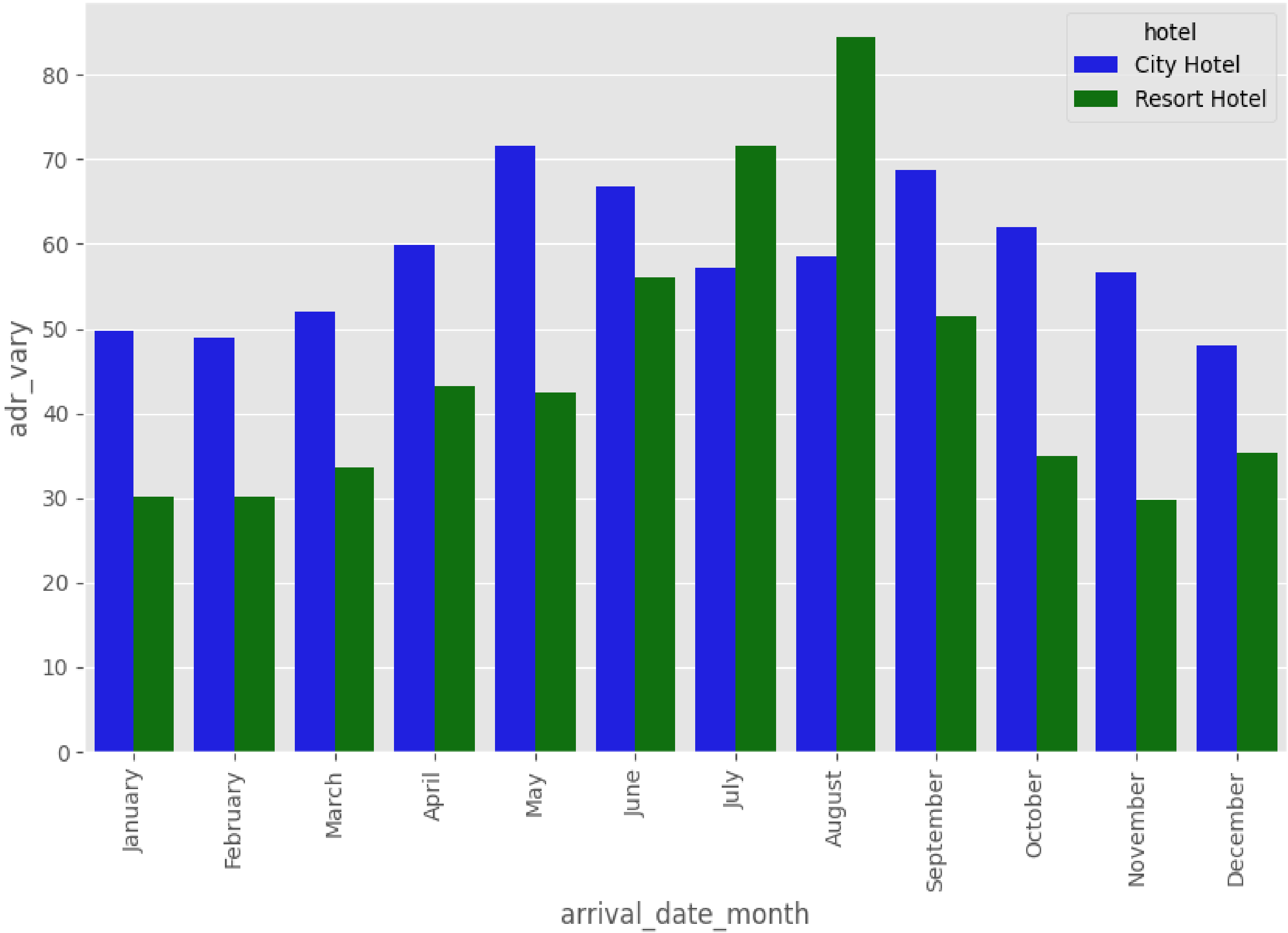
Cancellation based on Lead_time Analysis



- We have used scatter plot to check the relation between lead time and cancellation.

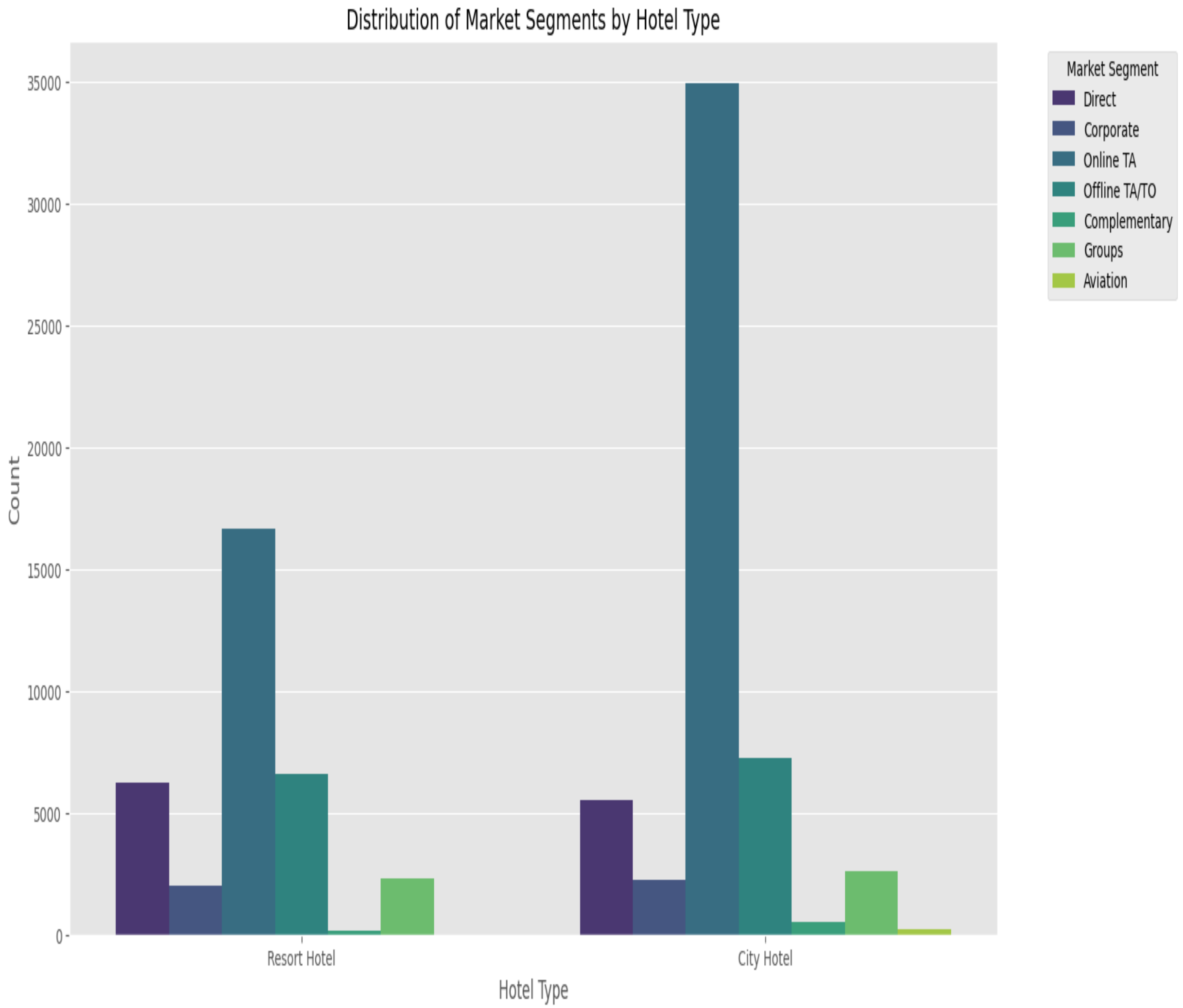
- it suggests that as the lead time (the time between booking and the actual stay) increases, the likelihood of cancellations also increases. In other words, there is a tendency for customers who book farther in advance to be more likely to cancel their reservations.

Price fluctuation over the year analysis



- We picked grouped bar chart because , Grouped bar charts are well-suited for comparing multiple categories (hotels) across different subcategories (months) in a visually intuitive way, also Bar charts are effective for visualizing variations in a numerical variable (ADR) across different groups, providing a clear representation of how values change over time.
- The chart provides insights into the average daily rate variation for each hotel type over the months.
- It allows for the identification of months with significant ADR fluctuations and the comparison of these fluctuations between the two hotel categories. Patterns or trends in ADR variations can be easily observed, providing valuable insights into the seasonal patterns of pricing.
- From the above plot we can see that in the month of august Resort hotel has more price flactuation and on themonth of may City hotel has more price flactuation.

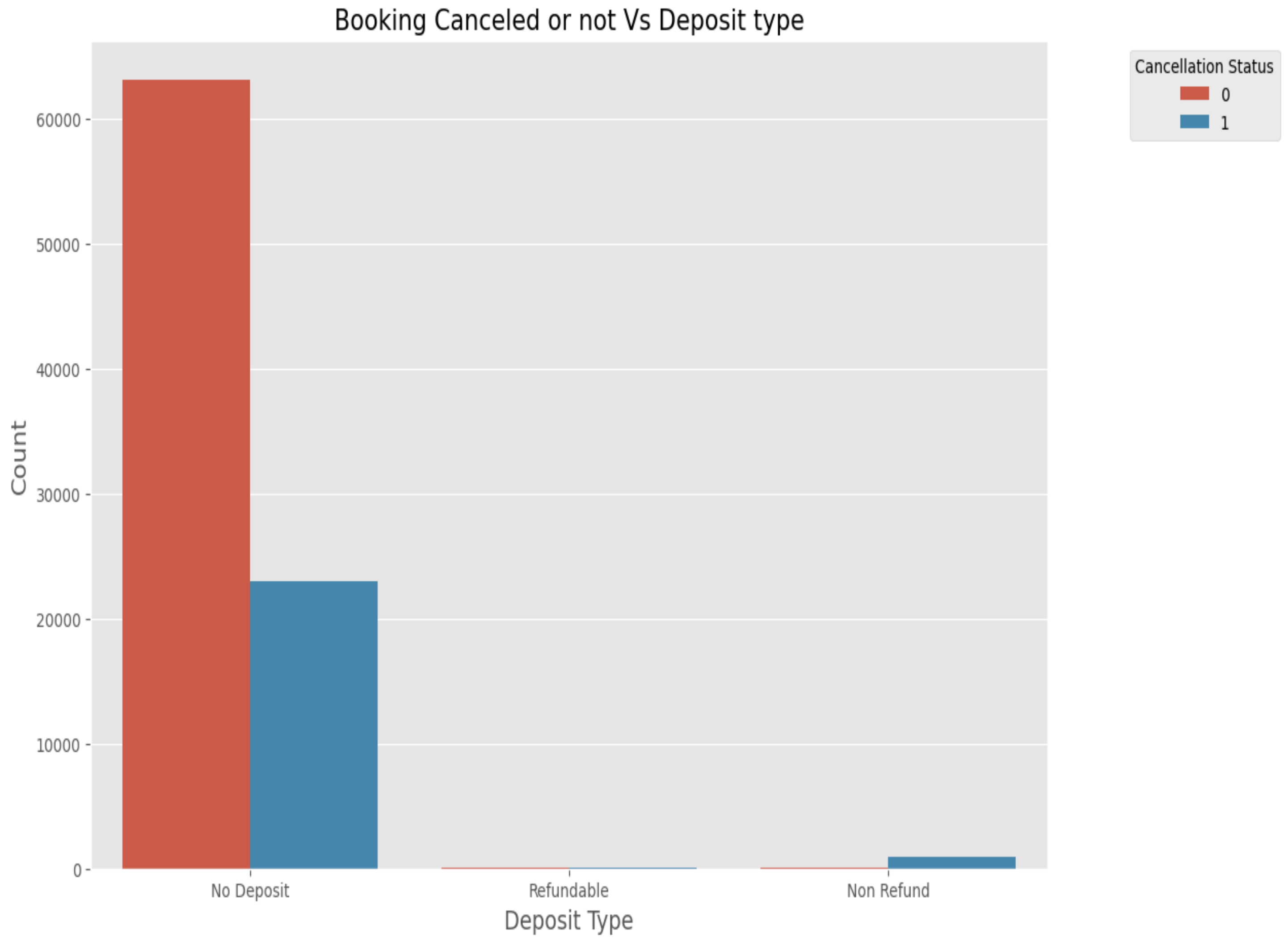
Distribution of Market Segments by Hotel Type



- I chose a clustered bar chart because it allows for the visual comparison of the distribution of market segments with respect to each hotel type. Each hotel type is represented as a separate cluster of bars, and within each cluster, bars are grouped by market_segment.

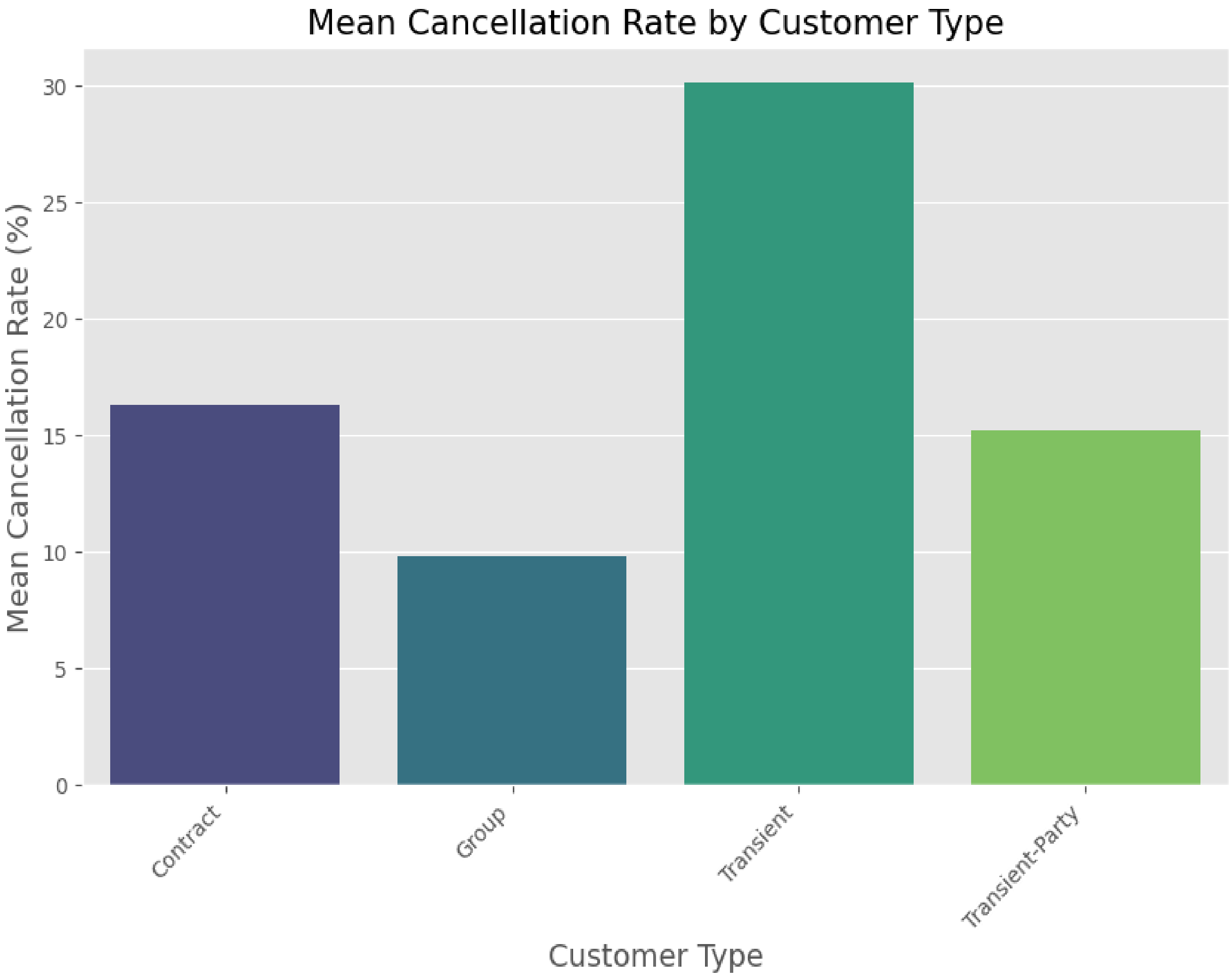
- The chart provides insights into how market segments are distributed across different hotel types. It allows us to compare the popularity of various market segments within each hotel category.

Booking Canceled or not based on Deposit type



- The chosen chart, a count plot comparing booking cancellations across different deposit types, is suitable for visualizing categorical data. It effectively shows the distribution of cancellations within each deposit type, making it easy to compare and analyze.
- We can infer from the plot above that about 25% of reservations were canceled by customers with No Deposit. It follows that visitors who make reservations without paying a deposit are probably going to cancel more of them. Surprisingly, non-refundable deposits experienced a higher rate of cancellation than refundable deposits. Given that hotel rates are typically higher for rooms with refundable deposits and that customers pay more in anticipation of cancellation, it would make sense to assume that refundable deposits have higher cancellation rates.

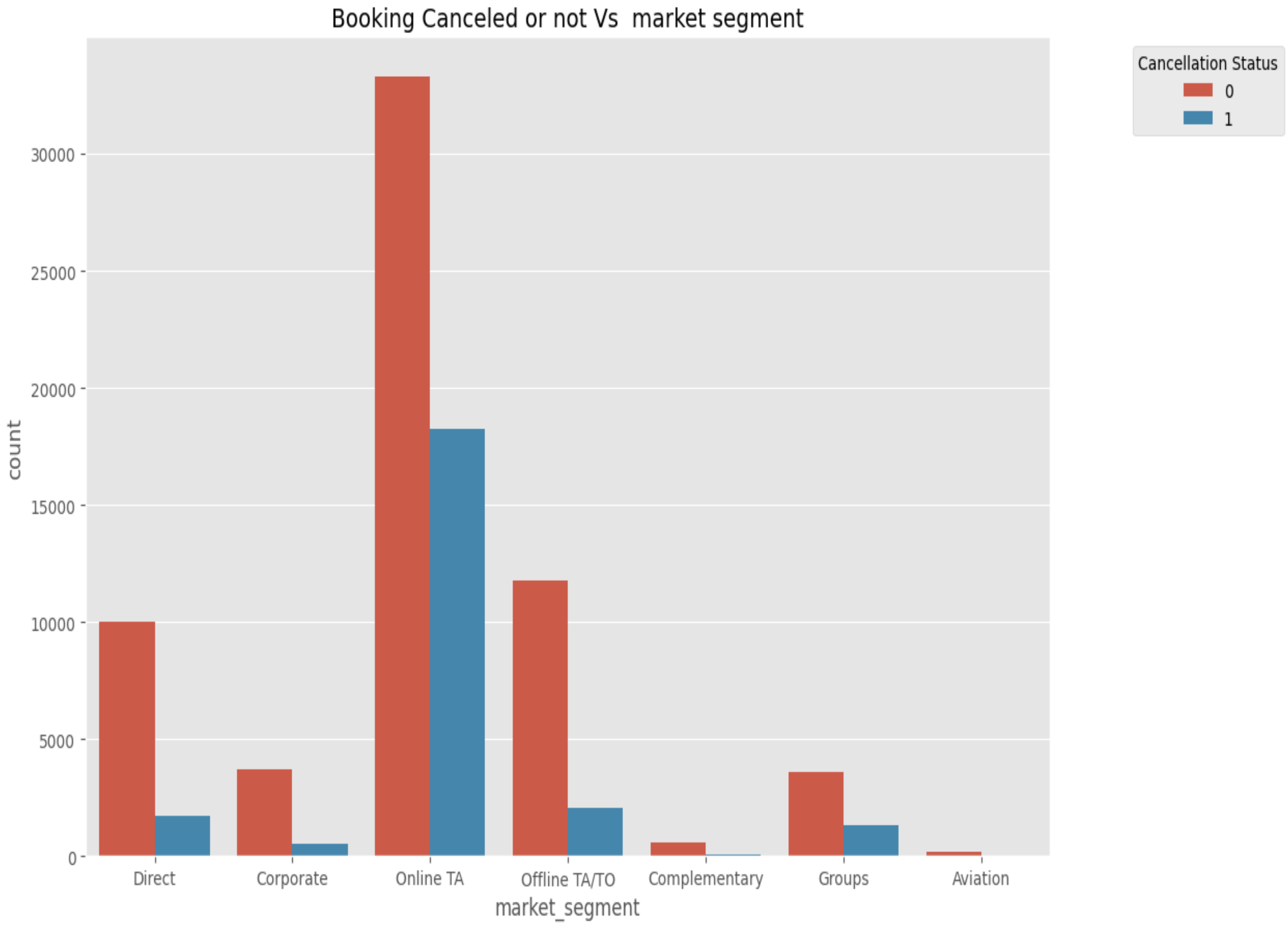
Cancellation Rate and Customer Type distribution



- The specific chart chosen is a bar chart. A bar chart is suitable for visualizing the mean cancellation rate for different customer types. It provides a clear comparison of cancellation rates among various customer types using horizontal bars.

- From the above plot we can say that **Transient** customer types have higher cancellations.

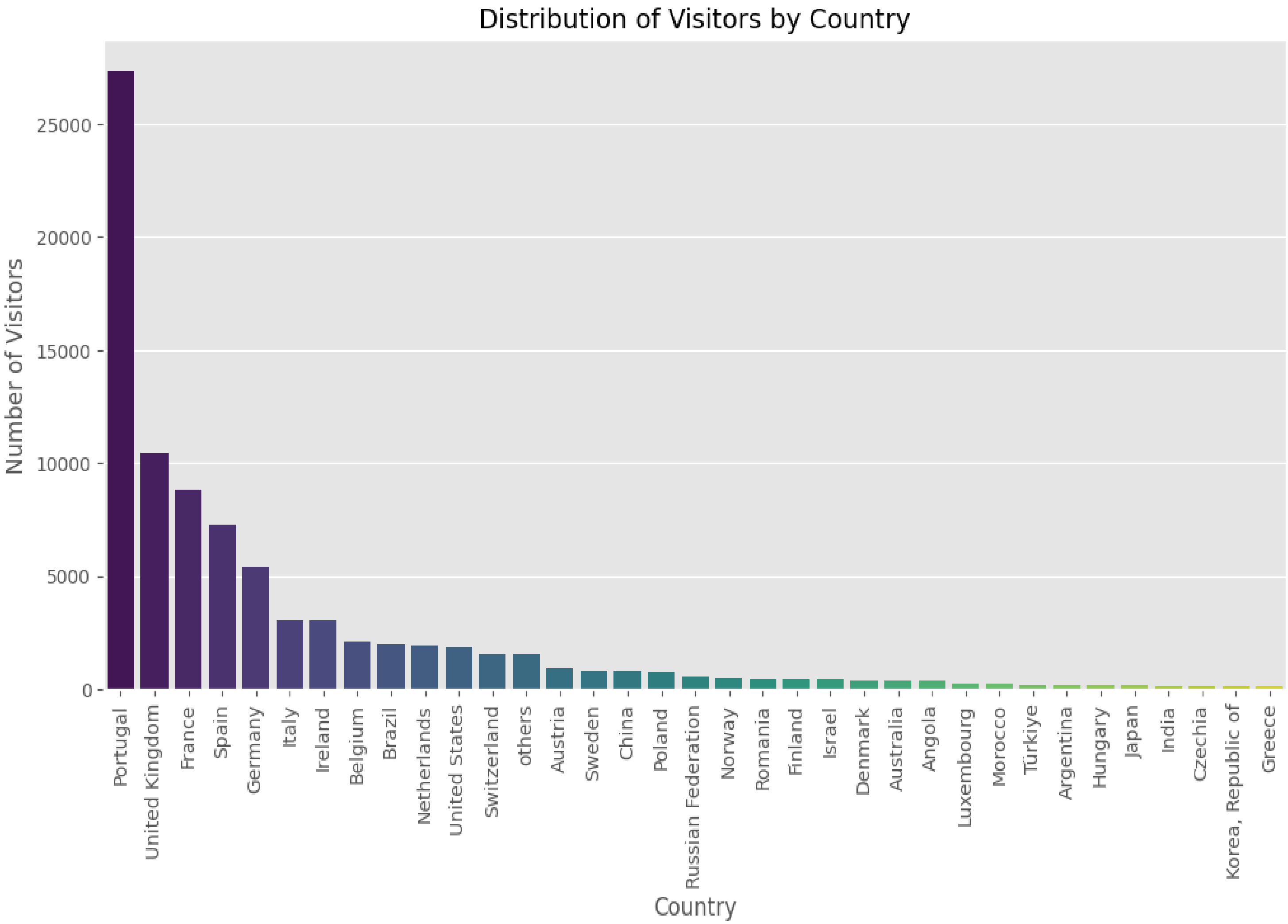
Market_Segment Vs Booking Cancelled or not



- Count plot has been chosen because this chart is appropriate for visualizing the count of occurrences of each category (market segment) and distinguishing cancellations (by using the hue parameter for 'is_canceled’).

- The group segment cancelation rate is approximately 50%. The cancellation rate for both online and offline travel agents and tour operators, or TA/TO, is more than 33%.The direct segment has a cancellation rate of less than 20%, which is noteworthy to observe.

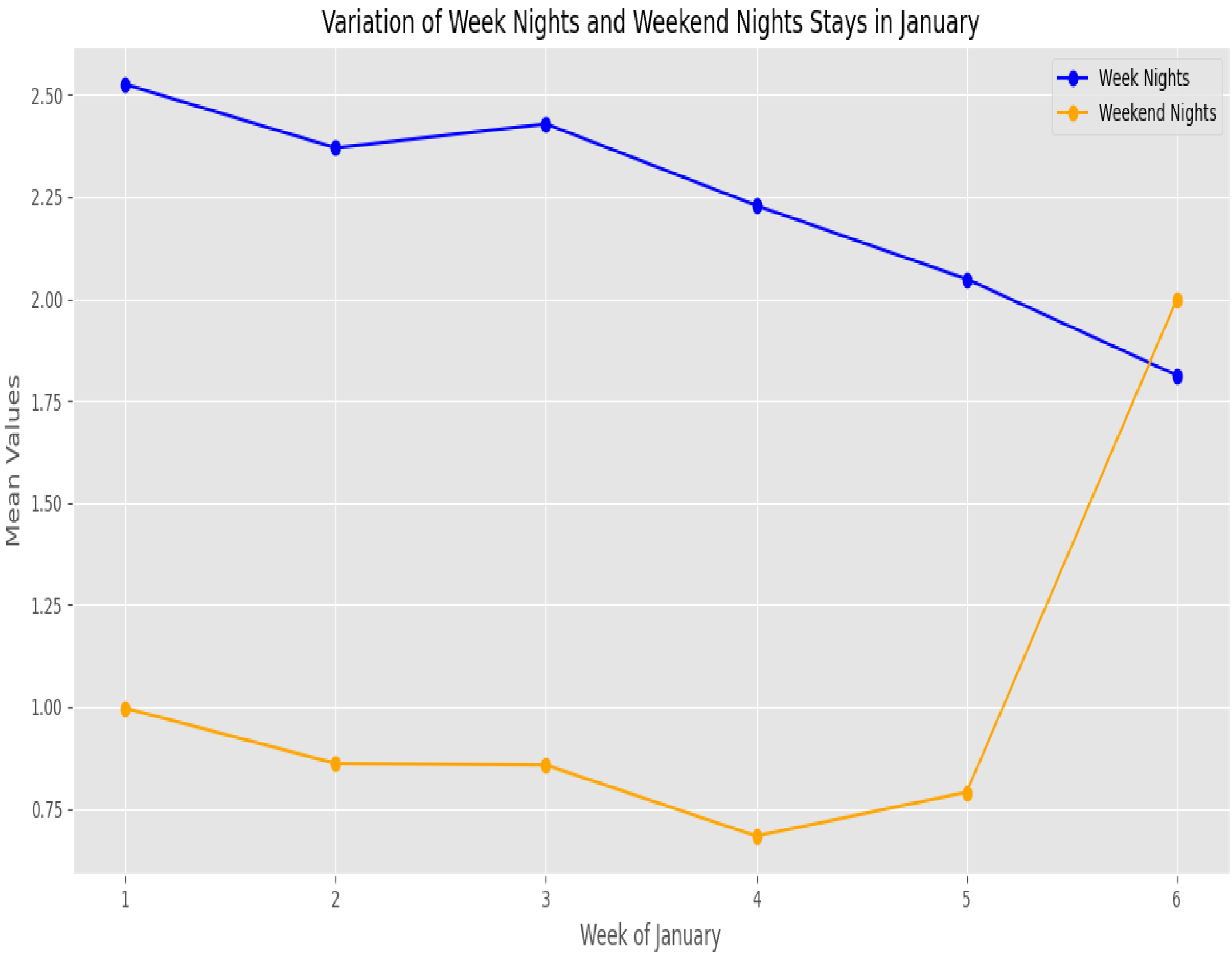
Distribution of Visitors by Country



- We have used bar chart because is an effective choice for visualizing the distribution of categorical data, such as the number of visitors from different countries.

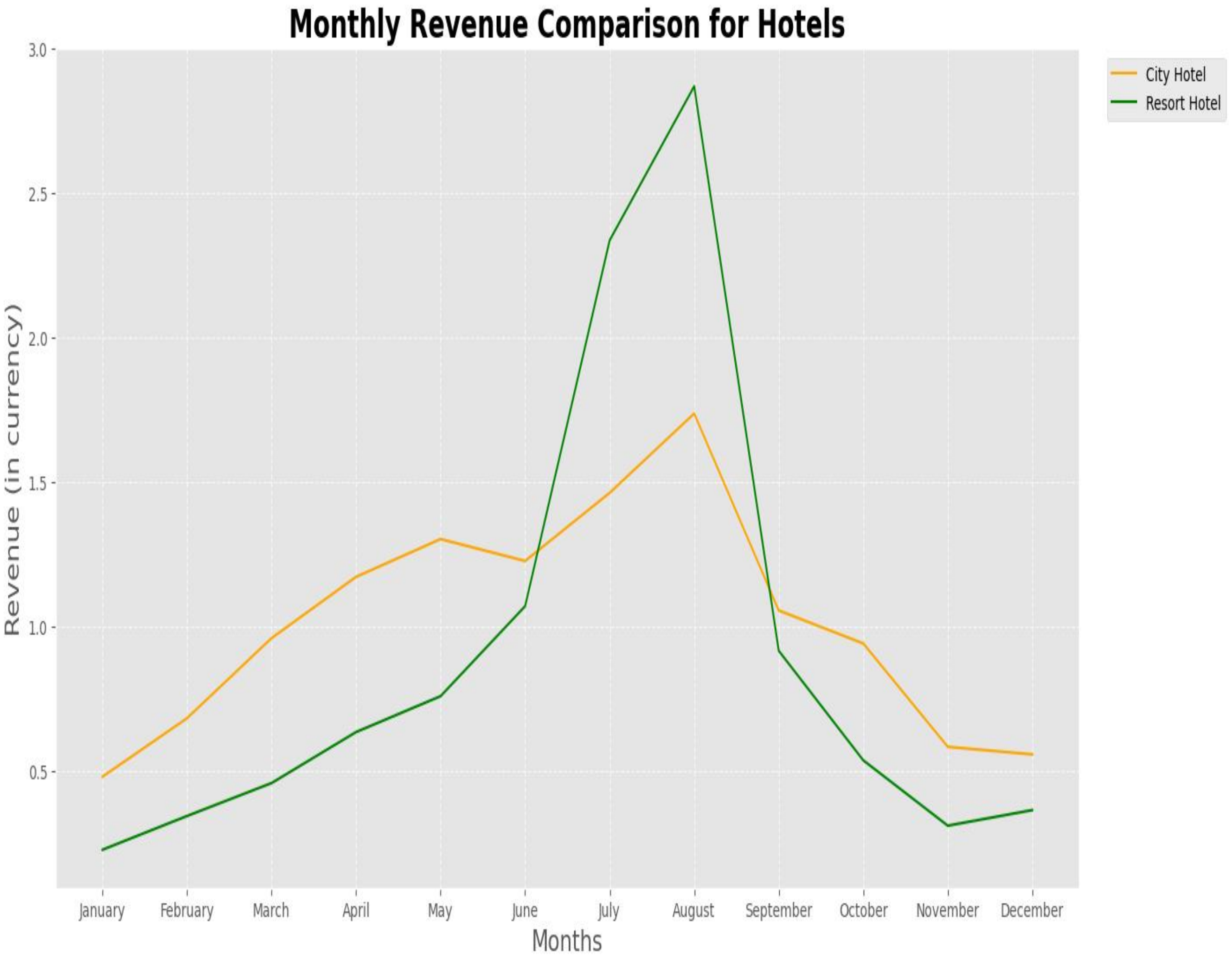
- The chart provides a visual representation of the distribution of visitors by country, showcasing the top 35 countries with the highest number of visitors.we can easily from the graph that highest no of visitors are coming from portugal.

Variation of Week Nights and Weekend Nights Stays with respect to each month



- A line plot is chosen because it is suitable for visualizing the variation of stays over time (in this case, weeks). It allows for the identification of trends, patterns, and fluctuations.
- The chart allows for a clear comparison between the mean values of stays in week nights and weekend nights throughout the weeks of the specified month.
- By observing the line plot, one can potentially identify seasonal trends or patterns in the stays throughout the month, providing insights into booking behavior.

Monthly Revenue in Comparison with Hotels

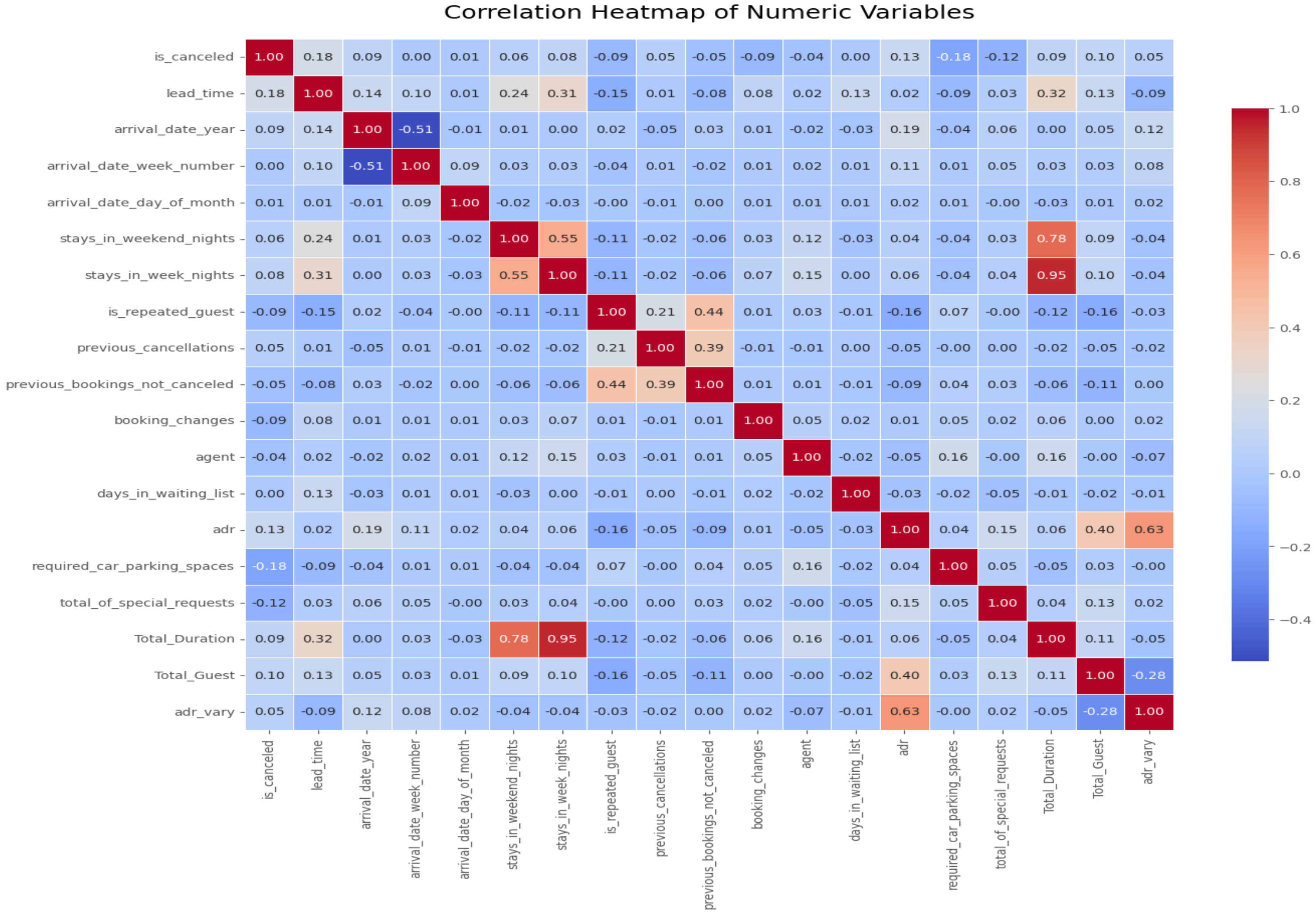


- We have chosen line plot here to clearly understand the distribution and extract the insights .

- From the graph we can say that Resort hotels and City hotels both are getting higher revenue between June to September.

Multi-Variate Analysis:-

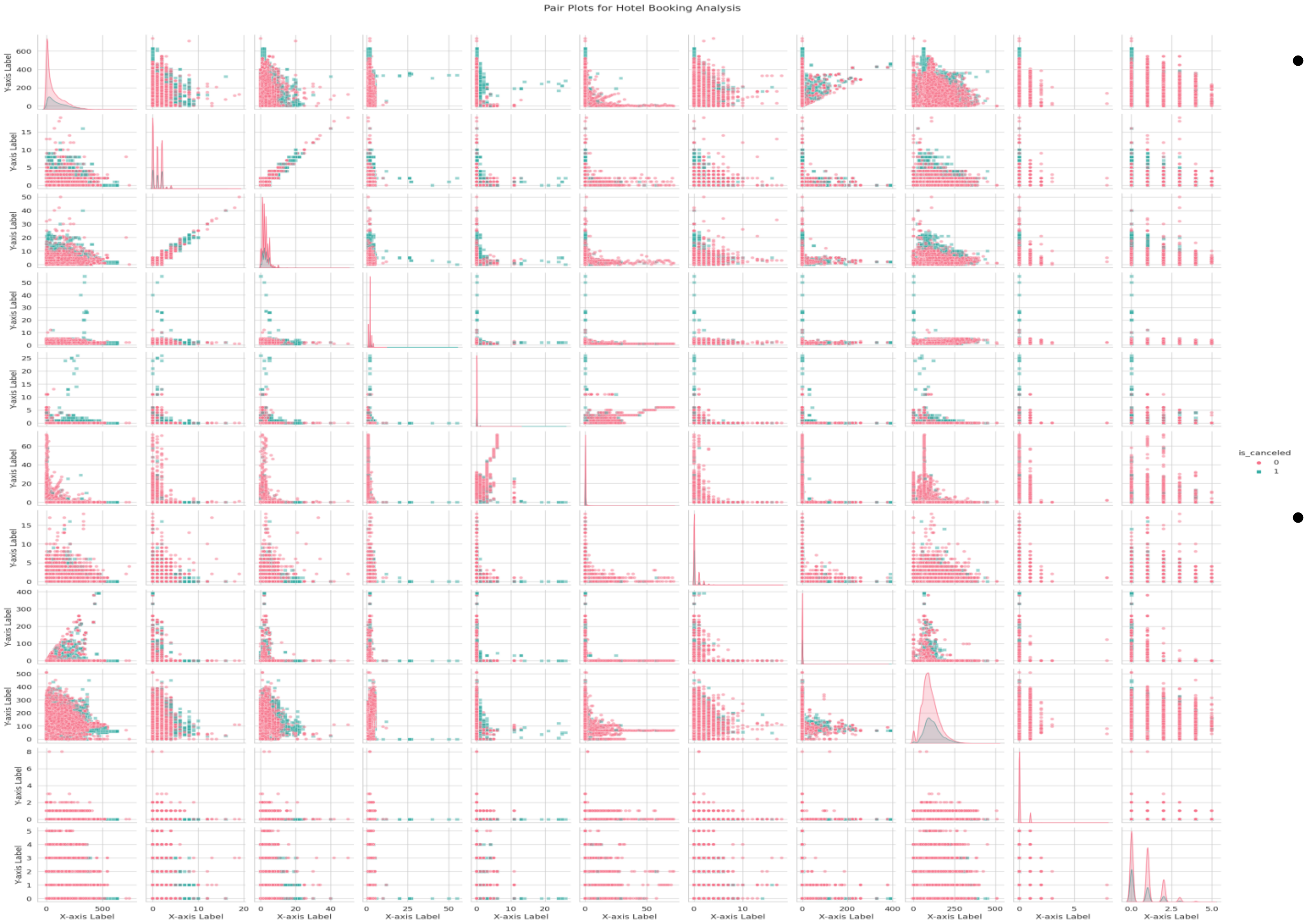
Correlation Heatmap of Numeric Variables


Correlation Heatmap of Numeric Variables

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	Total_Duration	Total_Guest	adr_vary
is_canceled	1.00	0.18	0.09	0.00	0.01	0.06	0.08	-0.09	0.05	-0.05	-0.09	-0.04	0.00	0.13	-0.18	-0.12	0.09	0.10	0.05
lead_time	0.18	1.00	0.14	0.10	0.01	0.24	0.31	-0.15	0.01	-0.08	0.08	0.02	0.13	0.02	-0.09	0.03	0.32	0.13	-0.09
arrival_date_year	0.09	0.14	1.00	-0.51	-0.01	0.01	0.00	0.02	-0.05	0.03	0.01	-0.02	-0.03	0.19	-0.04	0.06	0.00	0.05	0.12
arrival_date_week_number	0.00	0.10	-0.51	1.00	0.09	0.03	0.03	-0.04	0.01	-0.02	0.01	0.02	0.01	0.11	0.01	0.05	0.03	0.03	0.08
arrival_date_day_of_month	0.01	0.01	-0.01	0.09	1.00	-0.02	-0.03	-0.00	-0.01	0.00	0.01	0.01	0.01	0.02	0.01	-0.00	-0.03	0.01	0.02
stays_in_weekend_nights	0.06	0.24	0.01	0.03	-0.02	1.00	0.55	-0.11	-0.02	-0.06	0.03	0.12	-0.03	0.04	-0.04	0.03	0.78	0.09	-0.04
stays_in_week_nights	0.08	0.31	0.00	0.03	-0.03	0.55	1.00	-0.11	-0.02	-0.06	0.07	0.15	0.00	0.06	-0.04	0.04	0.95	0.10	-0.04
is_repeated_guest	-0.09	-0.15	0.02	-0.04	-0.00	-0.11	-0.11	1.00	0.21	0.44	0.01	0.03	-0.01	-0.16	0.07	-0.00	-0.12	-0.16	-0.03
previous_cancellations	0.05	0.01	-0.05	0.01	-0.01	-0.02	-0.02	0.21	1.00	0.39	-0.01	-0.01	0.00	-0.05	-0.00	0.00	-0.02	-0.05	-0.02
previous_bookings_not_canceled	-0.05	-0.08	0.03	-0.02	0.00	-0.06	-0.06	0.44	0.39	1.00	0.01	0.01	-0.01	-0.09	0.04	0.03	-0.06	-0.11	0.00
booking_changes	-0.09	0.08	0.01	0.01	0.01	0.03	0.07	0.01	-0.01	0.01	1.00	0.05	0.02	0.01	0.05	0.02	0.06	0.00	0.02
agent	-0.04	0.02	-0.02	0.02	0.01	0.12	0.15	0.03	-0.01	0.01	0.05	1.00	-0.02	-0.05	0.16	-0.00	0.16	-0.00	-0.07
days_in_waiting_list	0.00	0.13	-0.03	0.01	0.01	-0.03	0.00	-0.01	0.00	-0.01	0.02	-0.02	1.00	-0.03	-0.02	-0.05	-0.01	-0.02	-0.01
adr	0.13	0.02	0.19	0.11	0.02	0.04	0.06	-0.16	-0.05	-0.09	0.01	-0.05	-0.03	1.00	0.04	0.15	0.06	0.40	0.63
required_car_parking_spaces	-0.18	-0.09	-0.04	0.01	0.01	-0.04	-0.04	0.07	-0.00	0.04	0.05	0.16	-0.02	0.04	1.00	0.05	-0.05	0.03	-0.00
total_of_special_requests	-0.12	0.03	0.06	0.05	-0.00	0.03	0.04	-0.00	0.00	0.03	0.02	-0.00	-0.05	0.15	0.05	1.00	0.04	0.13	0.02
Total_Duration	0.09	0.32	0.00	0.03	-0.03	0.78	0.95	-0.12	-0.02	-0.06	0.06	0.16	-0.01	0.06	-0.05	0.04	1.00	0.11	-0.05
Total_Guest	0.10	0.13	0.05	0.03	0.01	0.09	0.10	-0.16	-0.05	-0.11	0.00	-0.00	-0.02	0.40	0.03	0.13	0.11	1.00	-0.28
adr_vary	0.05	-0.09	0.12	0.08	0.02	-0.04	-0.04	-0.03	-0.02	0.00	0.02	-0.07	-0.01	0.63	-0.00	0.02	-0.05	-0.28	1.00

- The correlation heatmap is chosen because it visually represents the strength and direction of relationships between numeric variables in your dataset. It's a valuable tool for understanding how different features correlate with each other, which can provide insights into potential patterns or dependencies.
- Analyzing the heatmap, we can identify pairs of variables that have strong correlations. For example, a positive correlation between 'stays_in_weekend_nights' and 'stays_in_week_nights' may suggest that guests who stay more on weekends also tend to stay more on weekdays. also Total_Duration and stays_in_week_nights and stays_in_weekend_night has strong correlation.

Pair Plot Pair Plots for Hotel Booking Analysis



- Pair plots are chosen because we are trying to visualize relationships between pairs of numerical variables. In the context of hotel booking analysis, it's a good choice to understand how different numerical features correlate with each other, and how these relationships may vary based on the cancellation status.
- Understand the relationship between the history of cancellations and the bookings that were not canceled. This may provide insights into customer behavior.

Observations:-

1. The most popular type of hotel among visitors is a city hotel.
2. Despite having nearly twice as many reservations as a resort hotel, the city hotel makes less money overall.
3. We can observe that reservations for city hotels peak in August and peak from April to July. Additionally, there are two peak months for resort hotels: June and September. July, August, and October see the highest booking volumes, therefore reservations are typically made 30 to 60 days in advance.
4. From June to September, revenue at both resort and city hotels increased. This is also due to the fact that, as the previous slide illustrates, ADR is high for both kinds of hotels at the same time. Therefore this period is best for hotels to generate more revenue.
5. Here, we can observe that the ADR falls with increasing lead time. This implies that a customer can obtain a better rate if he books a hotel in advance.
6. The majority of reservations are made by sporadic clientele.
7. The majority of reservations are made by transient clientele.
8. The majority of reservations and cancellations are handled by tour operators and online and offline travel agents.
9. Compared to resort hotels, there are more cancellations at city hotels.
10. When hotels don't collect deposits, there is a greater likelihood of cancellation. Therefore, hotels ought to require minimum deposits in order to lower the cancellation rate.
11.ADR falls as overall stay duration rises. This implies that a better offer for the clients might be arranged for a longer stay.

Conclusion:-

In conclusion, our thorough analysis of hotel booking trends uncovered key insights crucial for smart decisions in the hospitality industry. To curb cancellations, we recommend requiring minimum deposits, as data shows high cancellation rates when no deposits are taken.

Understanding customer types is essential; transient customers have higher cancellation rates. Encouraging group bookings with special incentives can cut cancellations and boost customer loyalty.

Peak booking months are from May to August, suggesting hotels can run special promotions during off-peak times to attract more bookings and increase revenue.

Geographically, Western European countries like Portugal, France, and the UK drive significant bookings and revenue. Targeted marketing in these regions can further boost customer acquisition.

In essence, our analysis provides a solid foundation for strategic decisions, offering practical insights to increase revenue, enhance customer satisfaction, and improve operations. Continuous adaptation to market changes will be key for sustained success. These insights serve as a roadmap for hotels to navigate industry trends and stay competitive.

thank you