Capstone Project-1

EDA On Hotel Booking Analysis

BY

Subham Behera

(Cohort Enlighten)

Problem Statement:

- For this project we will be analyzing Hotel Booking data. This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and babies, and the number of available parking spaces.
- Hotel industry is a very volatile industry and the bookings depends on above factors and many more.
- The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management ,which can perform various campaigns to boost the business and performance.

❖ Work Flow :

AI

➢ So we will divide our work flow into following 3 steps.

EDA will be divided into following 3 analysis.

1) **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.
2) **Bivariate analysis:** Bivariate analysis is where you are comparing two variables to study their relationships.
3) **Multivariate anlysis:** Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

❖ Data Collection and Understanding:

➢After collecting data, it is very important to understand the data. So we had hotel Booking analysis data Which had 1,19,390 rows and 32 columns. So let's understand these 32 columns.

Data Description:

hotel :Resort Hotel or City Hotel
is_canceled : Value indicating if the booking was cancelled (1) or not (0)
lead_time : Number of days that elapsed between the entering date of the booking and the arrival date
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
stays_in_week_nights : Number of week nights.
adults : Number of adults
children : Number of children
babies : Number of babies
meal : Type of meal booked.
country : Country of origin.

❖ Data Collection and Understanding:

market_segment : Market segment designation (TA/TO)
distribution_channel : Booking distribution channel(TA/TO)
is_repeated_guest : is 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 to the booking.
booking_changes : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type : No Deposit, Non Refund , Refundable.
agent : ID of the travel agency that 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 customer. Contract,Group,transient,Transient party.
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 (e.g. twin bed or high floor)
reservation_status : Reservation last status.

❖ Data Cleaning and Manipulation:

Step 1: Removing duplicate rows if any

```
[12] hb1[hb1.duplicated()].shape   # Show no. of rows of duplicate rows duplicate rows

     (31994, 32)

[71] # Dropping duplicate values
     hb1.drop_duplicates(inplace = True)

[85] hb1.shape

     (87229, 35)
```

Step 2: Handling missing values

```
# Columns having missing values.
hb1.isnull().sum().sort_values(ascending = False)[:6]
```

```
company                 82137
agent                   12193
country                   452
children                    4
reserved_room_type          0
assigned_room_type          0
dtype: int64
```

Null values in columns company and agent were replaced by 0.

Null values in column children were replaced by the mean of the column.

Null values in column country were replaced by 'others'.

```
hb1[['company','agent']] = hb1[['company','agent']].fillna(0)

hb1['children'].fillna(hb1['children'].mean(), inplace = True)

hb1['country'].fillna('others', inplace = True)
```

❖ Data Cleaning and Manipulation:

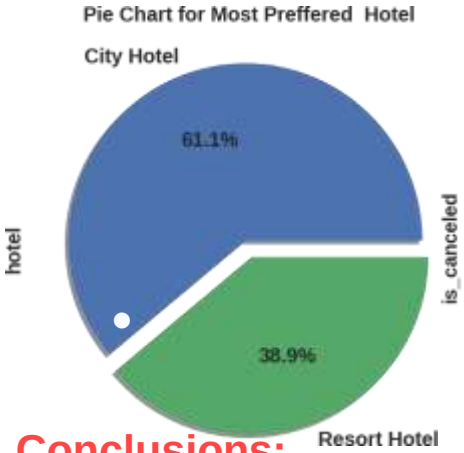
Step 3: Converting columns to appropriate datatypes.

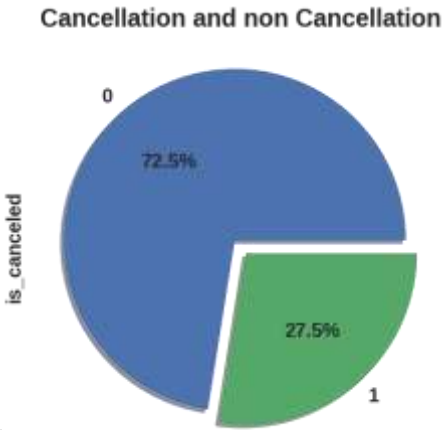
```
[113] # Converting datatype of columns 'children', 'company' and 'agent' from float to int.
      hb1[['children', 'company', 'agent']] = hb1[['children', 'company', 'agent']].astype('int64')

[114] # changing datatype of column 'reservation_status_date' to data_type.
      hb1['reservation_status_date'] = pd.to_datetime(hb1['reservation_status_date'], format = '%Y-%m-%d')
```

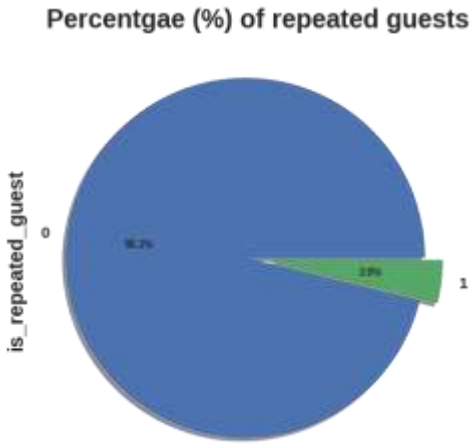
Step 4: Adding important columns.

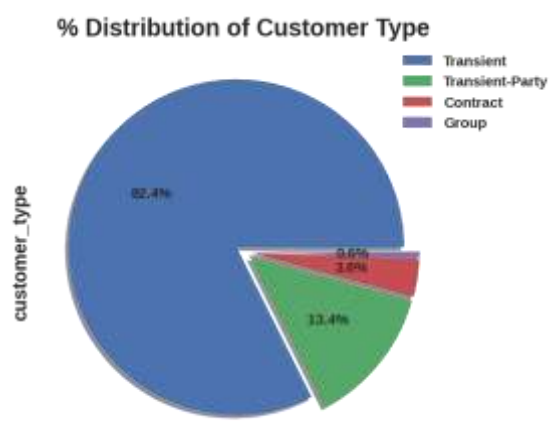
```
[115] # Adding total staying days in hotels
      hb1['total_stay'] = hb1['stays_in_weekend_nights']+hb1['stays_in_week_nights']

      # Adding total people num as column, i.e. total people num = num of adults + children + babies
      hb1['total_people'] = hb1['adults']+hb1['children']+hb1['babies']
```

❖ Exploratory Data Analysis (EDA) :







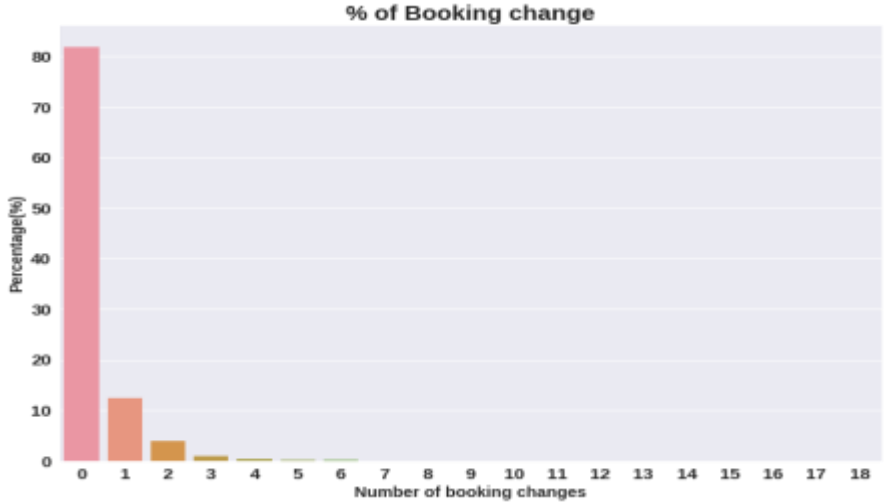


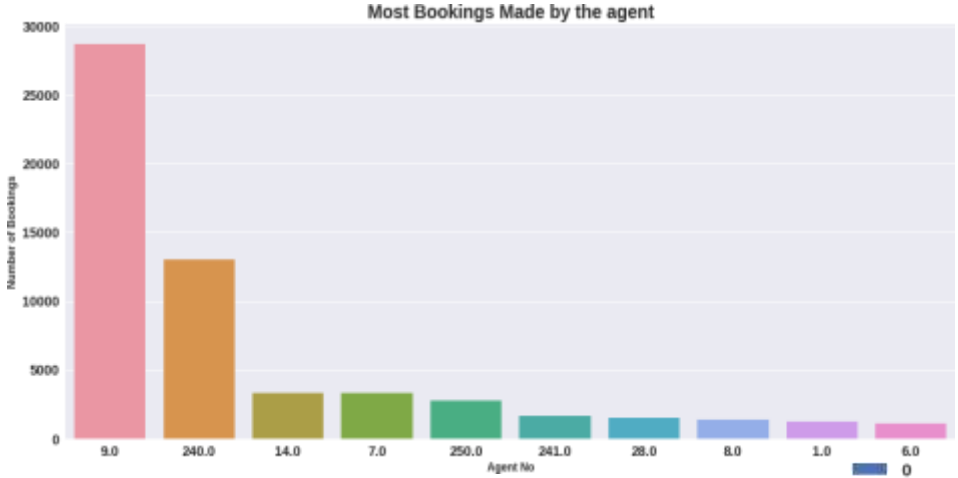
Conclusions:

- City hotels is the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- 27.5 % bookings were cancelled out of all the bookings
- .Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- Most of the customers/guests were Transient type(82.4%). And transient party were 13.4% and 0.6 belongs to group. Remaining guests belongs to Contract type.

Contract-when the booking has an allotment or other type of contract associated to it
Group -when the booking is associated to a group
Transient-when the booking is not part of a group or contract, and is not associated to other transient booking
Transient-party-when the booking is transient, but is associated to at least other transient booking

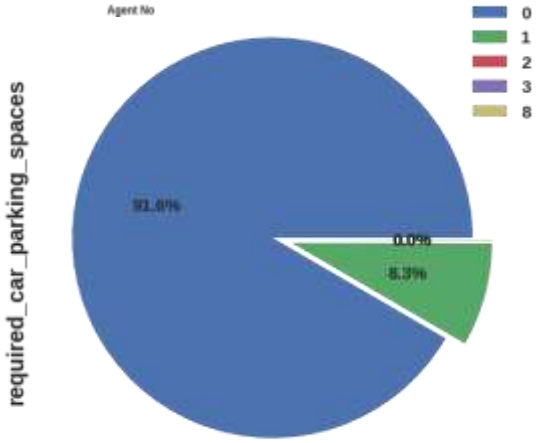
❖ Exploratory Data Analysis (EDA) :



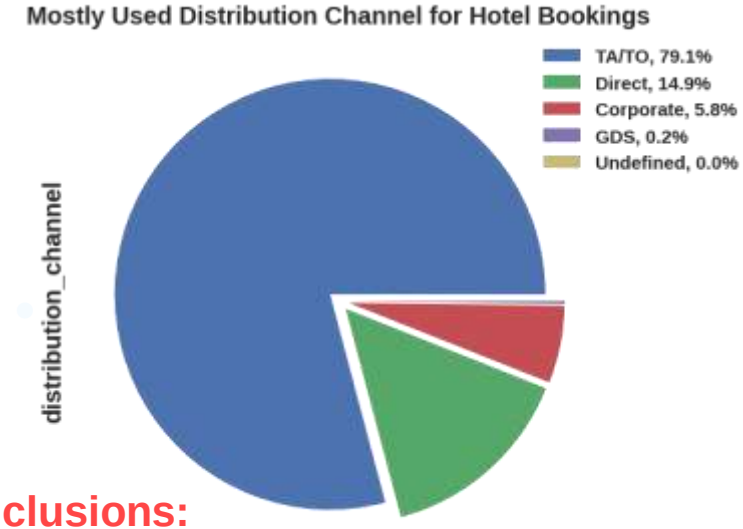


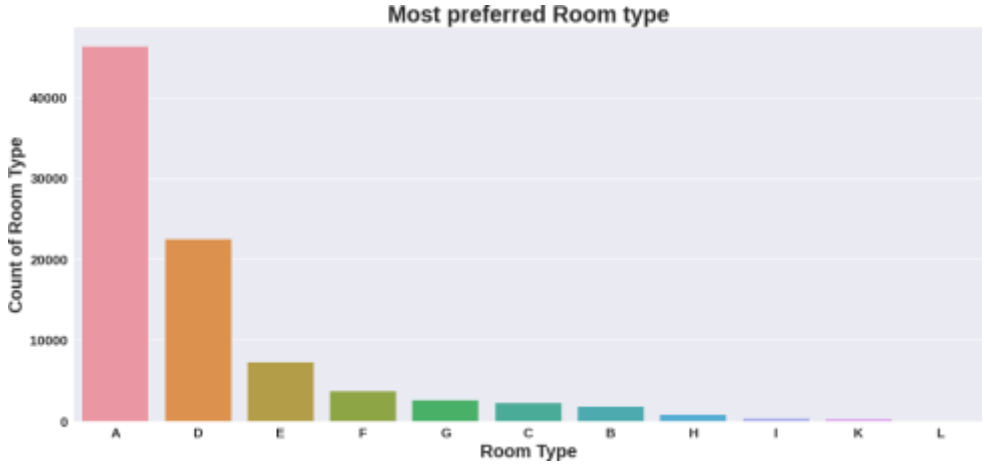
Conclusions:

- The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.
- Agent Id no -9 made the highest bookings which is more than 28721.
- Most of the customers(91.6%) do not require car parking spaces. Only 8.3 % people required only 1 car parking space.



❖ Exploratory Data Analysis (EDA) :



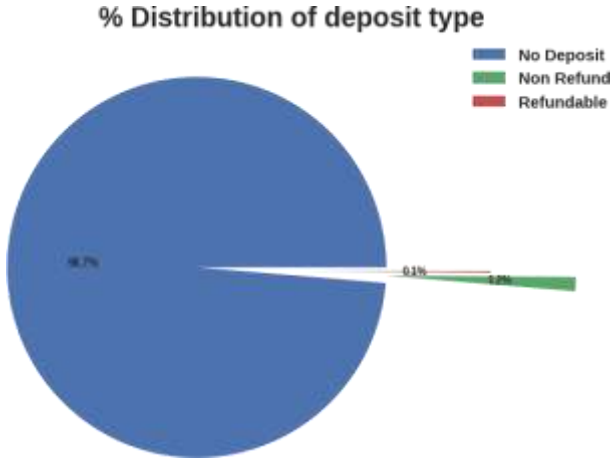


Conclusions:

➢ 79.1 % bookings were made through TA/TO (travel agents/Tour operators).Second most channel is direct.

➢ Room type ‘A’ is most preferred by the guests second most preferred is ‘D’.

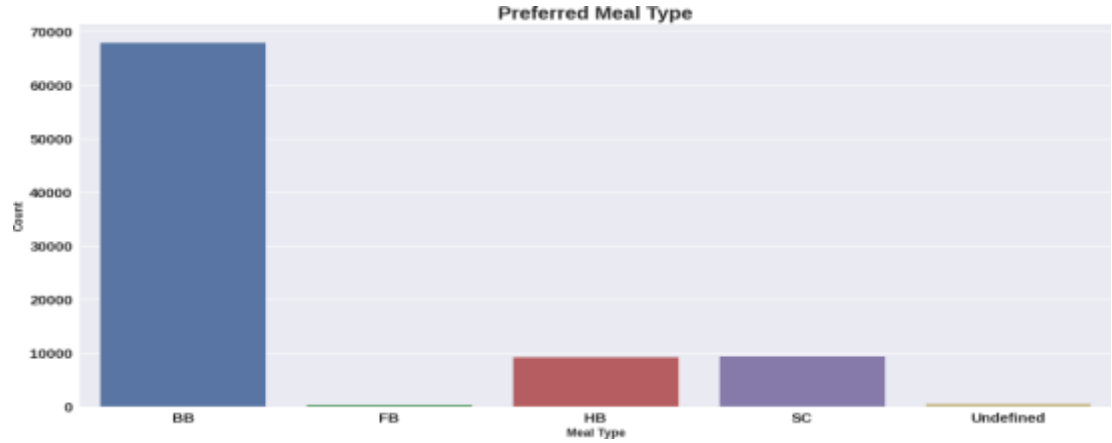
➢ Almost 98.7% of the guests prefer ‘No deposit’ type of criterion while booking hotels.



❖Exploratory Data Analysis (EDA) :

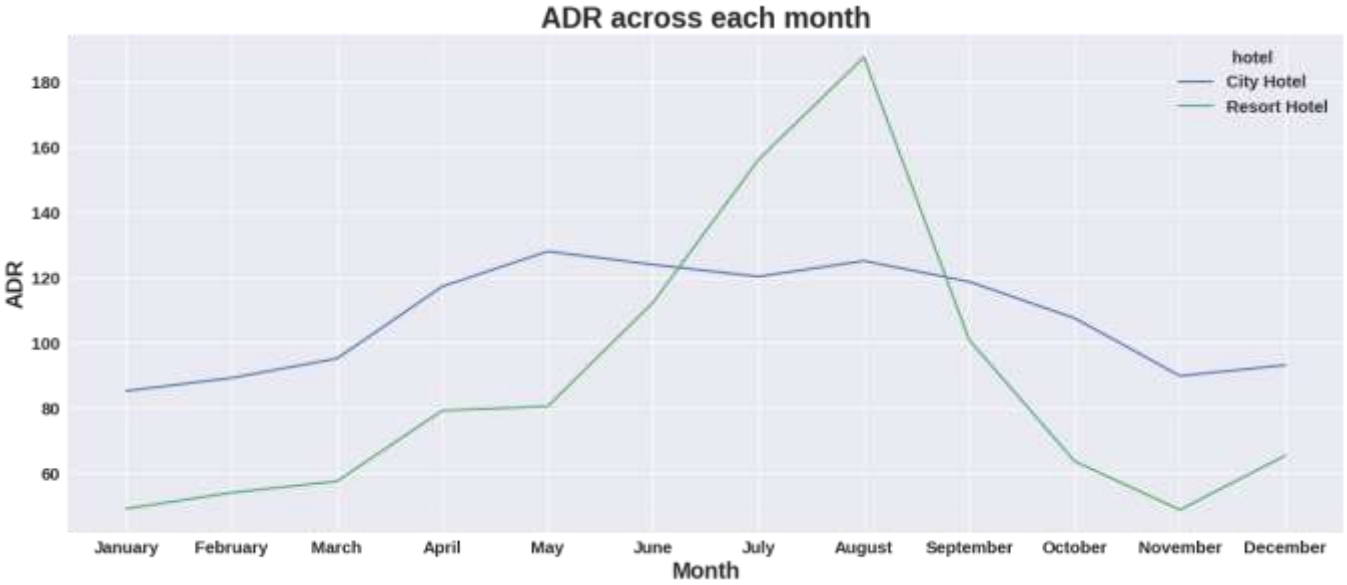
Conclusions:

- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Full Board i.e. FB is least preferred.
- HB (Half Board) and SC(Self Catering) are equally prefered.



- As we can see in the line chart, from June to September most of the bookings happened. It's Summer time. After September bookings Starts declining.

❖Exploratory Data Analysis (EDA) :

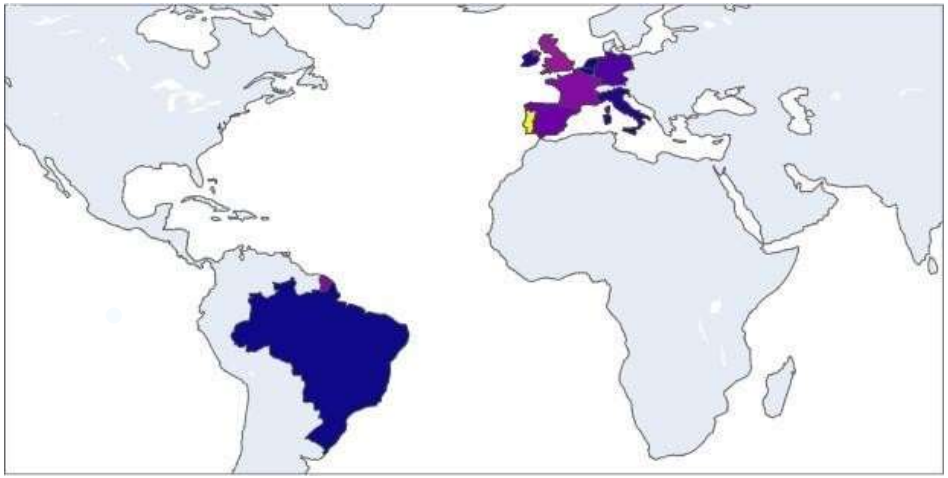


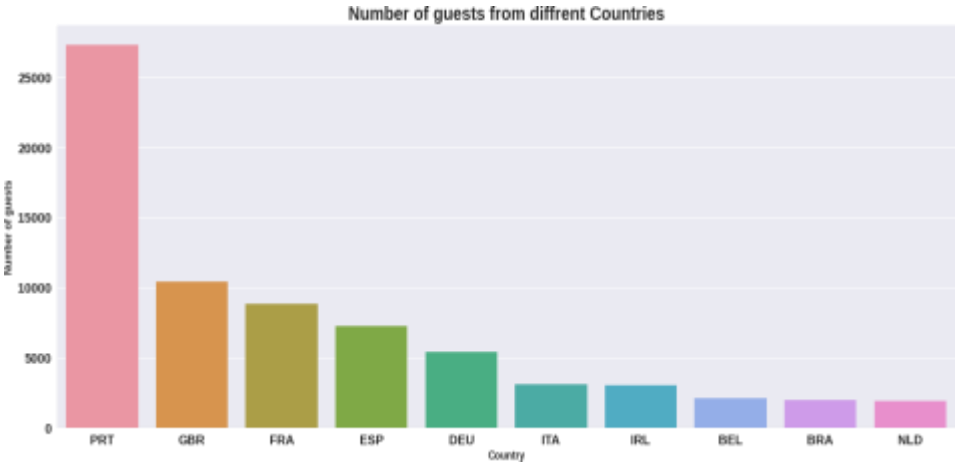
Conclusions:

➢Resort hotels had the highest adr in June ,July and August than the City hotels. But in other months adr of Resort hotel was less than the City hotels.

➢Thus we can say that, the January, February, March, April ,November and December are the good months for customers to get good adr

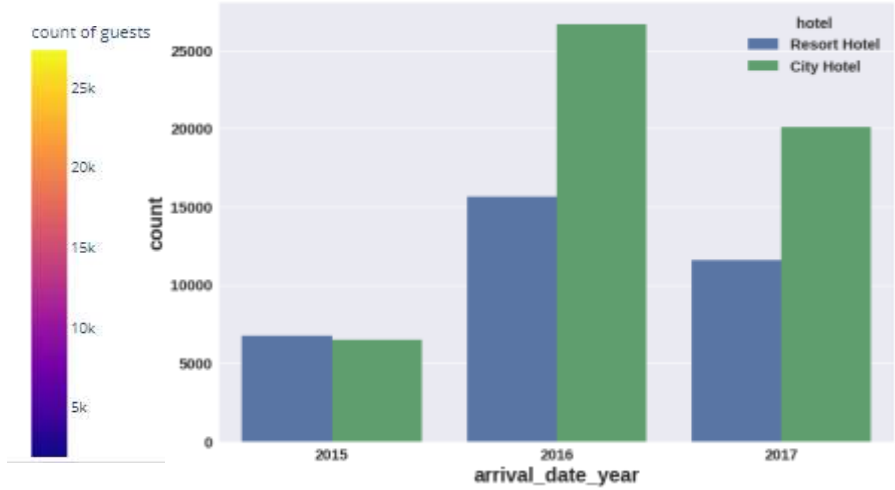
❖Exploratory Data Analysis (EDA) :



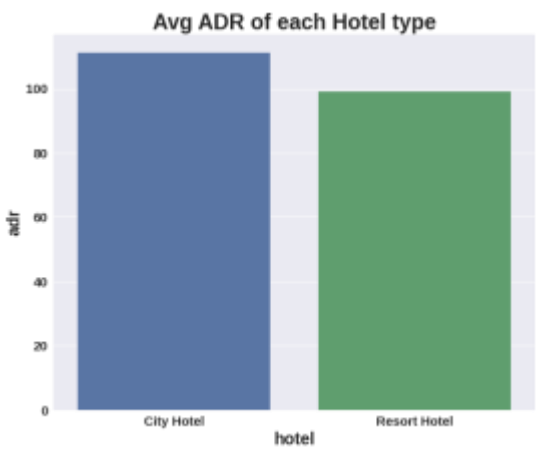


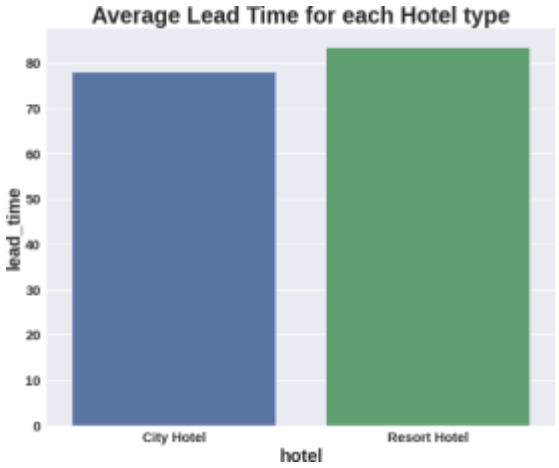
Conclusions:

- Maximum number of guests were from Portugal. i.e. more than 25000 guests.
- After Portugal, GBR(Great Brittan),France and Spain are the countries from where most of the guests came.
- Most of the bookings for City hotels and Resort hotel were happened in 2016. As we can see Most of the bookings were for City hotels.



❖Exploratory Data Analysis (EDA) :





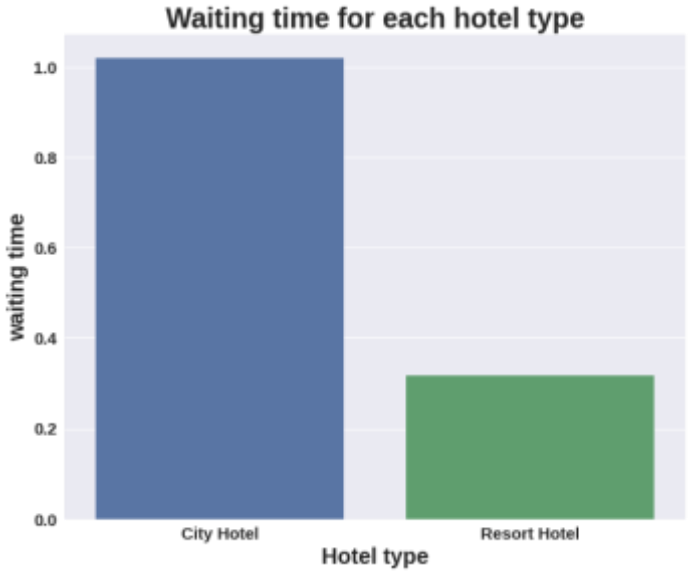
Conclusions:

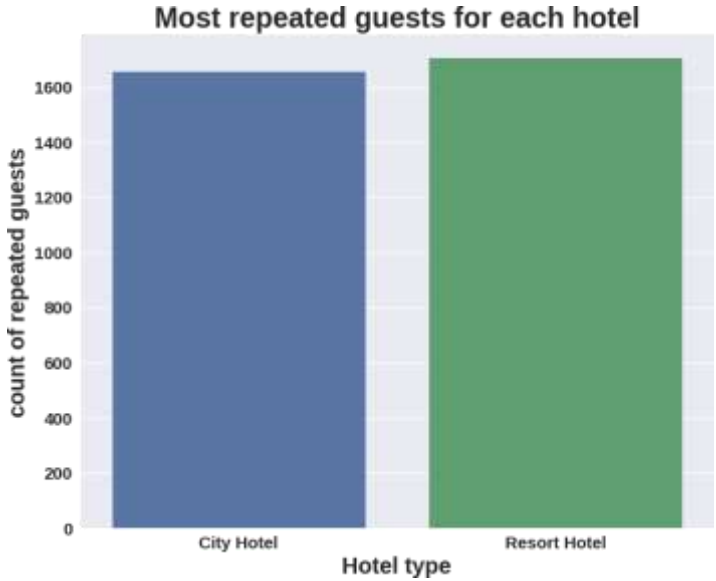
➢ Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.

➢ Average lead time for resort hotel is high. It means people plan their trip too early. Usually people prefer resort hotels for longer stays. That's why people plan early

➢ Booking cancellation rate is high for City hotels which almost 30 %.

❖Exploratory Data Analysis (EDA) :



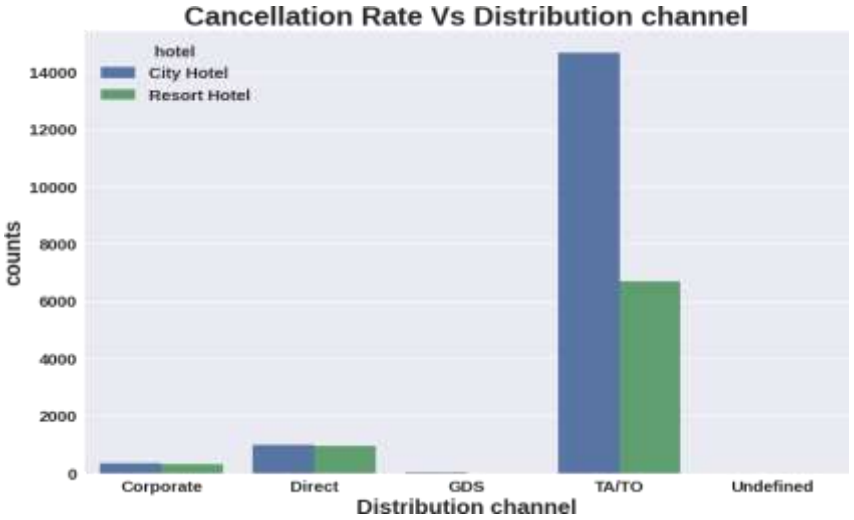


Conclusions:

➢Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.

➢Resort hotels has the most repeated guests. In order to get increase the count of repeated guests hotel management need to take the valuable feedbacks from the guests and try to give good service.

❖Exploratory Data Analysis (EDA) :



Conclusions:

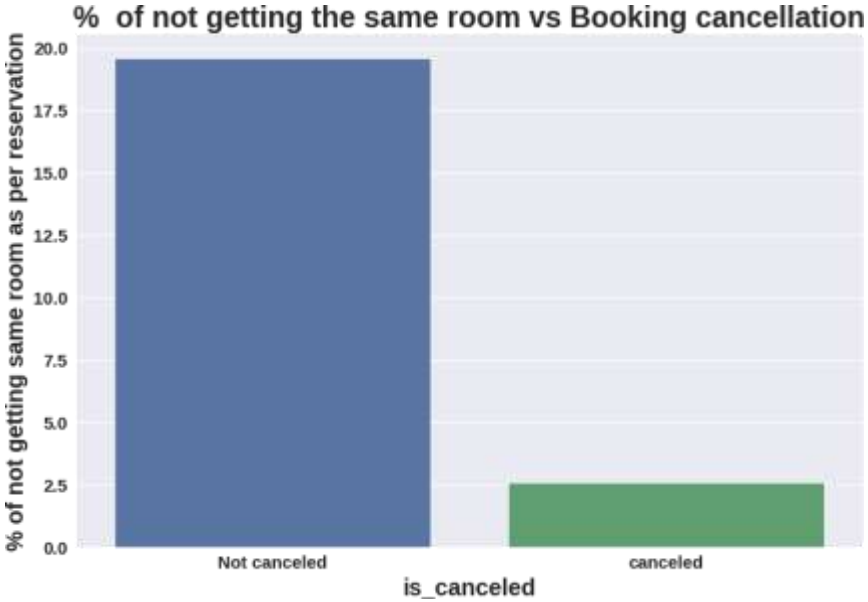
Distribution channel:

➢'TA/TO' distribution channel has highest cancellations for city hotels and more than 6000 cancellations for resort hotels. In order to reduce the cancellations they should improve their cancellation policies and deposit policies.

Market Segment:

➢ 'Online TA/TO' market segment has highest cancellations for city hotels.

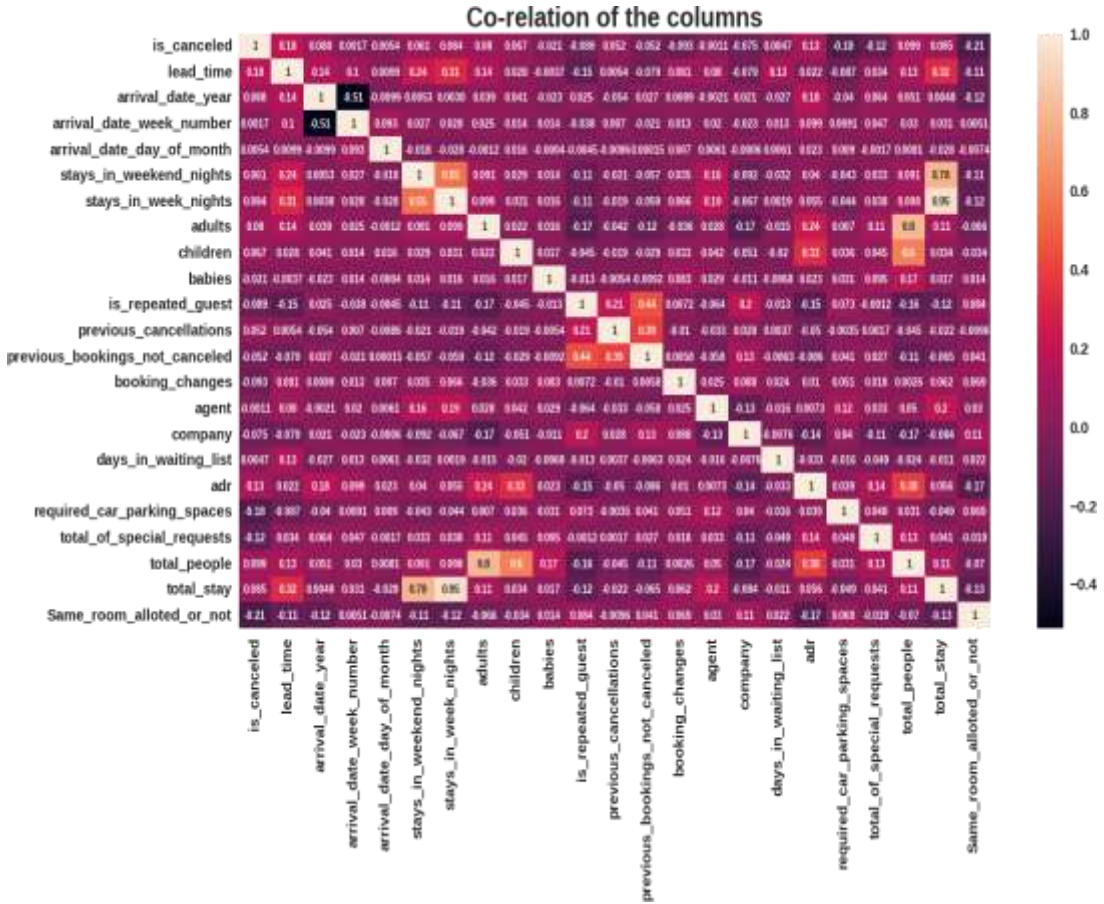
❖Exploratory Data Analysis (EDA) :



Conclusions:

- Almost 19 % people did not canceled their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.
- Thus not getting the same room as per reserved room is not the reason for booking cancellations.

❖Exploratory Data Analysis (EDA) :



Conclusions:

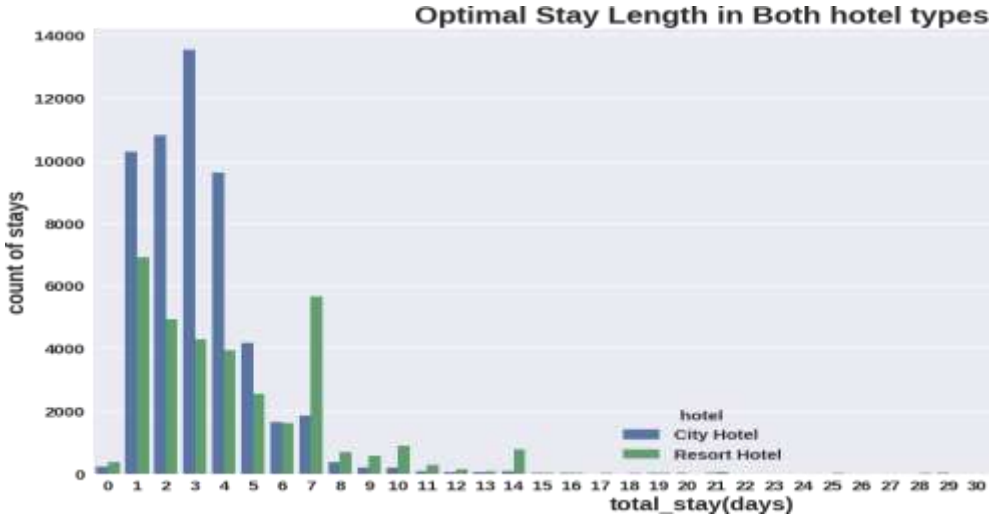
➢is canceled and same_room_alloted_or_not are negatively correlated. Not getting the same room as per reserved room is not the reason for booking cancellations.

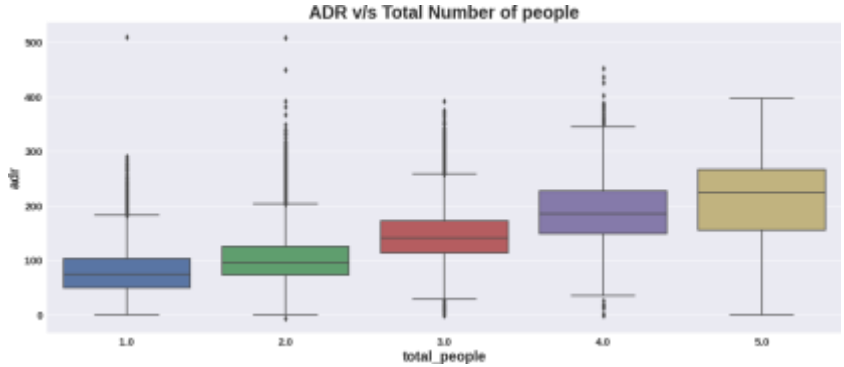
➢lead-time and total stay is positively correlated means that the more the guest's stay, more will be the lead time.

➢ADR and total people are highly correlated. That means more the people more will be the adr. High adr means high revenue

➢is_repeated_guest and previous_bookings Not_canceled has strong correlation. May be repeated guests are not more likely to cancel their bookings.

❖Exploratory Data Analysis (EDA) :





Conclusions:

- Optimal stay in both the type hotel is less than 7 days. Usually people stays for a week.
- For stay more than 7 days people likes to stay in Resort hotels. As we can see after 7 days City Hotel Bookings are very less as compared to Resort hotels.

- As we saw in Correlation heatmap, total people and adr are positively correlated. Thus for 2 people ,adr is almost 100 and for 5 people it is more than 200.
- Thus more the people, more will be the revenue of the hotels.

Signing off…

THANK YOU