

# Capstone Project

## EDA

### Hotel Bookings Analysis

SUBMITTED BY -

Team Member 1 – Mr. Kumar Ankit

Team Member 2 – Ms. Charmi Patel

Team Member 3 – Mr. Aman Verma





# OVERVIEW OF THE PROJECT

---

## INTRODUCTION

---

## WORKFLOW

---

## CONCLUSION

# INTRODUCTION

- In today's scenario hospitality is one of the emerging business. The travel and tourism contributes around 5.81 trillion U.S. dollars to the global economy
- Today most of hotel owners would like to run their business at its full capacity helping them to bring in sizeable revenue.
- We are here with a compact data study about the hotel industry, mainly the booking . This dataset contains information two types of hotels and its other related parameters.
- Keeping this into mind the purpose of the study is to find out meaningful insight from the data which can help them to understand their business better way , which in turn help them in providing better services.
  
- By the end we will conclude the study by providing following insights:-
  - ✓ Distribution Segment and market segments to be focused to increase revenue
  - ✓ Factors like preferred meals, room type, extra services for improving efficiency etc. and thereby help in increasing its ADR and revenue.



## WORKFLOW OF THE PROJECT

### PROBLEM STATEMENT AND COLLECTION OF RAW DATA

Finding the problem

### DATA CLEANING AND MANIPULATING

Removing all null values, outliers and out of format data

### EXPLORATORY DATA ANALYSIS (EDA)

Using python technical skills to extract results from data

### DATA VISUALISATION

Displaying output results in the form of graphs

### ANALYSIS AND INTERPRETATION

Our findings and interpretation from the EDA and visualization

# WORKFLOW

- For the purpose of analysis we divided the workflow in following five steps:

## PROBLEM STATEMENT AND COLLECTION OF RAW DATA

- Finding the problem

## DATA CLEANING AND MANIPULATING

- Removing all null values, outliers and out of format data

## EXPLORATORY DATA ANALYSIS (EDA)

- Using python technical skills to extract results from data

## DATA VISUALISATION

- Displaying output results in the form of graphs

## ANALYSIS AND INTERPRETATION

- Our findings and interpretation from the EDA and visualization

# PROBLEM STATEMENT

- In today's scenario hospitality is one of the emerging business. The travel and tourism contributes around 5.81 trillion U.S. dollars to the global economy
- Today most of hotel owners would like to run their business at its full capacity helping them to bring in sizeable revenue.
- We are here with a compact data study about the hotel industry, mainly the booking . This dataset contains information two types of hotels and its other related parameters.
- Analysing data on this hotels will help them understand the factors which are favorable and unfavorable in running their business.
- Keeping this into mind the purpose of the study is to find out meaningful insight from the data which can help them to understand their business better way , which in turn help them in providing better services.

# DATA COLLECTION AND UNDERSTANDING

- ❑ This data contains **119390** rows and **32** columns.
  
- ❑ Attributes of the data is as following:
  - **Hotel** : There were two category of hotels- Resort Hotel and City Hotel
  - **is\_canceled** : If the booking was canceled (1) or not (0)
  - **lead\_time** : Number of days before the actual arrival of the guest
  - **arrival\_date\_year** : Year of arrival date
  - **arrival\_date\_month** : Month of arrival date
  - **arrival\_date\_week\_number** : Week number of year for arrival date
  - **arrival\_date\_day\_of\_month** : Day of arrival date
  - **stays\_in\_weekend\_nights** : Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests
  - **stays\_in\_week\_nights** : Number of week nights (Monday to Friday) spent at the hotel by the guests
  - **adults** : Number of adults

- **children** : Number of children among guests
- **babies** : Number of babies among guests
- **meal** : Type of meal booked
- **country** : Country of origin i.e. Country of guests
- **market\_segment** : Designation of the market segment into - “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **distribution\_channel** : Name of the booking distribution channel into - “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **is\_repeated\_guest** : If the booking name was from a repeated guest (1) or not (0)
- **previous\_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
- **previous\_bookings\_not\_canceled** : Number of previous bookings not cancelled by the customer prior to the current booking
- **reserved\_room\_type** : Code of room type reserved
- **assigned\_room\_type** : Code for the type of room assigned



- **booking\_changes** : Number of changes/amendments made to the booking
- **deposit\_type** : Type of the deposit made by the guest
- **agent** : ID of the travel agency who made the booking
- **company** : ID of the company/entity that made the booking
- **days\_in\_waiting\_list** : Number of days the booking was in the waiting list before it was confirmed to the customer
- **customer\_type** : Type of booking, assuming one of four categories
- **adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- **required\_car\_parking\_spaces** : Number of car parking spaces required by the customer
- **total\_of\_special\_requests** : Number of special requests made by the customer
- **reservation\_status** : Reservation last status, assuming one of three categories (Canceled , Check-Out , No-Show)
- **reservation\_status\_date** : Date at which the last reservation was updated. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel

# DATA CLEANING AND MANIPULATING

## Data Cleaning:

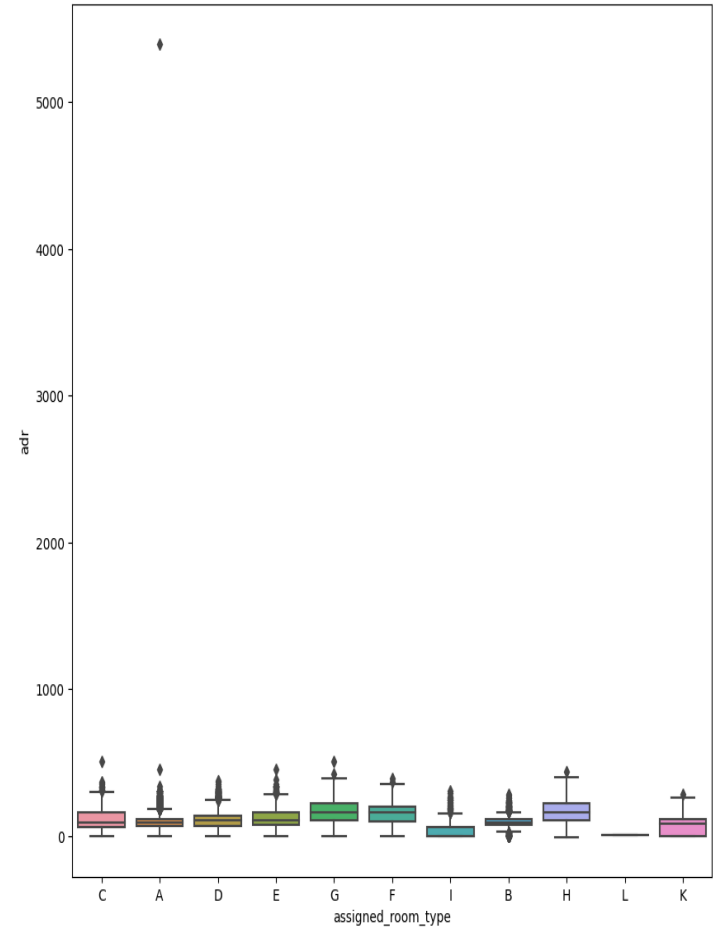
- This given data has duplicate values which are dropped using **drop\_duplicates()**
- We dropped column with maximum null values - company (112593)
- In agent column, we have 16340 missing values, so we are replacing those values with 'unknown'. Also we are typecasting the ID of Agent to integer values.
- In Children column we have 4 missing values, so we are filling those missing values with average value of children column and we type-casted float values to integer value.
- Adults, babies, and children can't be zero at the same time, so we are **dropping the rows having sum of all these is zero.**

## Data Manipulation:

- For effective study we have manipulated data by **combining few columns-**
  - $\text{total\_people} = \text{adults} + \text{babies} + \text{children}$
  - $\text{total\_stay} = \text{stays\_in\_weekend\_nights} + \text{stays\_in\_week\_nights}$
  - $\text{Revenue} = \text{total\_stays} * \text{ADR}$

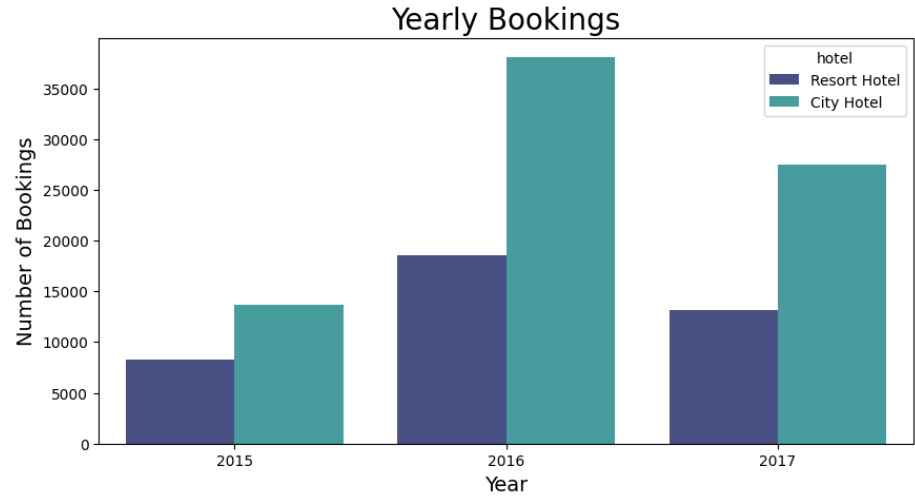
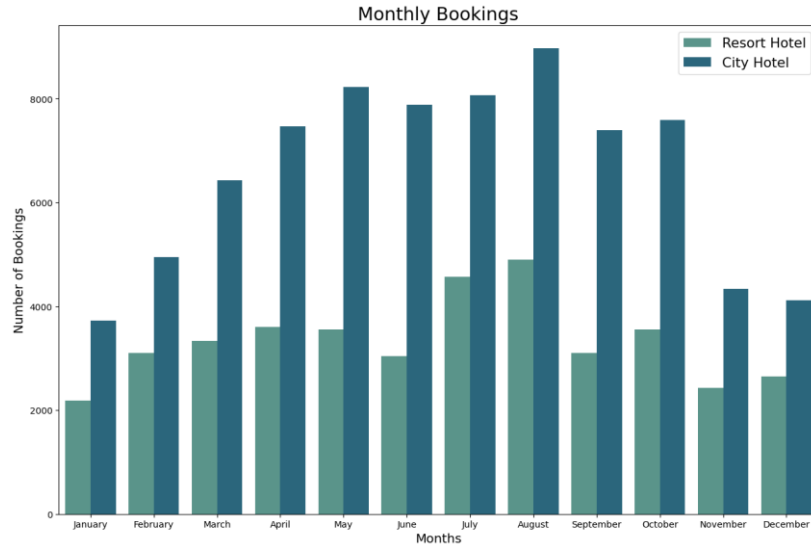
# DATA CLEANING AND MANIPULATING

- Following graph shows outliers in average daily rate (ADR) column. Maximum values were in the range of 1000.
- Hence we are removing outlier data from dataset where ADR is greater than 1000.

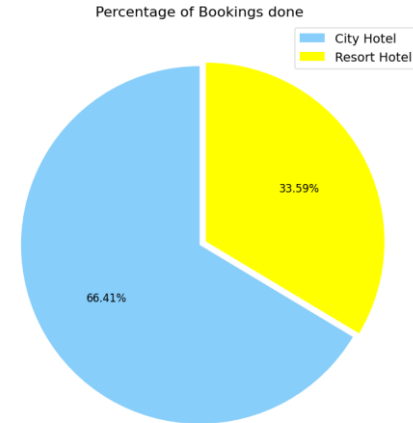


## **EXPLORATORY DATA ANALYSIS (EDA)**

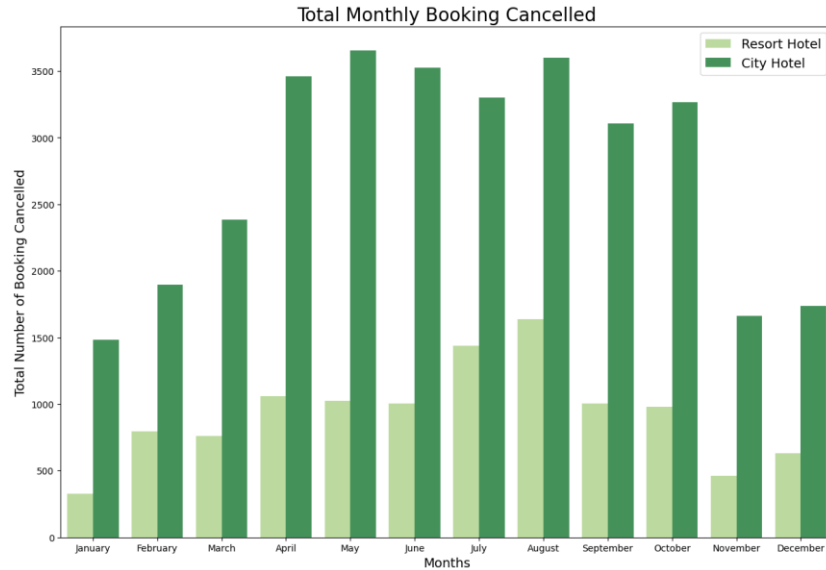
## PERCENTAGE CONTRIBUTION OF DIFFERENT HOTEL TYPES



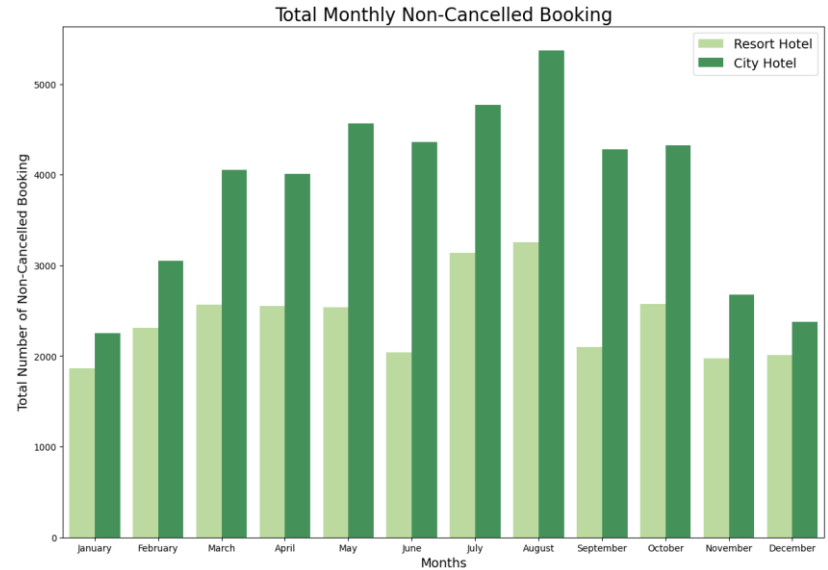
- From the above Pie-Chart, it is observed that 66.41 % of booking were done in City Hotel and 33.59 % of booking were done in Resort Hotel.
- The above visualization states that, in year 2016 most bookings have been made in City Hotel and Resort Hotel.
- Month wise booking indicates that most of the bookings done in the Month of August in Resort Hotel and City Hotel.



# TOTAL CANCELLED AND NON-CANCELLED BOOKING OF DIFFERENT HOTEL TYPES

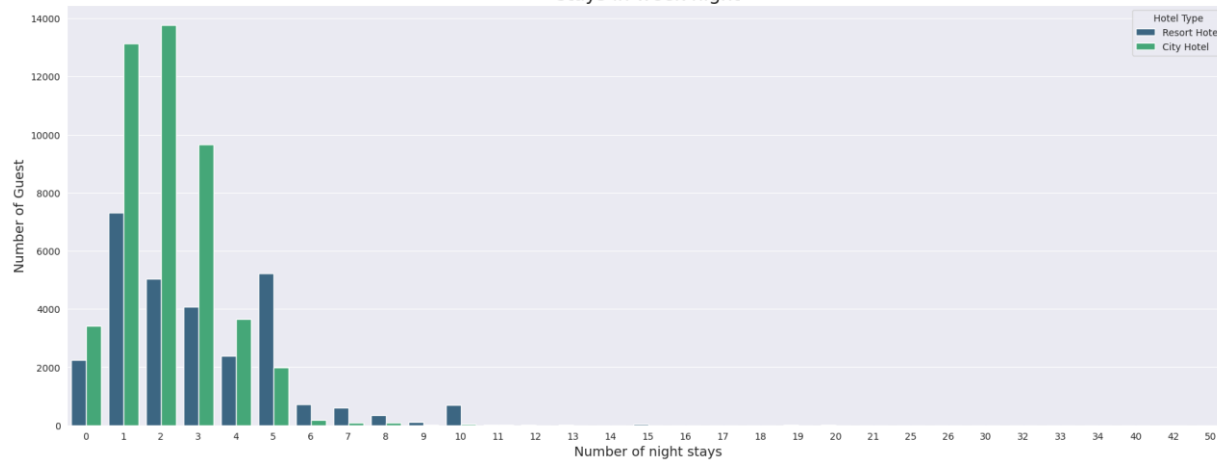


- As seen from the plot that in the month of May, most of the bookings were cancelled in City Hotel while in the month of August most of the cancellation done in Resort.



- Maximum number of Booking done in August in City Hotel and Resort Hotel that are Not-Cancelled. Summer vacation can be considered as the reason for this increased bookings.

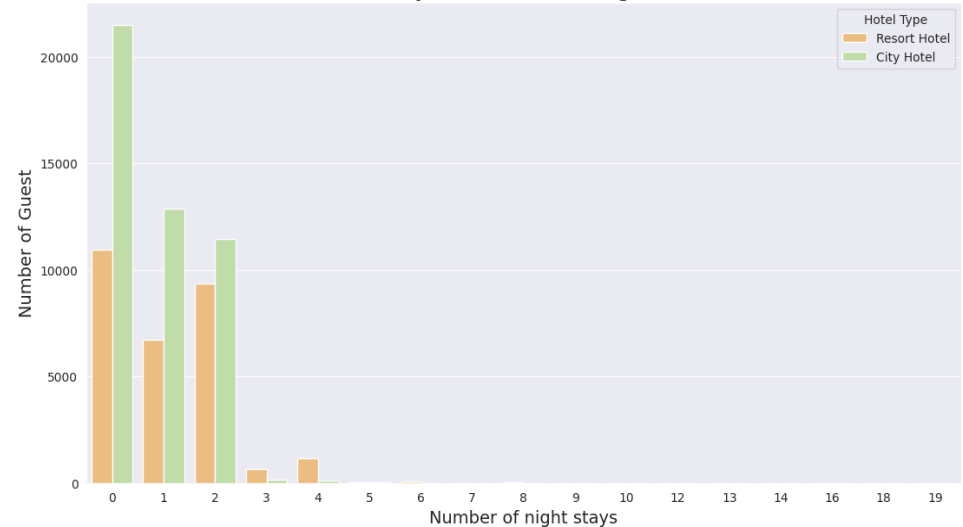
stays in week night



## OPTIMAL STAY IN EACH HOTEL TYPE

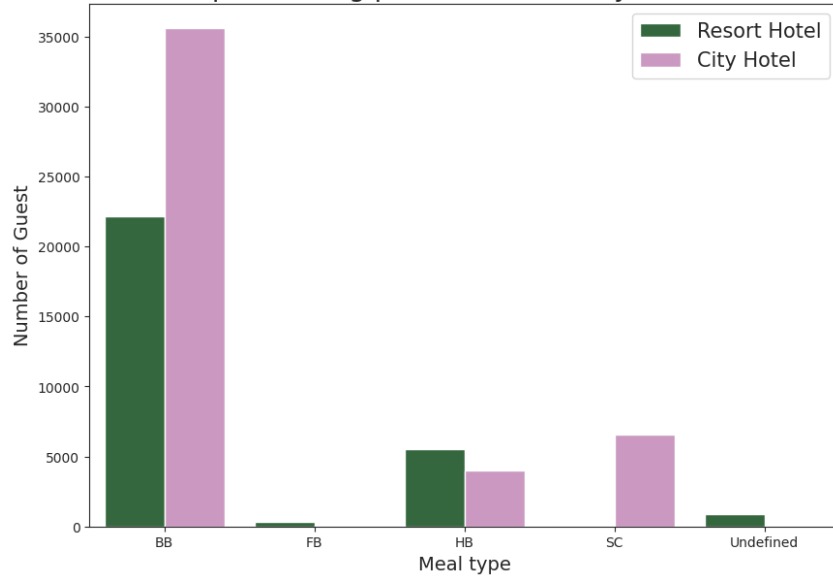
- It was observed that during weeknight most of the customers preferred to stay for 2 days in City Hotel and 1 day in Resort Hotel.
- While most of the customer stays for 0 nights in Resort and City hotels during weekdays.
- From this we can conclude that maximum people prefer to stay out during week night.

stays in weekend nights



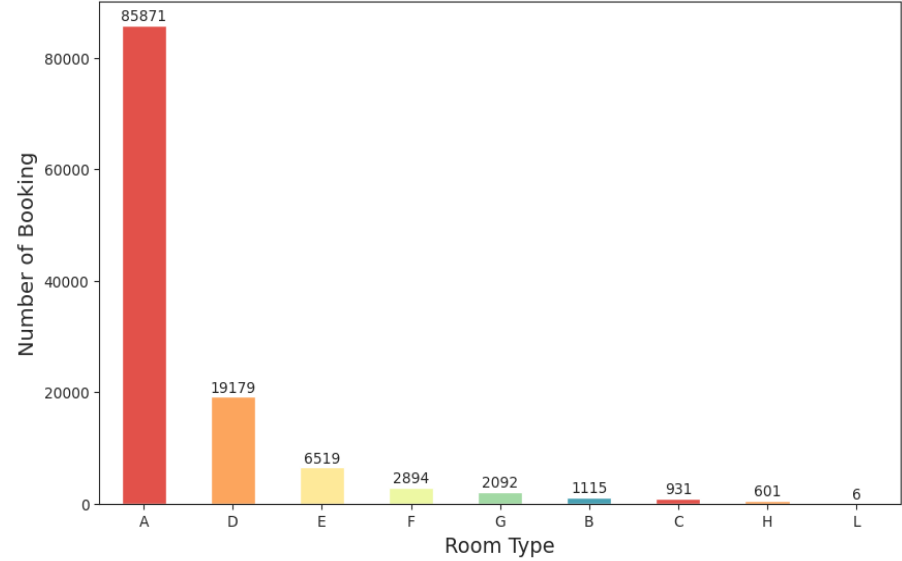
## PREFERRED MEAL AND ROOM TYPE BY THE CUSTOMER

Graph showing preferred meals by customer



- As seen from the above graph we can infer that BB( Bed & Breakfast) is the most preferred type of meal by the guests

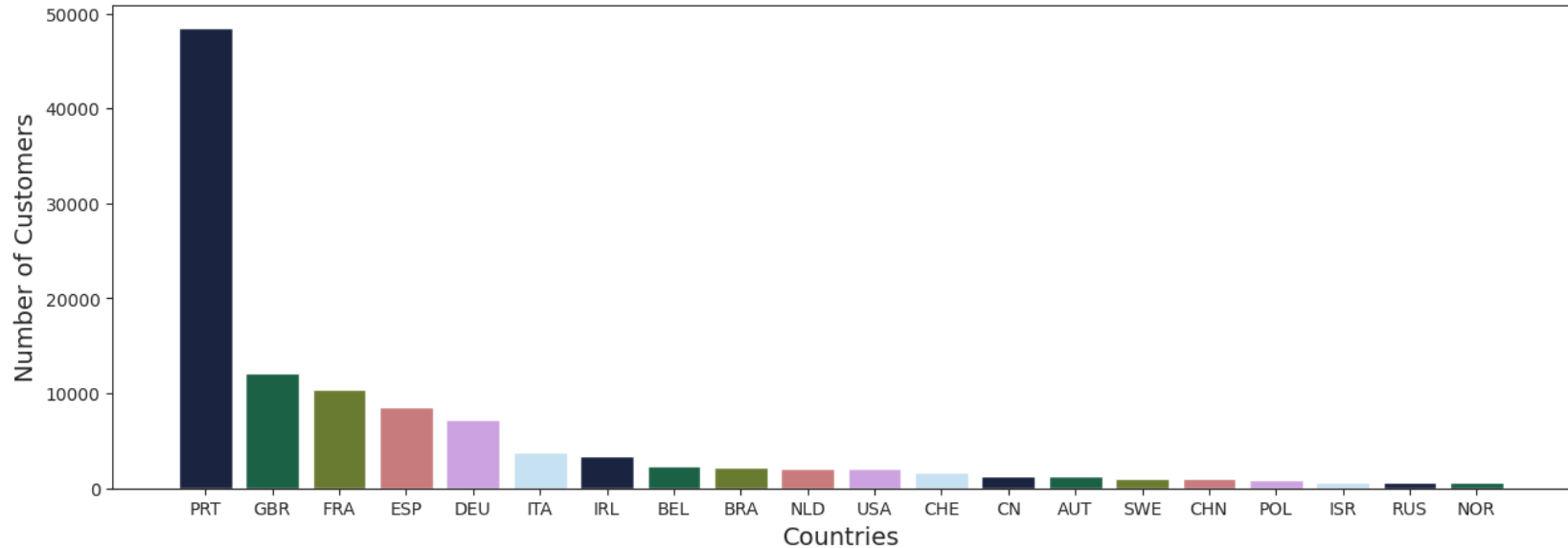
Bar Graph of hotel



- Similarly we can say that the most preferred Room type by the customer is type "A" bookings

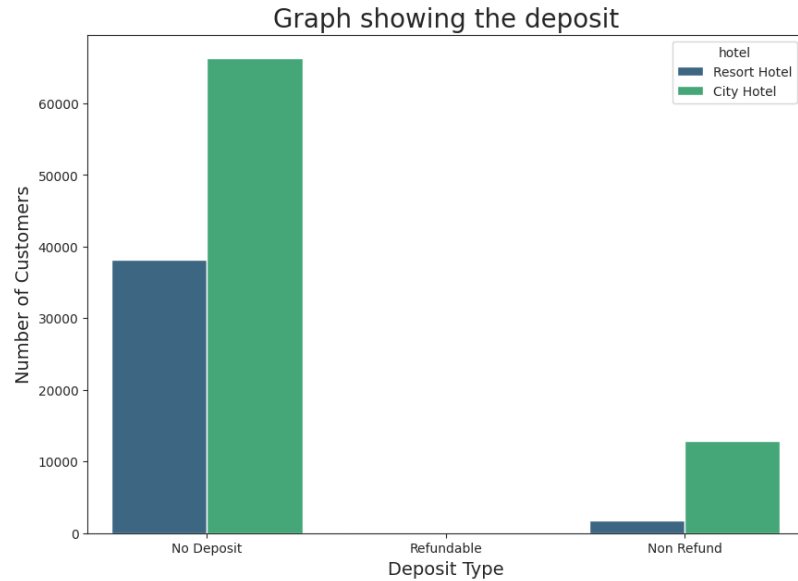
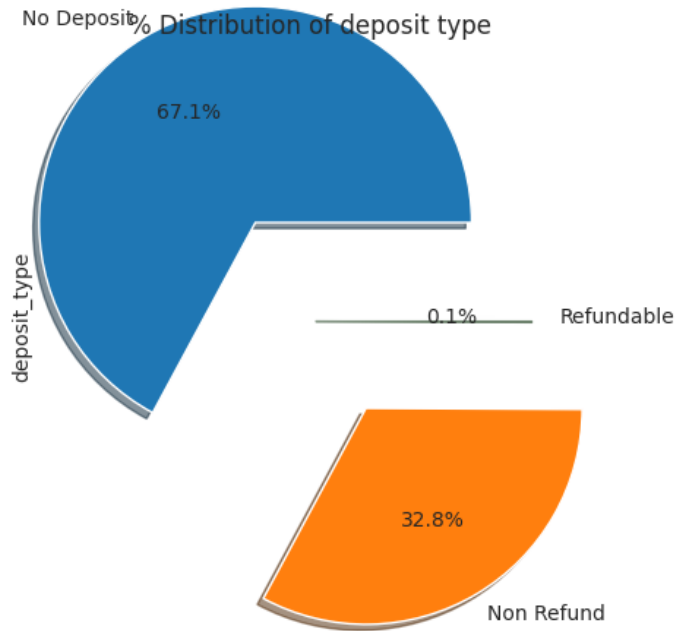


## Customer and their countries



- It is observed that most of the guests are coming from Portugal that is more 25000 guests are from Portugal

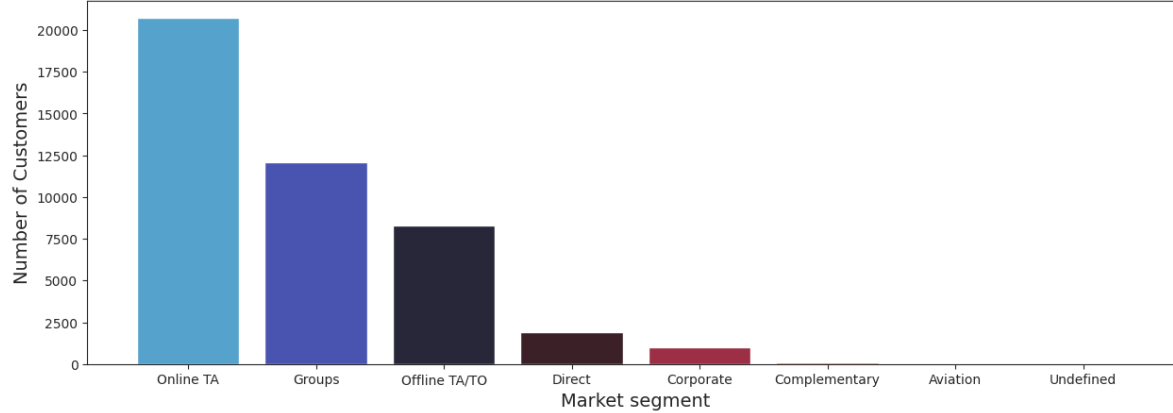
## DISTRIBUTION OF DEPOSIT TYPE



- It is observed that most of the customers prefer “No Deposit” (67.1%) for their booking

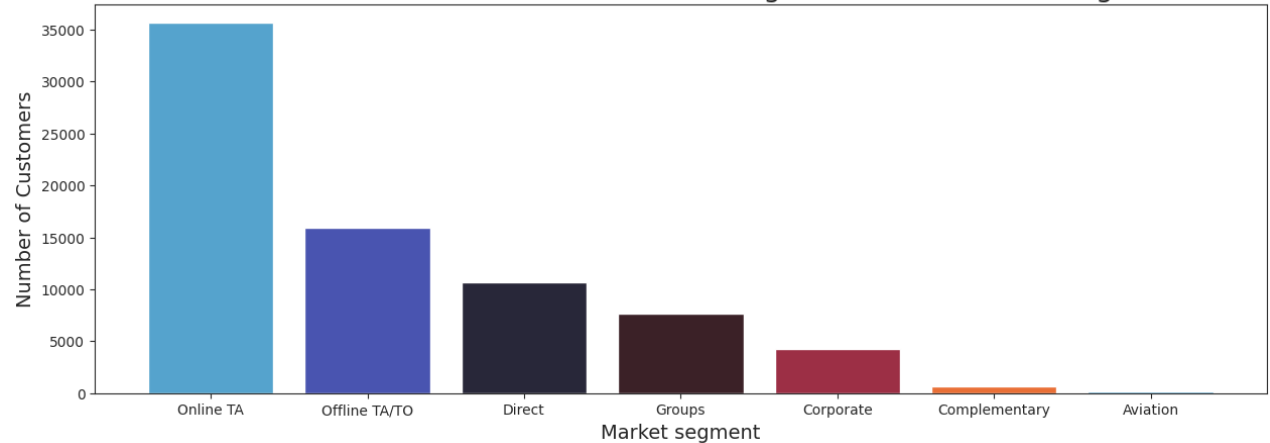
## AVERAGE ADR ACR MARKET SEGMENT V/S AVERAGE ADR EACH HOTEL TYPE

Number of customers cancelling booking and related to the below market segment

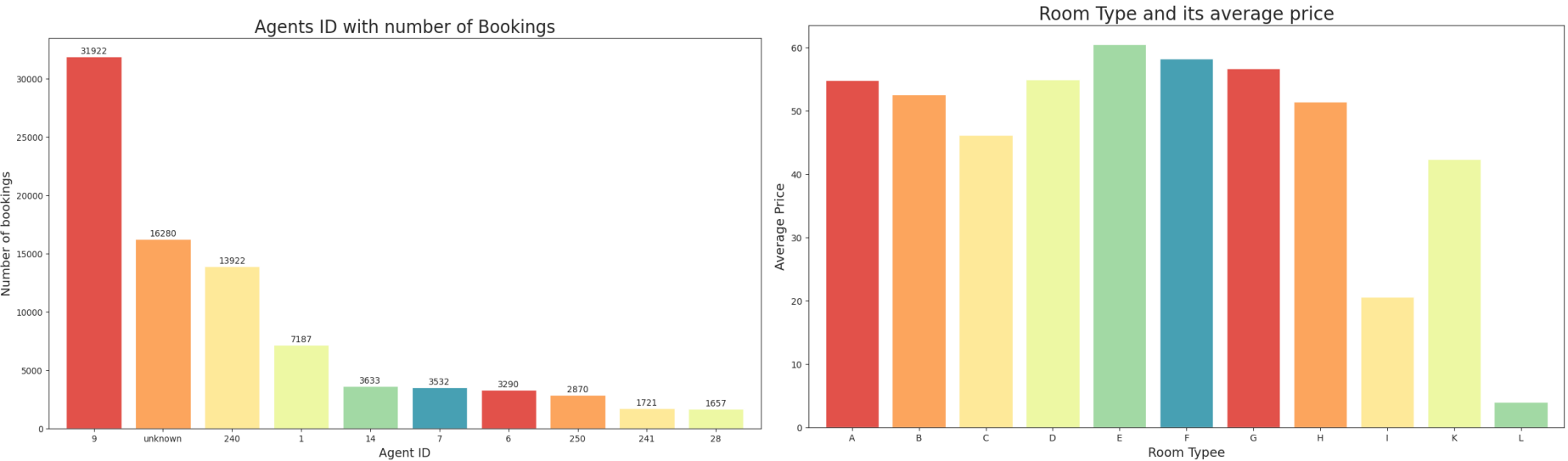


- It is observed that most of the customer from 'Online TA' are not cancelling the booking, one need to keep into mind that's its cancelling rate is also high.

Number of customers not cancelled booking from below market segments

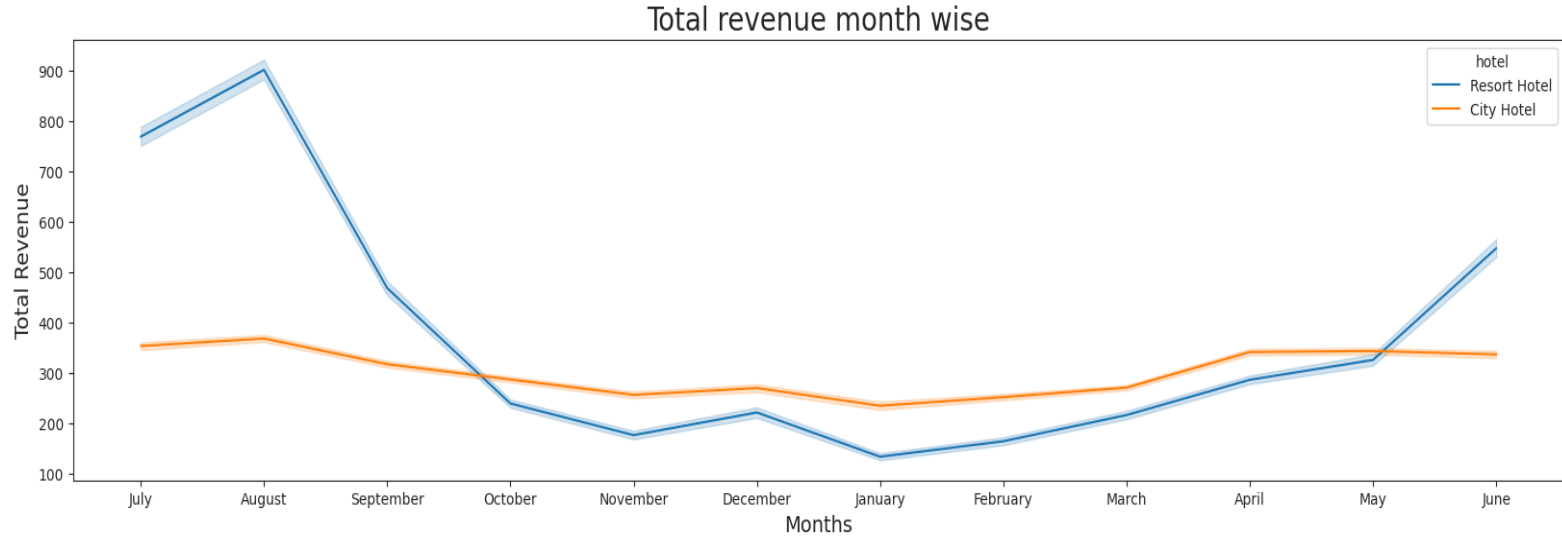


# CONTRIBUTION OF AGENT IN BOOKING AND PREFERRED ROOM TYPE V/S ITS AVERAGE PRICE



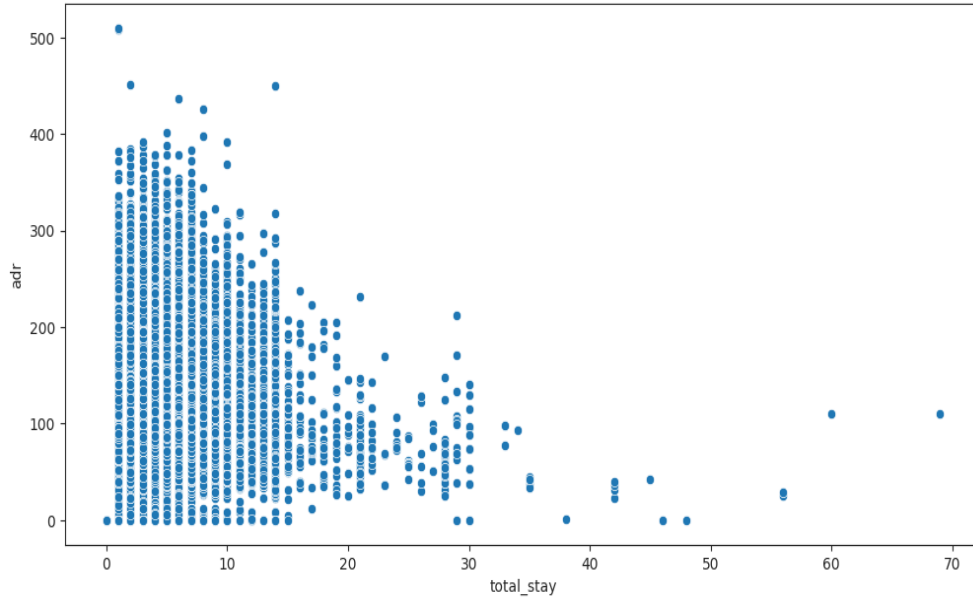
- It is observed that Agent ID-9.0 has contributed in the getting maximum number of bookings
- The given visualization show that the average price of Room type E is the maximum and the average price of Room type L is minimum.

## TOTAL REVENUE PER MONTH PER HOTEL TYPE

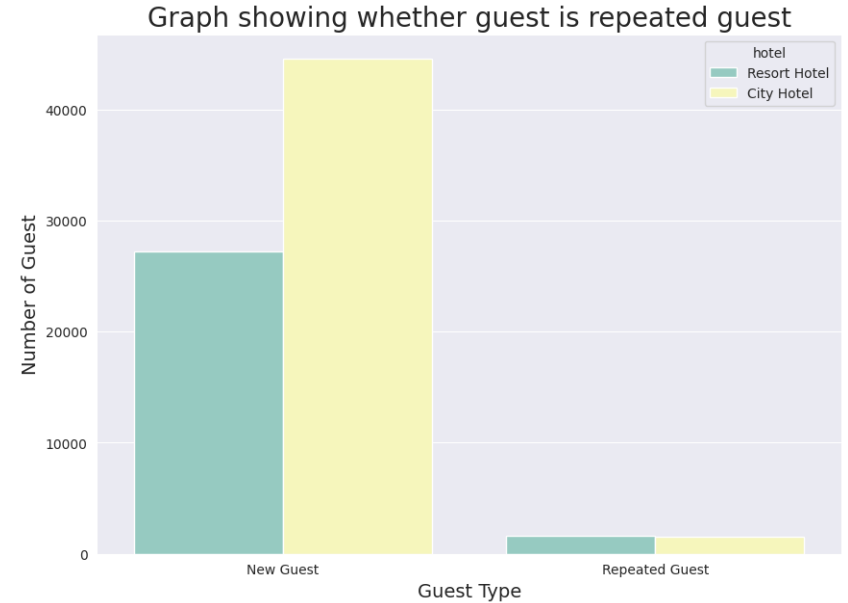


- Revenue of Resort Hotel are maximum in the month of August and least in the month of January, and there is no high fluctuation in the price of city Hotels throughout the Year.
- Summer vacation can be considered as the reason for this increased bookings.

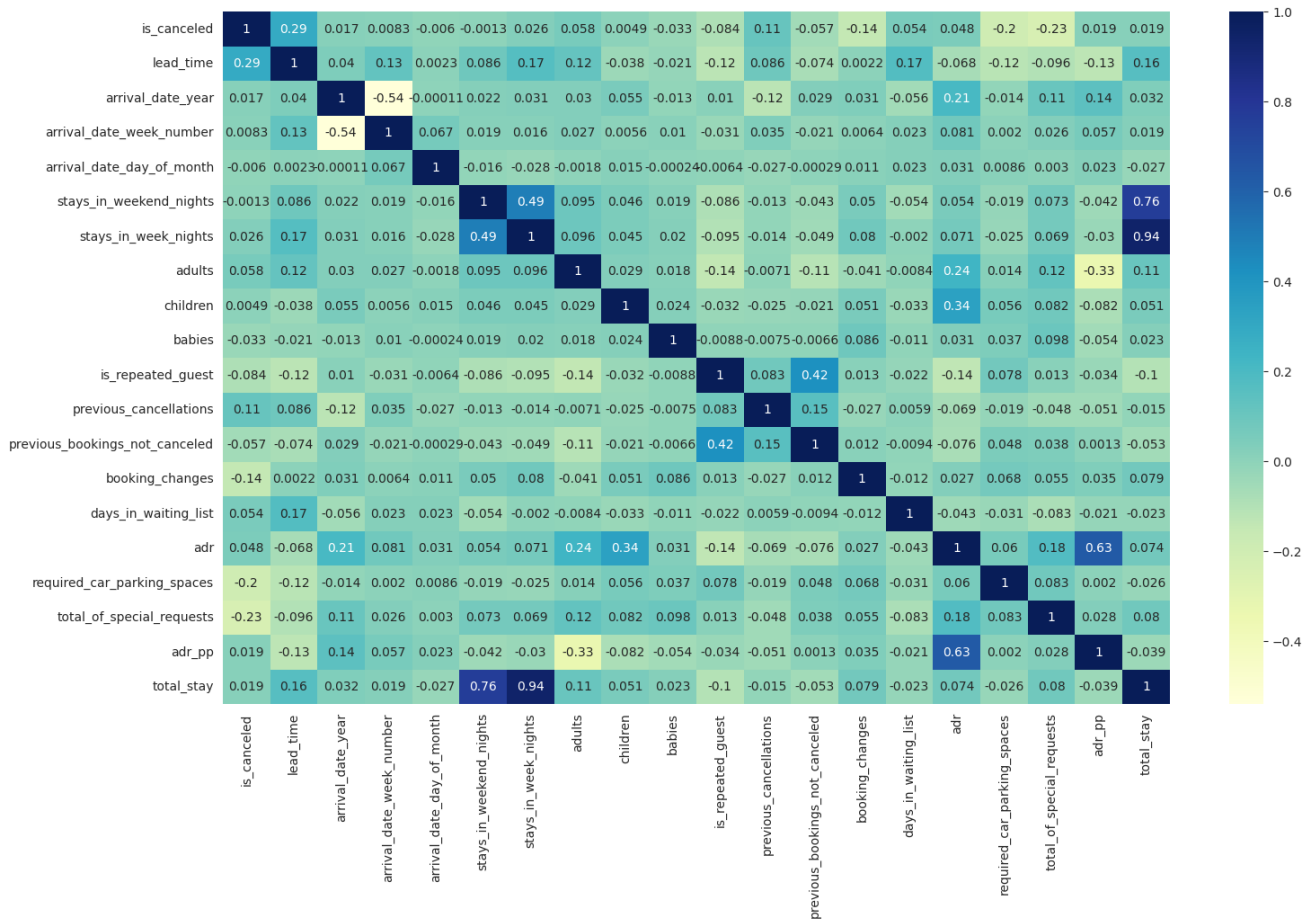
## ADR V/S TOTAL STAY



## REPEATED GUEST



- The scatter plot indicates that as the **stay increases ADR is decreasing**. Thus for **longer stays customer can get good ADR**
- The bar chart indicates that the hotel business have more number of New Guest in both type of hotels.
- In other words the ratio old repeated guest is less as compared to the repeated guest indicating need to proved better services for customer retention and expansion.



- It is observed that `is_canceled` and `same_room_alloted_or_not` are **negatively correlated**. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.
- `lead_time` and `total_stay` is **positively correlated**. That means more is the stay of customer more will be the lead time.
- Total stay is highly correlated with `stays_in_weekend_nighrs` and `stays_in_week_nights`.
- `is_repeated_guest` and `previous_bookings_not_canceled` has **strong correlation**. which may be repeated guests are not more likely to cancel their bookings.

# MAJOR FINDINGS / CONCLUSION

- Majority of the people prefer **City hotels**. We can say City hotel is the busiest hotel.
- **Most of the bookings** for City hotels and Resort hotel were happened in **2016**.
- When we look at Month wise booking, it indicates that most of the bookings done in the Month of August in Resort Hotel and City Hotel.
- Booking **cancellation rate is high for City hotels**
- It is observed that in **City hotel** most people stay for **3 days** and in **Resort hotel** most people stay for **only 1 day**.
- It was observed that during weeknight most of the customers preferred to stay for 2 days in City Hotel and 1 day in Resort Hotel. While While most of the customer stays for 0 nights in Resort and City hotels during weekdays, indicating that people preference to stay out during week night.
- **BB( Bed & Breakfast)** is the **most preferred type of meal** by the guests and most preferred **Room type is "A"** bookings
- It was observed that **Maximum number of guests were from Portugal**
- It is observed that most of the customers prefer "No Deposit" (67.1%) for their booking
- It is observed that '**Direct**' and '**Online TA**' are contributing the **maximum** in both types of hotels.
- Agent ID-9.0 has made the most number of bookings and the average price of Room type E is the maximum and the average price of Room type L is minimum.
- Revenue of Resort Hotel are maximum in the month of August and least in the month of January, and there is no high fluctuation in the price of city Hotels throughout the Year.
- It is observed that **July and August months** had the **most bookings Summer vacation** can be considered as the reason for this increased bookings.
- **Very few people have revisited the hotels**. Indicating lower **retention rate** and warning bell to **formulate strategies** to boost up the business.



# SUGESSTIONS

- Majority people prefer room A type, so in order to increase the revenue it should focus more on providing such rooms to its customers.
- Probability of cancellation is higher when there is no deposits taken by the hotel, it is advisable to take minimum deposits to reduce the cancellation rate.
- Maximum booking and revenue will be earned in the month of July and August months , because of Summer vacation. Hotels should create attractive offers to attract more customers.
- Customer feedback approach can be adopted to know the reason of less repetitive guests and thereby expand the business.
- Maximum number of guests were from Portugal. Efforts should be taken to attract the customers from other European countries also.

Thank  
you!