Hotel Booking Data Analysis Report

By Bidisha Pal

Contents

Introduction and Objective :

- ## Introduction :

This project delves into the booking data of a city hotel and a resort hotel collected between 2015 and 2017 to identify the key factors contributing to high cancellation rates. The dataset, , stripped of any personally identifiable information, encompasses a wide range of comprehensive booking details, including "Reservation_date", "Length_of_stay", "Room_Type", "Stay_on_weekday_and_weekend", "ADR", "Guest Demographics (Adults, Children, Babies)" and other amenities like "Parking_Availability" etc.

- ## Objective :

In recent years, both the City and Resort hotels have experienced a concerning surge in cancellation rate, negatively impacting revenue generation and leading to underutilized room capacity. By meticulously analysing cancellation patterns and associated variables, this project aims to provide insights and recommendations to reduce cancellation, boost revenue and optimize hotel operations, improving the financial performance.

Data Description :

- The given dataset contains 32 columns and 119390 rows.
- The dataset contains both numerical and categorical columns.
- Few columns "Country", "Agent", "Children" and "Company" have missing values.
- Rest of the columns do not have any missing values.
- "Reservation_Status_Date" column in is 'Object' datatype which is needed to convert in 'datetime' datatype.
- Similarly, "Children", "Agent" and "Company" columns are in 'float' datatype and is required to convert in 'integer' datatype.

```
booking_data = pd.read_csv('hotel_bookings.csv')
booking_data.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_i
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

Attributes of Dataset :

The attributes of the dataset are as follows :

1. **hotel** : The dataset contains the booking information of two hotel, City Hotel & Resort Hotel
2. **is_canceled** : Value indicating if the booking was cancelled (1) or not (0).
3. **lead_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. **arrival_date_year :** Year of arrival date
5. **arrival_date_month :** Month of arrival date with 12 categories: “January” to “December”
6. **arrival_date_week_number :** Week number of the arrival date
7. **arrival_date_day_of_month :** Day of the month of the arrival date
8. **stays_in_weekend_nights :** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. **stays_in_week_nights :** Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel BO and BL/Calculated by counting the number of week nights.

10. **adults :** Number of adults
11. **children :** Number of children
12. **babies** : Number of babies
13. **meal :** Preferred Meal Type amongst 'BB' 'FB' 'HB' 'SC'
14. **country :** Country of origin.
15. **market_segment :** Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
16. **distribution_channel :** Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
17. **is_repeated_guest :** Value indicating if the booking name was from a repeated guest (1) or not (0)
18. **previous_calcellation :** Number of previous bookings that were cancelled by the customer prior to the current booking
19. **previous_bookings_not_canceled :** Number of previous bookings not cancelled by the customer prior to the current booking
20. **reserved_room_type :** Code of room type reserved. Code is presented instead of designation for anonymity reasons
21. **assigned_room_type :** Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
22. **booking_changes :** Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
23. **deposit_type :** No Deposit – no deposit was made; non-refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
24. **agent :** ID of the travel agency that made the booking
25. **company :** ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
26. **days_in_waiting_list :** Number of days the booking was in the waiting list before it was confirmed to the customer
27. **customer_type :** Group , Transient , Transient-party , Contract
28. **adr :** Average Daily Rate (Calculated by dividing the sum of all lodging transactions by the total number of staying nights)
29. **required_car_parking_spaces :** Number of car parking spaces required by the customer
30. **total_of_special_requests :** Number of special requests made by the customer
31. **reservation_status :** Check-Out, Check-In, No Show
32. **reservation_status_date :** Date at which the last status was set.

Data Cleaning :

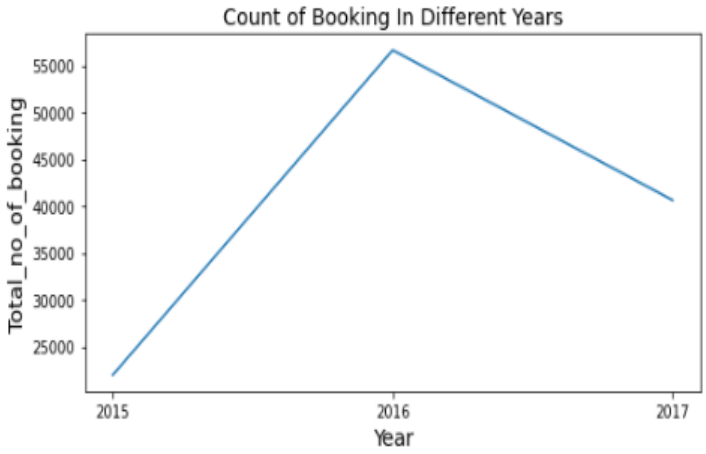
1. **Imputing Missing Values :**
   - **a.** There were missing values in 4 columns - ‘children’, ’country’, ’agent’ and ‘company’ .
   - **b.** ‘company’ column has been dropped due to high missing values (~94%).
   - **c.** 13% missing values in ‘country’ column has been imputed with ‘others’
   - **d.** Missing values in ‘children’ column has been filled with ‘0.0’ following mode approach.
   - **e.** Missing values in ‘agent’ column has been filled with ‘0.0’ creating a new category.
2. **Converting Data Types :**
   - **a.** Datatype of ‘reservation_status_date’ has been changed to ‘datetime’ from ‘object’ type.
   - **b.** ‘children’ and ‘agent’ columns are converted to ‘int’ datatype from ‘float’
3. **Creating New Features from the Existing Features :**
   - **a.** A new feature “total_guests” has been created by adding up the values of “adults", “children“ , “babies” and all the observations with ‘total_guests” = 0 has been dropped.

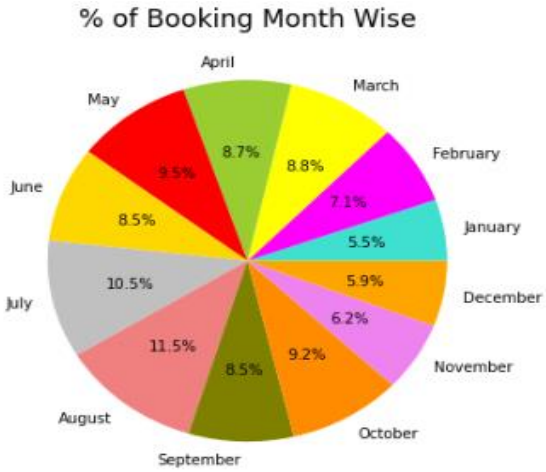
Exploratory Data Analysis :

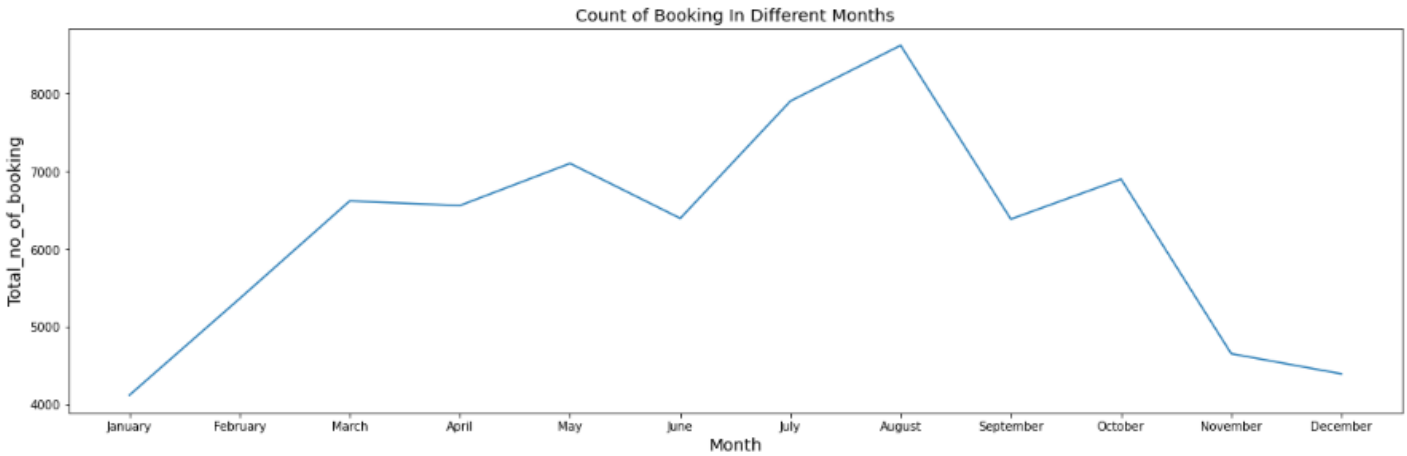
All the findings from EDA are as follows :

- Out of all the bookings 37% were cancelled.
- Most of the guests are from PTR followed by GBR and FRA.
- No. of booking is maximum when there is no children, babies and number of adults are 2.
- August was the busiest month of the year as no. of booking was highest and ADR is higher in August compared to other months of the year.
- The year 2016 received the maximum no of bookings(~64%).
- The cancellation rate was max in the year 2017 approximately 39%, yet ADR in 2017 is higher.
- The guests prefer City Hotels over Resort Hotels and this trend has maintained across years resulting a higher ADR for City Hotels in comparison to Resort Hotels.
- % of cancellation for City Hotels are higher ~41% and the longer waiting time can be a potential cause for this.
- Out of all the guests only 3.15% were repeated guests which indicates a low retention rate.
- Most of the bookings (~80%) has been done through Travel Agents/Travel Operators.
- Guests who have a record of cancelling more than 13 previous bookings has a very high probability of cancelling the current booking.

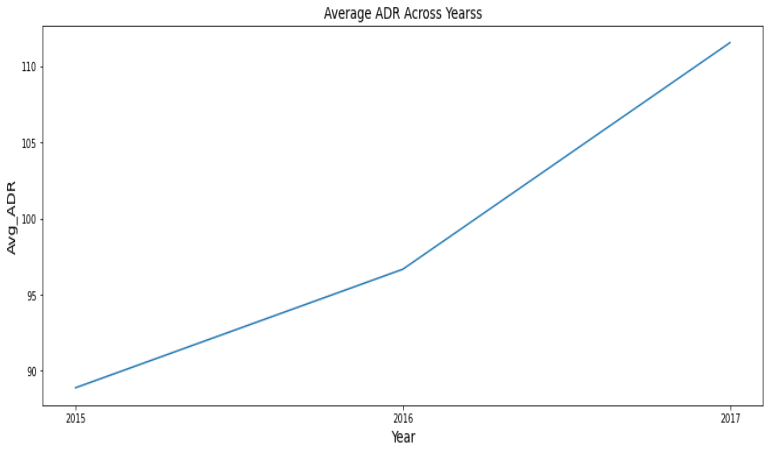
Booking Trend & Seasonality in booking :

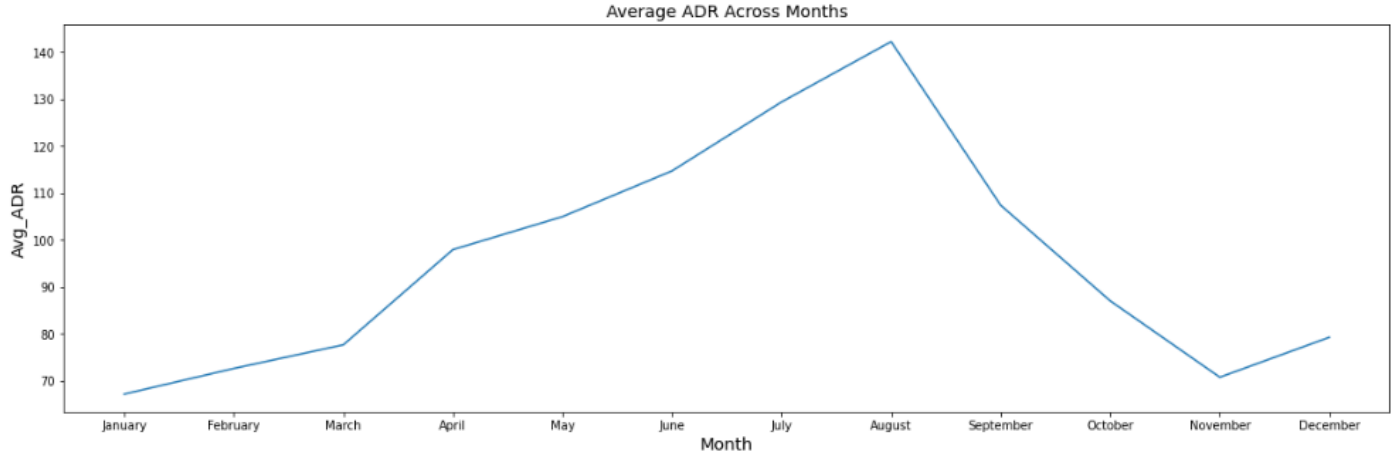




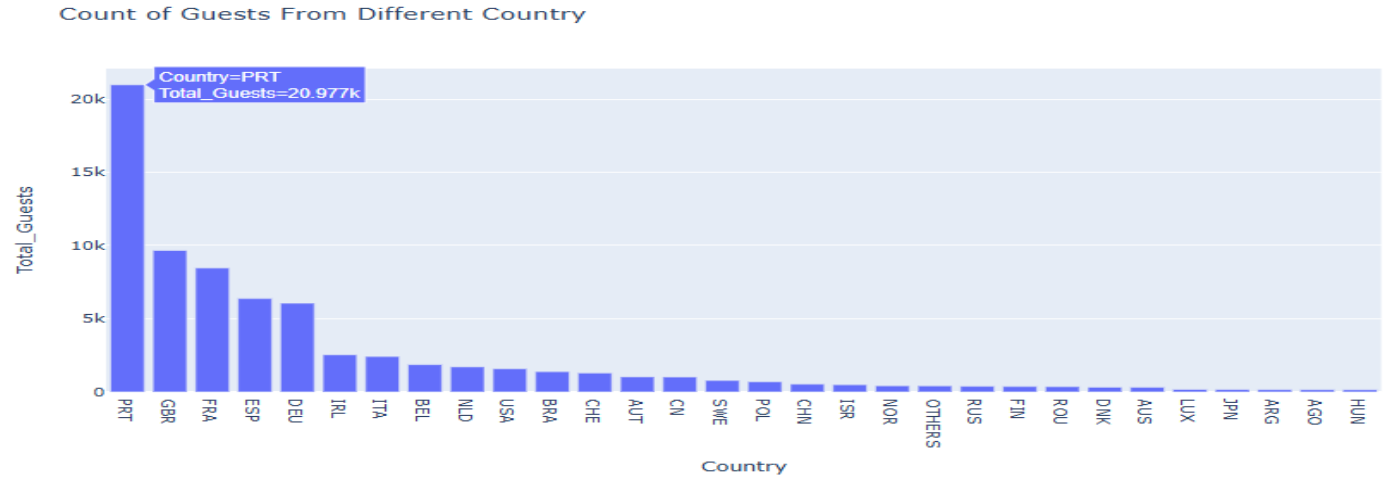


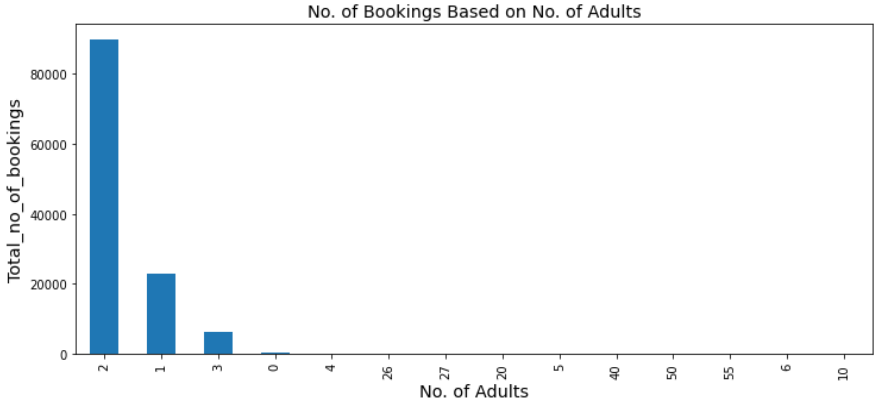
Average Daily Rate (ADR) Trend :

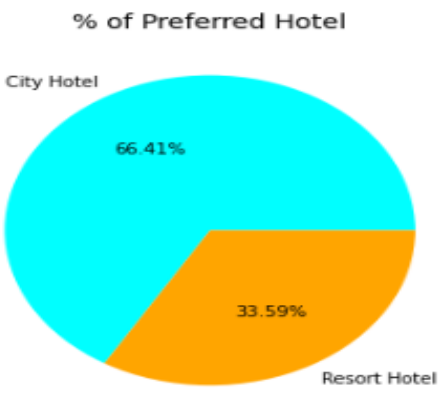




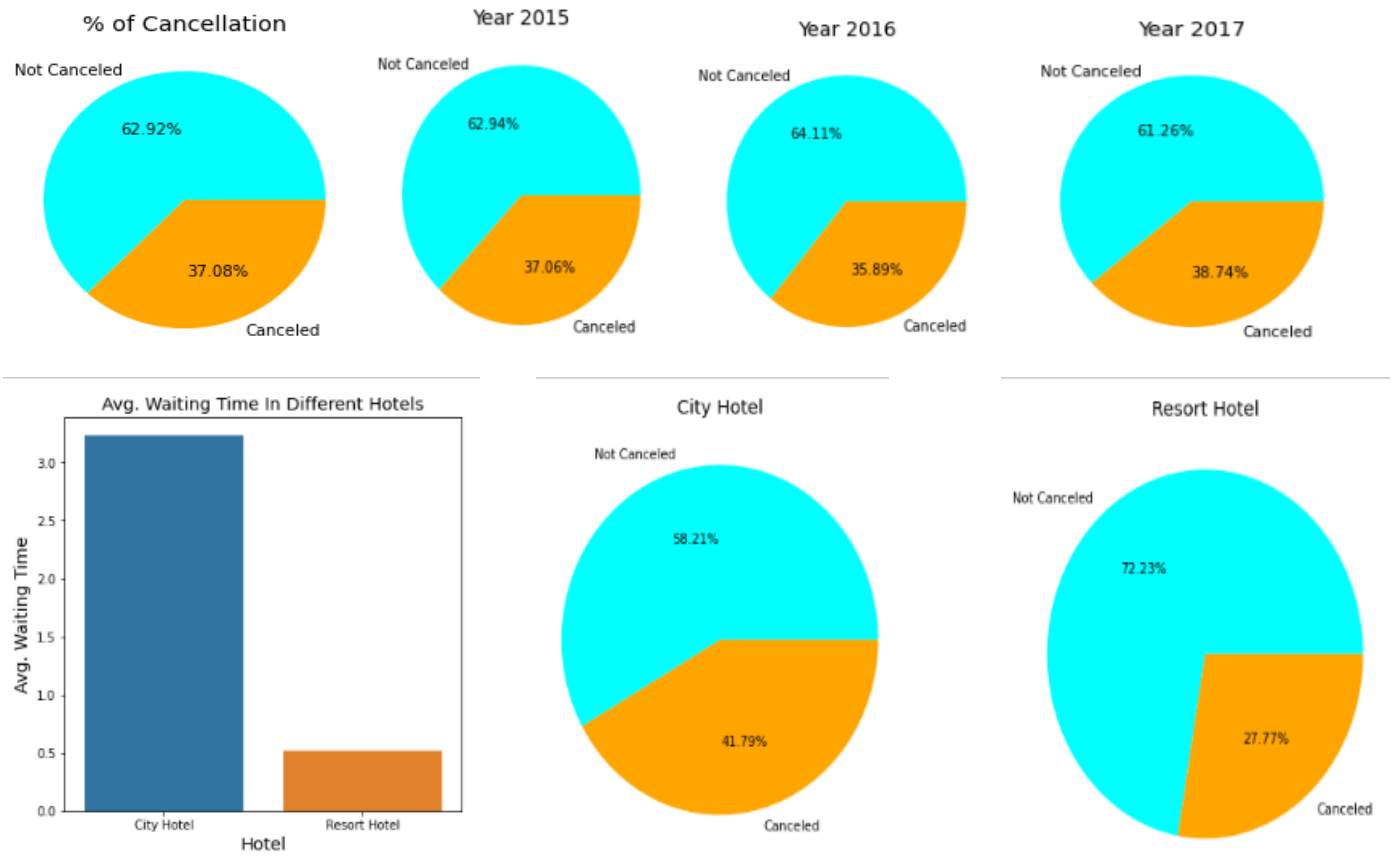
Customer Demographic :







Cancellation Rate, Pattern, and Influencing Factors :



Hypothesis Testing :

Here I have performed hypothesis testing on two statements.

1. "Customers booking more than 6 months in advance are more likely to cancel."
2. "Weekday bookings have a higher average daily rate than weekend bookings."

Statement_1 : "Customers booking more than 6 months in advance are more likely to cancel."

- H0: "Customers booking more than 6 months in advance are not likely to cancel booking."
- H1: "Customers booking more than 6 months in advance are more likely to cancel booking."

As both the variable are **Categorical** "Chi-Square Contingency" test has been used to test significant statistical relation between both the variables.

Conclusion of test:

- P-value = 0 indicates that two features “advanced_booking_category” and “is_canceled” are highly statistically significant and we can reject the null hypothesis.
- The conclusion of the above hypothesis test is that **"Customers booking more than 6 months in advance are more likely to cancel the booking."**

Statement_2: "Weekday bookings have a higher average daily rate than weekend bookings."

- H0: "Weekday bookings and Weekend Bookings have same average daily rate"
- H1: "Weekday bookings have a higher average daily rate than weekend bookings."

Here we need to compare mean of 'ADR' between two features "Weekday_booking" and "Weekend_booking".

In this case we can apply **Two Sample t-Test** or **Independent t-test** if both the features are normally distributed or we can apply **Mann-Whitney U Test** of any one of the features are not normally distributed.

To test Normal Distribution of Features **"Shapiro-Wilk"** test has been applied.

- As p-value = 0.0 we can say the value of 'ADR' is normally distributed for Weekend Data as well.
- As the Variable 'ADR' is normally distributed for both Weekday and Weekend data we will apply **Two Sample t-Test**

Conclusion of test :

- p-value is much lower than 0.05 (Value of Significance).
- So, we can reject the null hypothesis and the conclusion from the above hypothesis test is **"Weekday Booking have a higher average daily rate than Weekend Bookings."**

Predictive Modelling :

- **Data Preparation:**
- **Step 1 :** Encoding Categorical Variables
  - Label Encoding has been applied for Binary Categorical Variables.
  - One-Hot Encoding has been applied for Multi-class Categorical Variables.
- **Step 2 :** Splitting Data into Dependent and Independent Variables
- **Step 3 :** Handling Multicollinearity
  - All the columns with correlation coefficient > 0.8 are dropped.

```
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))
high_corr = [col for col in upper.columns if any(upper[col] >= 0.8)]
len(high_corr)
```

```
6
```

- **Step 4 :** Feature Scaling :
  - Values of a lot of the numerical columns vary over a very wide range. Hence, to avoid unwanted biasness in result we have scaled/normalize the independent variables with MinMaxScaler

- **Model Building and Evaluation:**

Different classification models like Logistic Regression, Decision Tree, Random Forest, XGBoost, AdaBoost etc. have been trained using the dataset and their performance is given in the following table,

	Model Name	Accuracy	ROC_AUC Score	F1 Score	Precision	Recall
0	Logistioc Regression	1.0	1.0	1.0	1.0	1.0
1	Decision Tree	1.0	1.0	1.0	1.0	1.0
2	Random Forest Classifier	1.0	1.0	1.0	1.0	1.0
3	Bagging Classifier	1.0	1.0	1.0	1.0	1.0
4	AdaBoost Classifier	1.0	1.0	1.0	1.0	1.0
5	Gradient Boosting	1.0	1.0	1.0	1.0	1.0
6	XGBoost Classifier	1.0	1.0	1.0	1.0	1.0

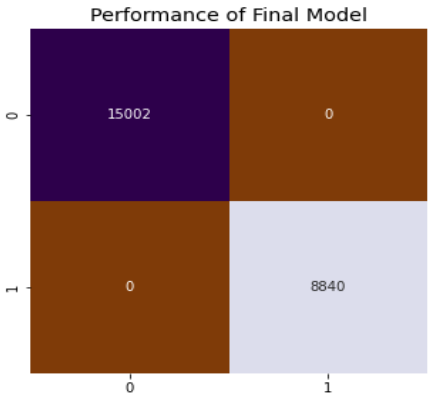
- **Model Selection :**
  - Performance of all the models is same using both Stratified k Fold and Simple Train Test Split method.
  - We have checked for Over Fitting problem using both Cross Validation and Train Test Split approach and in both the cases accuracy is consistent.

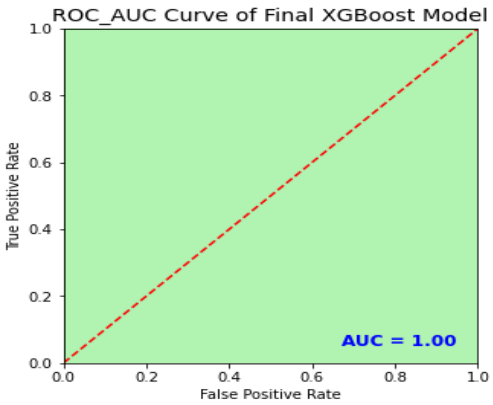
- So Logistic Regression is chosen as the final model as the model provide same performance as others with lesser complexity, computational cost, and high interpretability.

Performance of The Finally Selected Model:

Performance of Model is as follows :

1. Accuracy : 100%
2. ROC_AUC Score : 100%
3. Precision : 100%
4. Recall : 100%
5. F1 Score : 100%





Operational Insights :

1. Targeted Marketing : Segment customers and tailor marketing campaign with incentives, loyalty programs and family packages.

2. Focus on Direct Booking : Prioritize direct bookings and run seasonal campaigns to reduce dependencies on high-cancellation channels.

3. Operatonal Improvements : Balance flexible cancellation policies with protective measures and proactively engage with long-load-time bookings.

4. Dynamic Pricing : Implement a flexible pricing strategy based on demand, lead_time and guest loyalty.

Conclusion :

We have successfully trained our model to predict cancellation of a booking with an accuracy 100%. We proceeded step by step analysing, cleaning, and modelling the data. We have performed extensive Exploratory Data Analysis to find out the potential cause of increasing cancellation rate. Simultaneously we have applied various machine learning algorithm to achieve the desired result and finally we are able to build a model with a quiet good accuracy, precision, and recall.