Capstone Project

EDA On Hotel Booking Analysis

By

-Akash Sahu

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/or 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 :

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

Data Collection and Understanding	Data Cleaning and Manipulation	Exploratory Data Analysis(EDA)

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's very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 34 columns. So let's understand this 34 columns.

Data Description:
- **hotel** :Resort Hotel or City Hotel
- **is_canceled** : Value indicating if the booking was canceled (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 oforigin.
- **market_segment** : Market segment designation. (TA/TO)
- **distribution_channel** : Booking distribution channel.(T/A/TO)

❖ Data Collection and Understanding:

- **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.
- Service_rating** : *Customer gives rating on hotel service which is out 5 *
- Service_rating_text** : *Rating category*
  * 0 = No Rating,1 = Poor ,2 = Poor ,3 = Averge ,4 = Averge ,5 = Excellent

Data Cleaning and Manipulation:

➢ There were 4 columns company, agent, country and children with missing values.

```
#checking the null value
df.isnull().sum().sort_values(ascending=False)[:5]
```

	column	Null values
0	company	90396
1	agent	13422
2	country	460
3	children	4
4	hotel	0


```
# Filling/replacing null values with 0.
null_columns=['agent','children','company']
for col in null_columns:
    df[col].fillna(0,inplace=True)

#Successfully handled  Null Values
df.isna().sum().sort_values(ascending=False)[:5]
```

	Columns	Null values
0	hotel	0
1	days_in_waiting_list	0
2	reserved_room_type	0
3	assigned_room_type	0
4	booking_changes	0

➢Handling Duplicates: Data had 23125 duplicates values. So we dropped it from the data.

```
: df.duplicated().value_counts()

: False    96265
  True     23125
  dtype: int64
```

```
In [21]:  df= df.drop_duplicates()

In [22]:  # data set reduced
          df.shape

Out[22]:  (96265, 34)
```

❖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.
- 27.6 % bookings were got cancelled out of all the bookings
- .Only 3.6 % people were revisited the hotels. Rest 96.4 % were new guests. Thus retention rate is low.
- Most of the customers/guests were Transient type(78.8%). And transient party were 17.1% 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
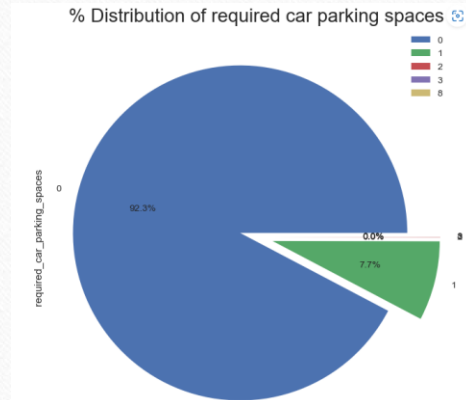**Transient-party-**when the booking is transient, but is associated to at least other transient booking

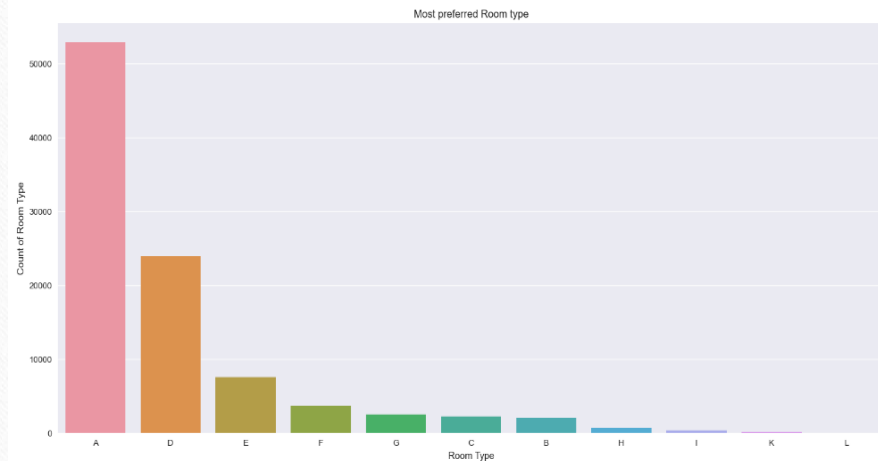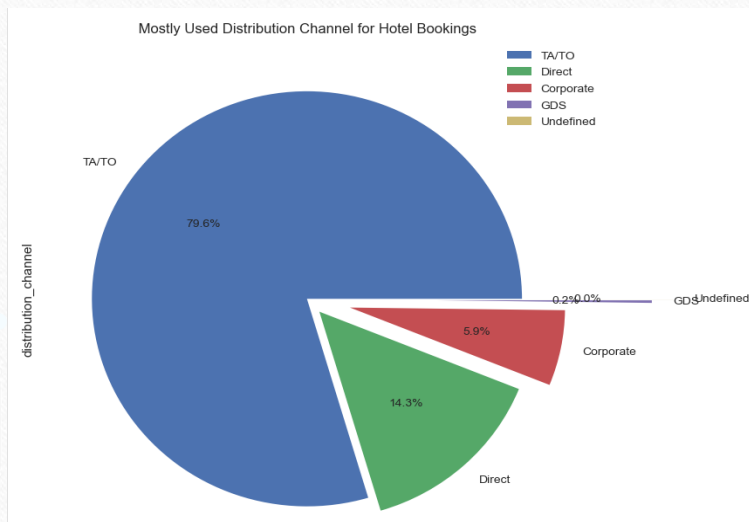# ❖Exploratory Data Analysis (EDA) :
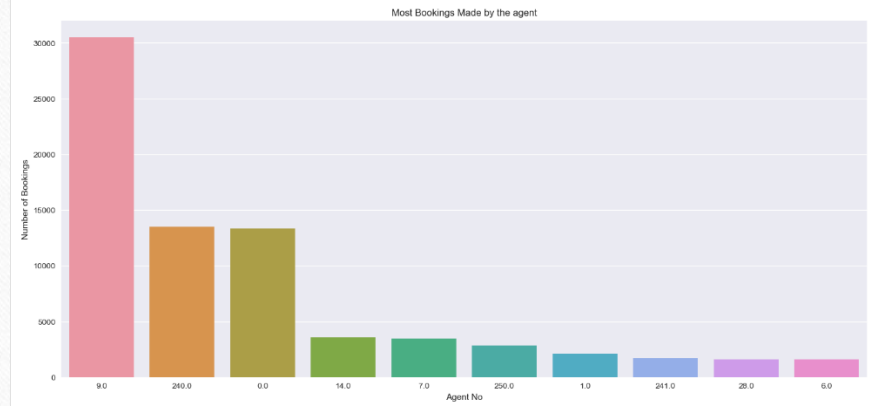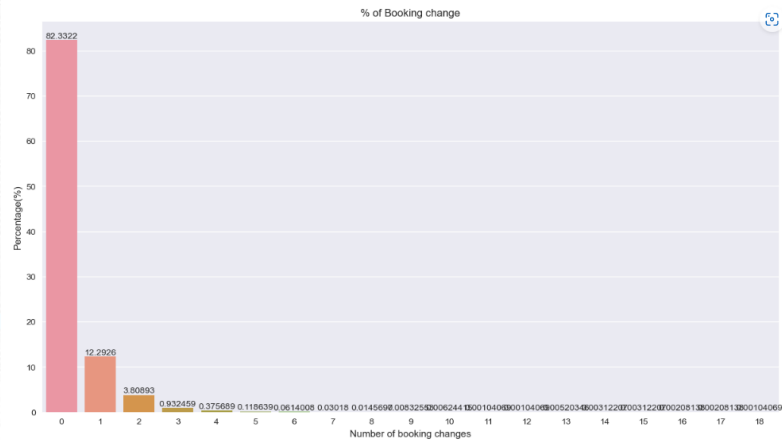
Conclusions:

➢The percentage of 0 changes made in the booking was more than 82.3 %. Percentage of Single changes made was about 10%.

➢ Agent Id no -9 made the highest bookings which is more than 30000.

➢ Most of the customers(91.6%) do not require car parking spaces. Only 7.7 % people required only 1 car parking space.
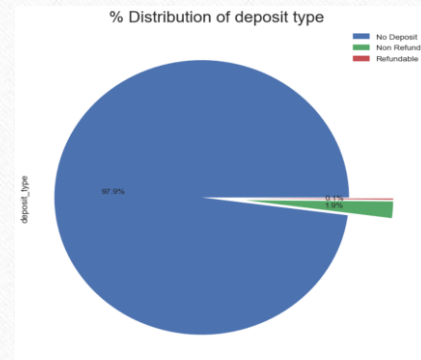
❖Exploratory Data Analysis (EDA) :

Mostly Used Distribution Channel for Hotel Bookings

Most preferred Room type

% Distribution of deposit type

Conclusions:

➢79.6 % 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 97.9% 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
   preferred.

Preferred Meal Type

➢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.

Number of bookings across each month
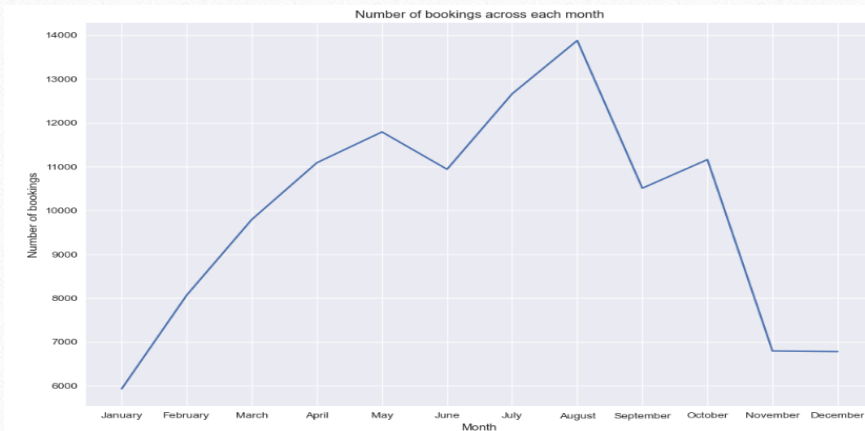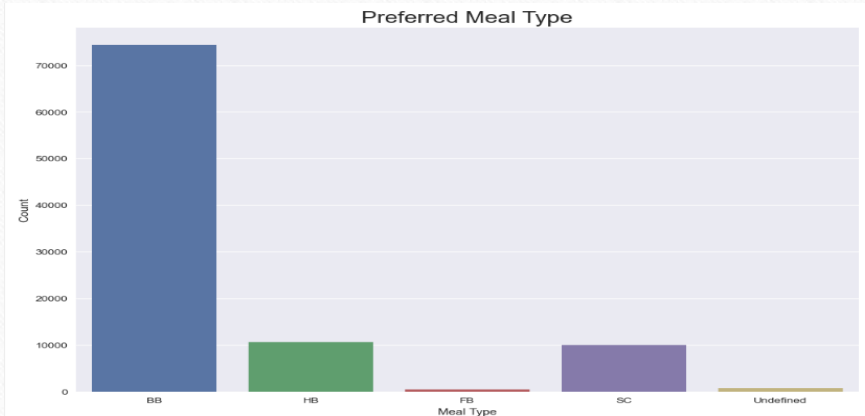
Exploratory Data Analysis (EDA) :
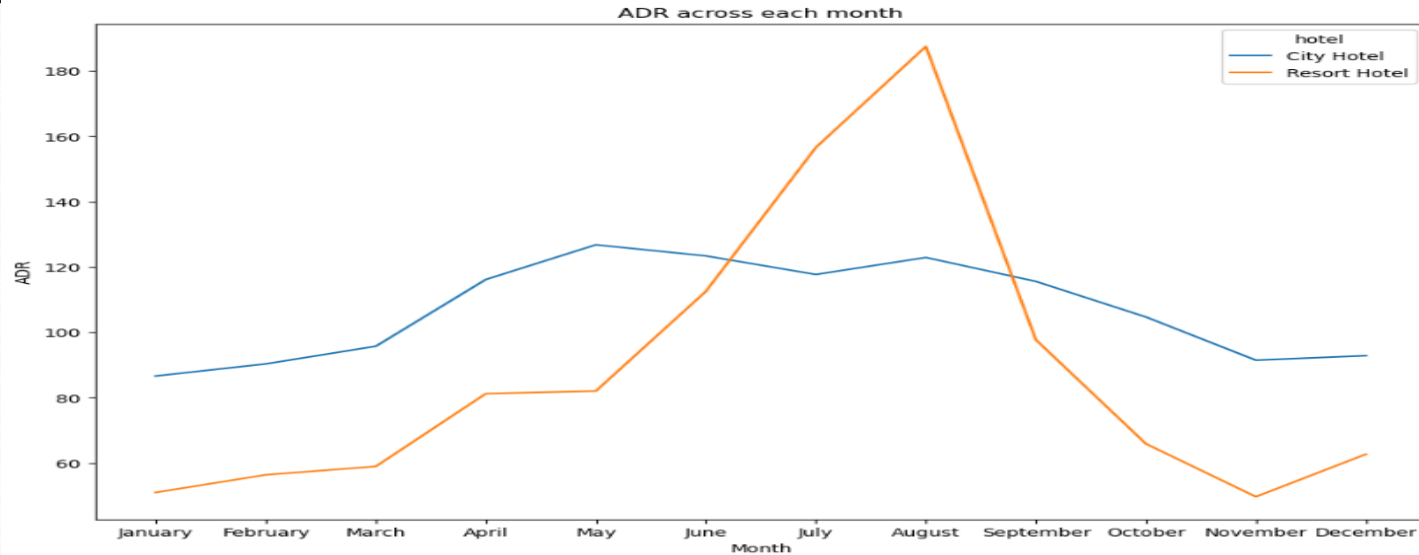
For Resort hotel is ADR is high in the months June, July, August as compared to City Hotels. May be Customers/People wants to spend their Summer vacation in Resorts Hotels.
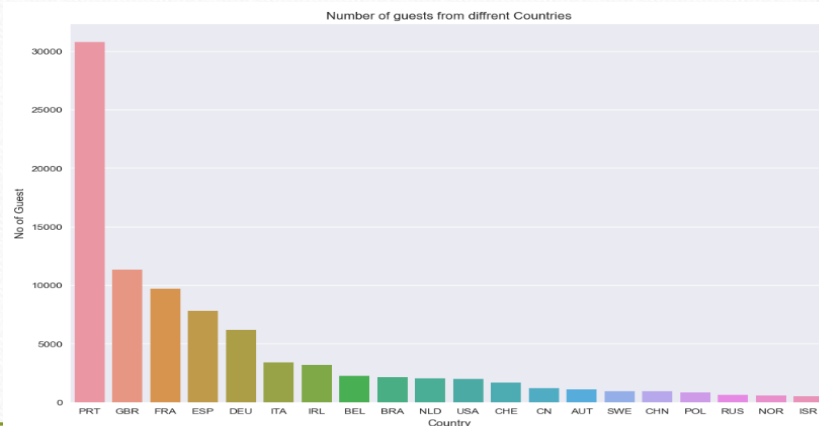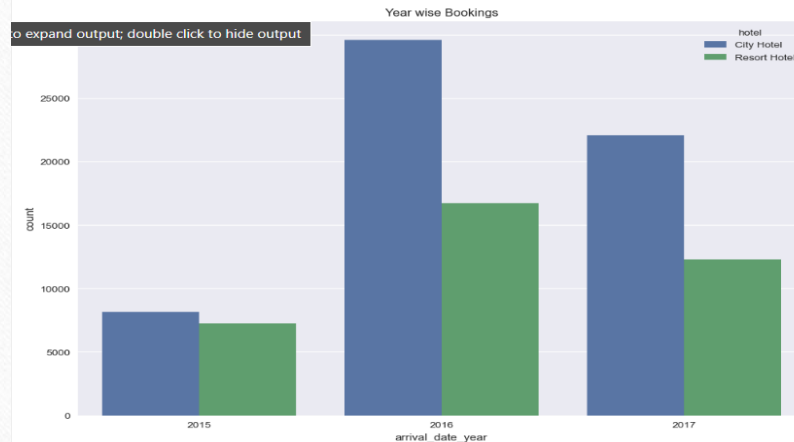
The best time for guests to visit Resort or City hotels is January, February, March, April, October, November and December as the average daily rate in this month is very low.
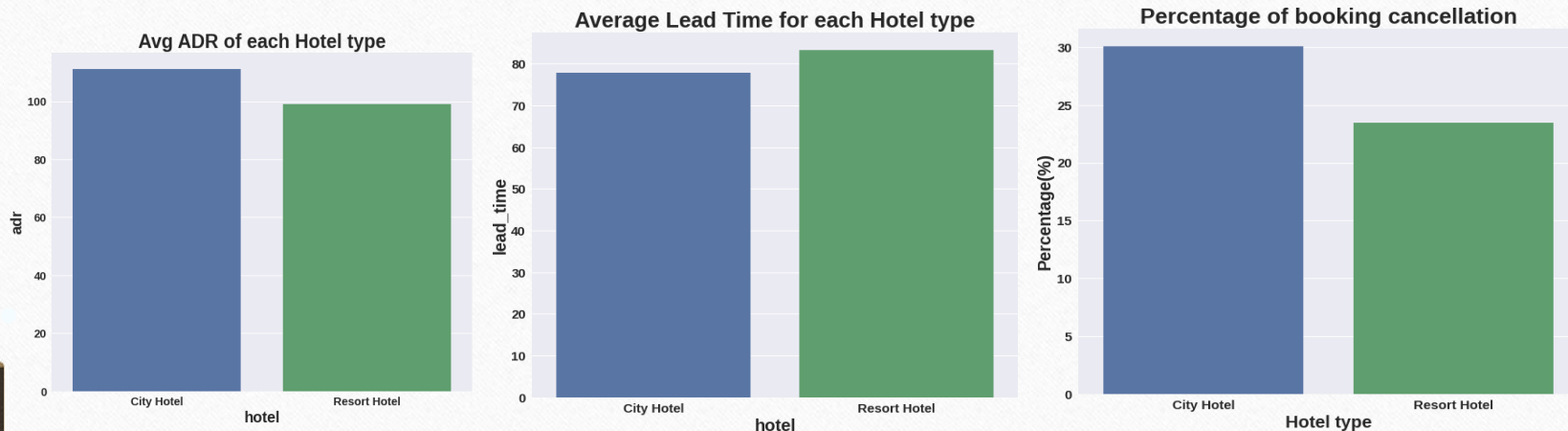
❖Exploratory Data Analysis (EDA) :

Year wise Bookings

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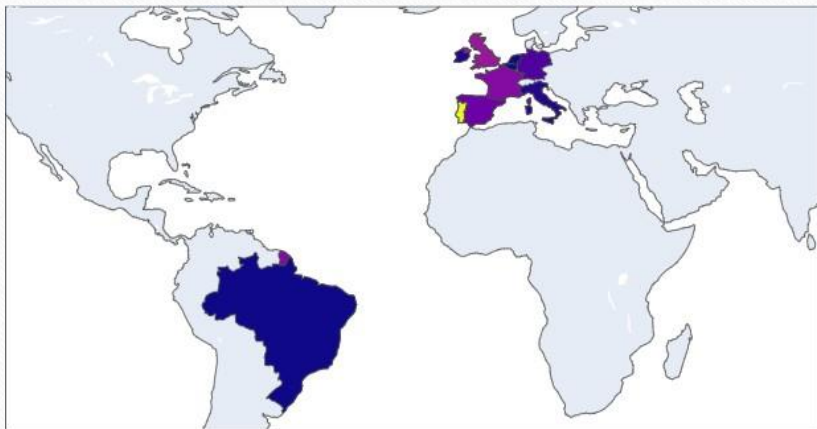
Conclusions:
➢ Maximum number of guests were from Portugal.
   i.e. more than 25000 guests.
➢After Portugal, GBR(Great Brittan),France and Spain are the countries from where most of the guests came.
➢ Most of the bookings for City hotels and Resort hotel were happened in 2016. As we can see Most of the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

❖ Exploratory Data Analysis (EDA) :

Avg ADR of each Hotel type — Average Lead Time for each Hotel type — Percentage of booking cancellation

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) :
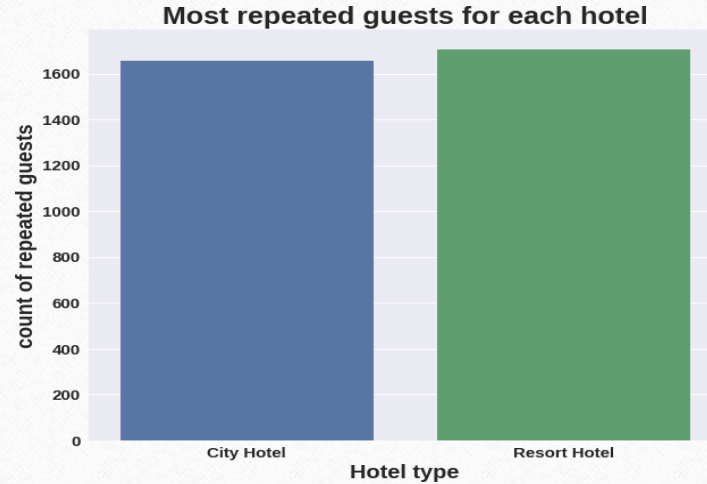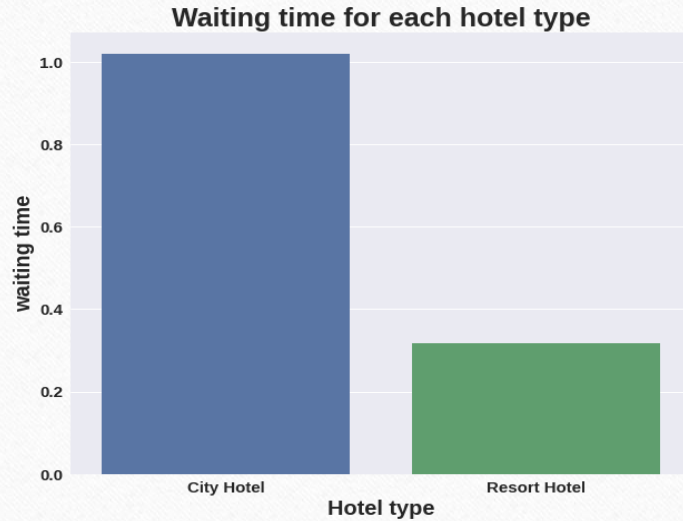
Conclusions:

➢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) :
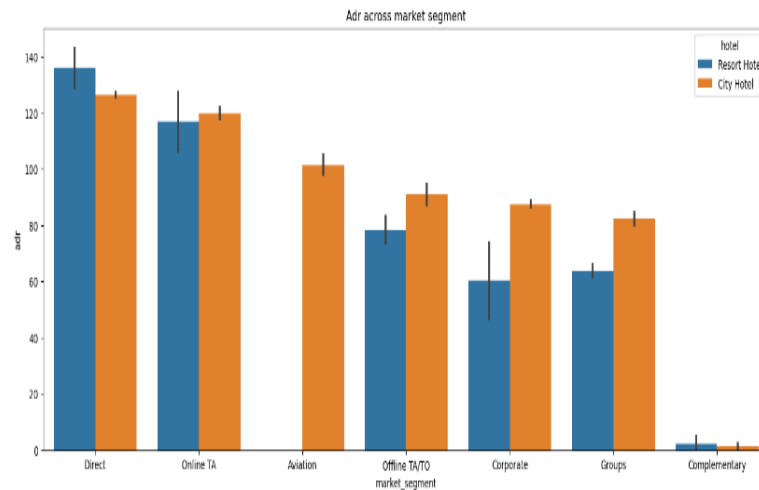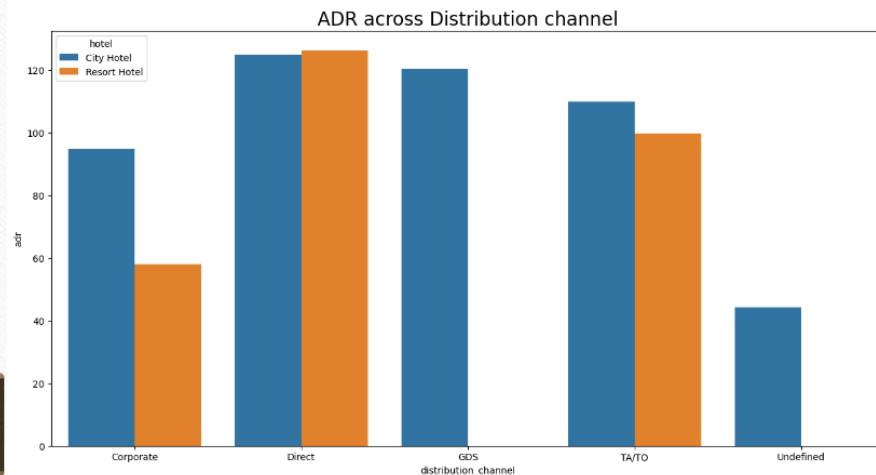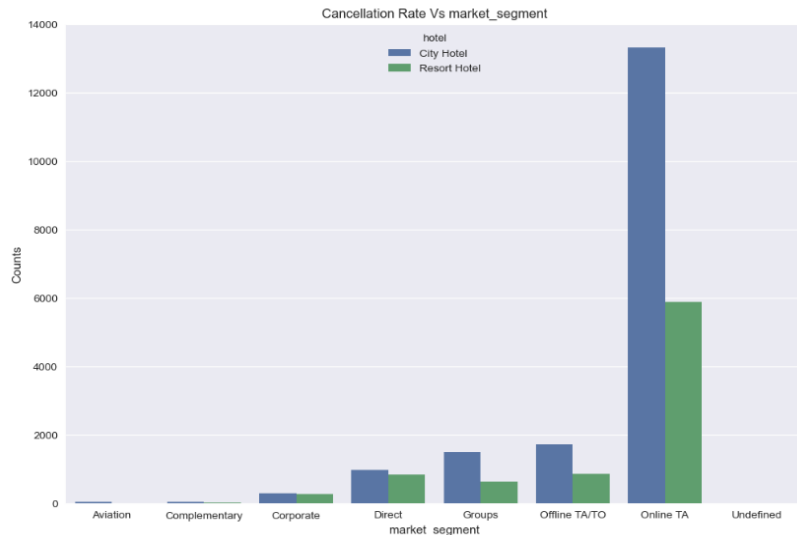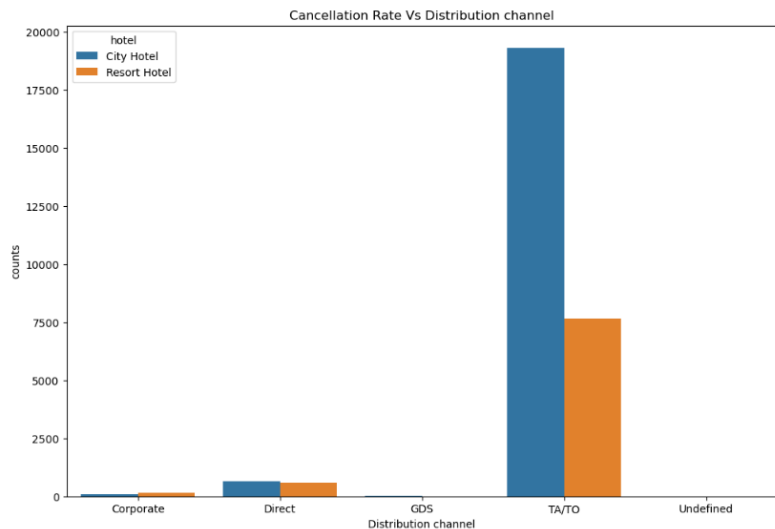

Conclusions:

Distribution channel:

➢ 'Direct' and 'TA/TO' has almost equal adr in both type of hotels which is high among other channels.

➢GDS has high adr in 'City Hotel' type. GDS needs to increase Resort Hotel bookings. From this we can say that "Direct" and 'TA/TO' are generating more revenue than the other channels.

Market Segment:

➢Here "Direct" and 'Online Travel Agency' has high adr for both hotel types. Aviation segment needs to increase Resort hotel bookings.

❖Exploratory Data Analysis (EDA) :
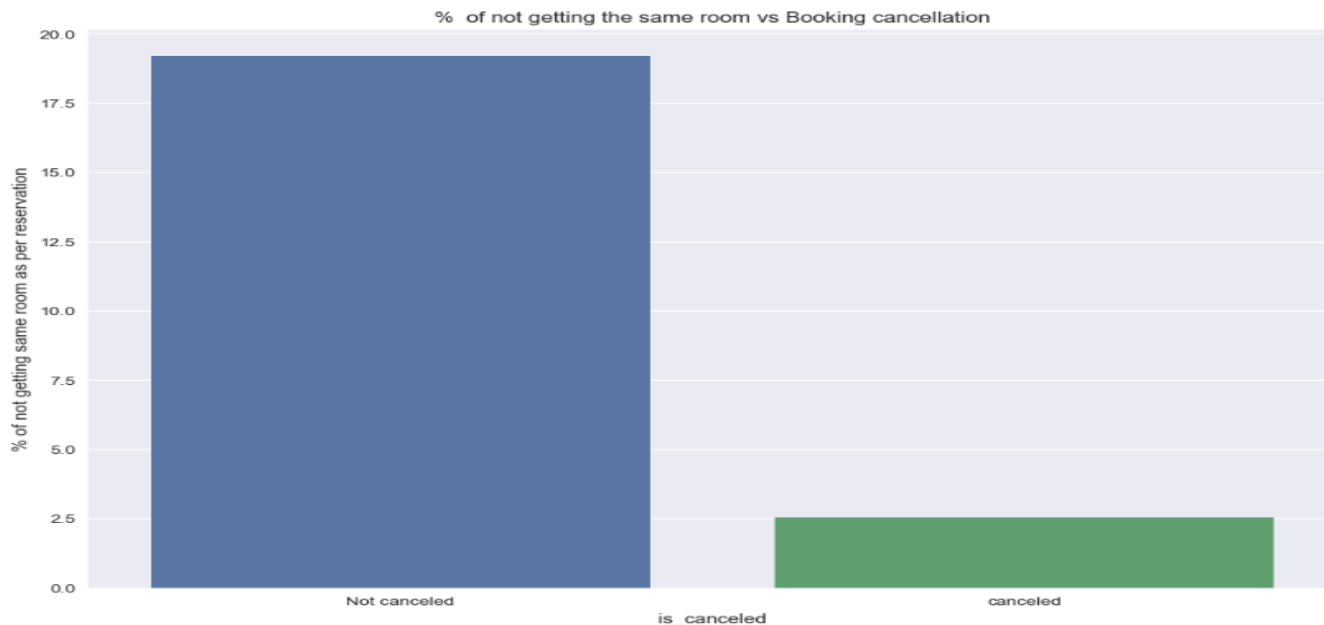

Conclusions:

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.

% of not getting the same room vs Booking cancellation
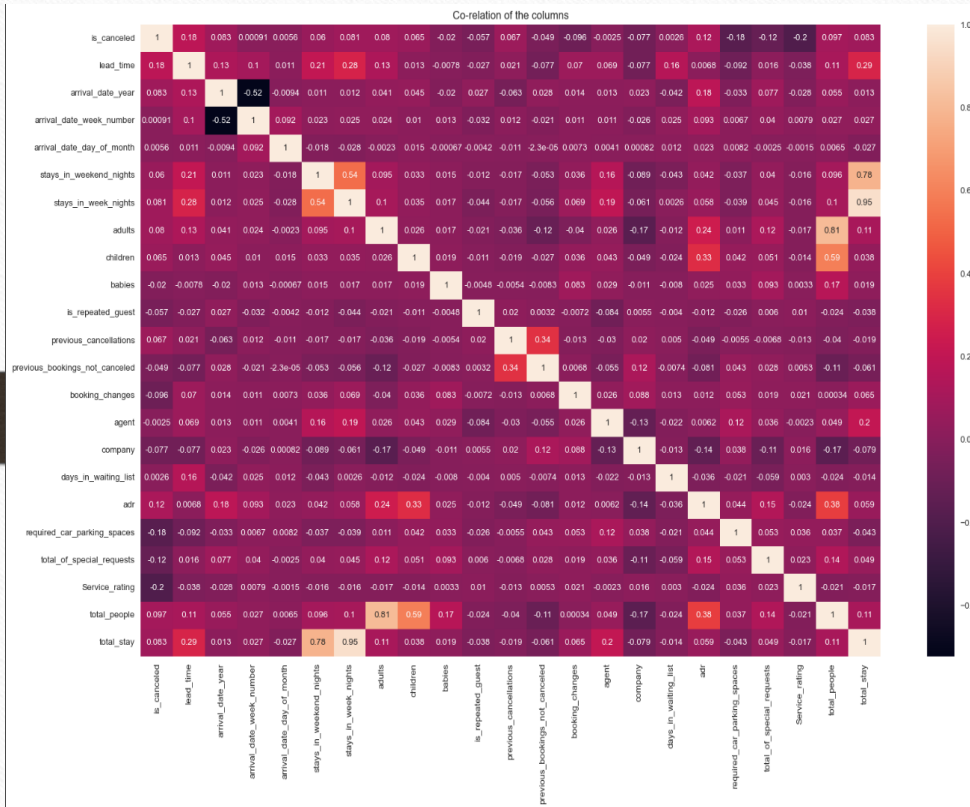
Conclusions:

➤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.

Co-relation of the columns

Conclusions:

➢ is_canceled and same_room_alloted_or_not are negatively corelated. 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.

➢adults, childrens and babies are correlated to each other. That means more the people more will be adr.

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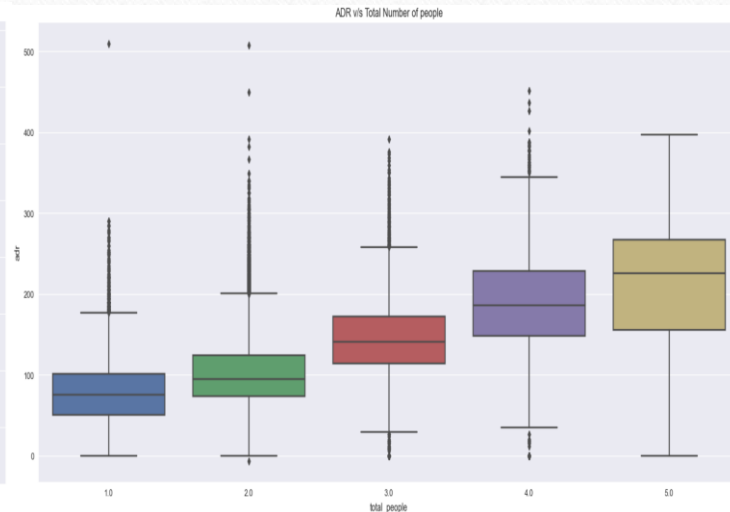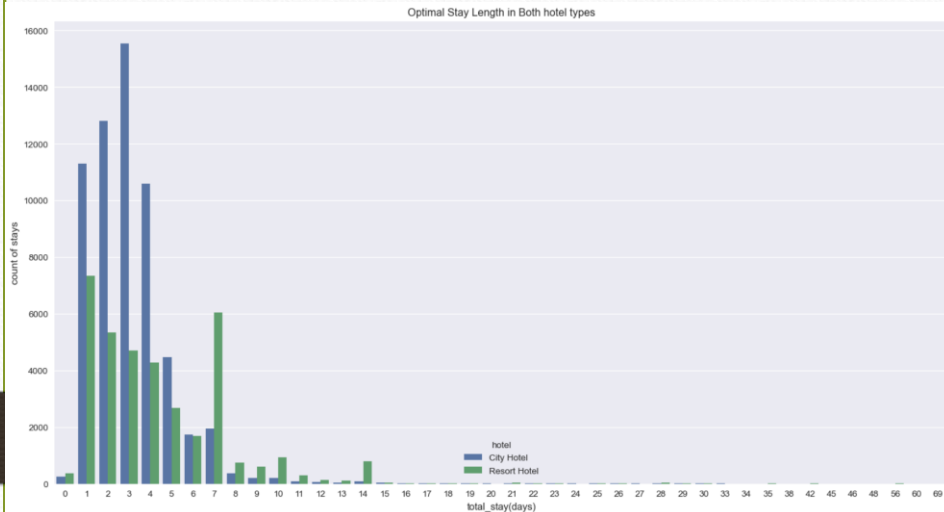
Conclusions:
➢ Optimal stay in both the type hotel is less than 7 days. Usually people stays for a week.
➢For stay more than 7 days people likes to stay in Resort hotels. As we can see after 7 days City Hotel Bookings are very less as compared to Resort hotels.

➢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 its more than 200.
➢ Thus more the people more will revenue of the hotels.

Conclusions

1. City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
2. 27.6 % bookings were got cancelled out of all the bookings.
3. Only 3.6 % people were revisited the hotels. Rest 96.4 % were new guests. Thus retention rate is low.
4. The percentage of 0 changes made in the booking was more than 82.3 %. Percentage of Single changes made was about 12%.
5. Most of the customers (92.3%) do not require car parking spaces.
6. 79.6 % bookings were made through TA/TO (travel agents/Tour operators).
7. BB(Bed & Breakfast) is the most preferred type of meal by the guests.
8. Maximum number of guests were from Portugal, i.e. more than 25000 guests.
9. July and August months had the most Bookings. Summer vaccation can be the reason for the bookings.
10. Most of the bookings for City hotels and Resort hotel were happened in 2016.
11. Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
12. Booking cancellation rate is high for City hotels which almost 30 %.
13. Average lead time for resort hotel is high.
14. Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
15. City hotels have the most repeated guests.
16. Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
17 Almost 18.5 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.

Signing off…

BY

Akash Sahu