Capstone Project 1

EDA On Hotel Booking Analysis

By

Abhishek Kannagi

Almabetter Trainee

➢ # 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 performancecontains.the business and performance.

AI

➢ 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:

AI

❖ After collecting data it's very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let's understand this 32 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 of origin.

➢ Data Collection and Understanding:

market_segment : Market segment designation. (TA/TO)

distribution_channel : Booking distribution channel.(T/A/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:

AI

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

```
[ ] #checking for Null Values
    Hotel_df1.isna().sum().sort_values(ascending=False)[:6]
```

	Columns	Null values
0	company	82137
1	agent	12193
2	country	452
3	children	4
4	reserved_room_type	0
5	assigned_room_type	0

```
# Replacing NA values with "others"
Hotel_df1['country'].fillna('others',inplace=True)
```

```
Hotel_df1.isna().sum().sort_values(ascending=False)[:6].reset_index().rename(columns={'index':'Columns',0:'Null values'})
```

	Columns	Null values
0	hotel	0
1	is_canceled	0
2	reservation_status	0
3	total_of_special_requests	0
4	required_car_parking_spaces	0

❖ Handling Duplicates: Data had 31994 duplicates values. So we dropped it from the

```
[ ]   # checking for the duplicate rows
      Hotel_df1.duplicated().value_counts()      #true means duplicate rows

      False      87396
      True       31994
      dtype:  int64
```

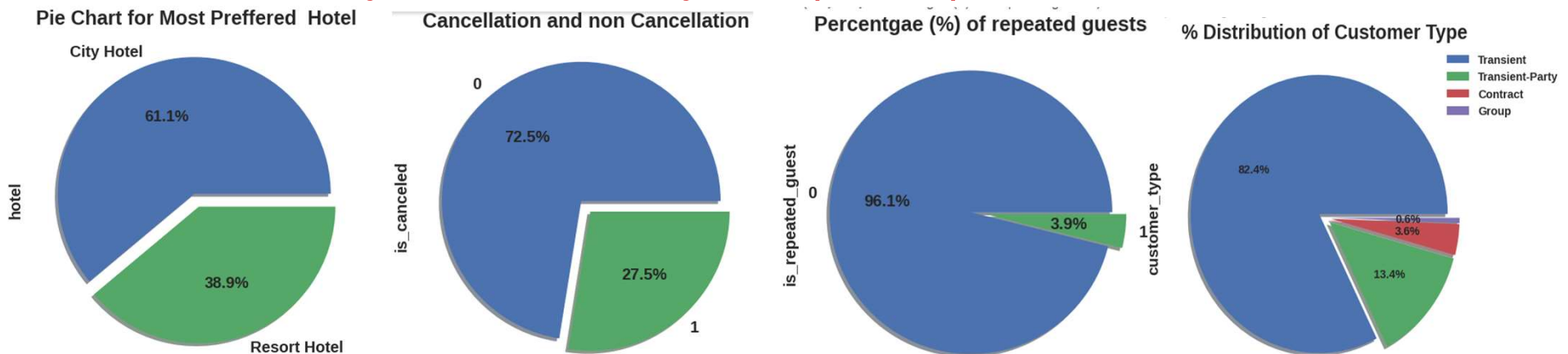❖ Feature Engineering:
We created 2 new columns 1)'Total_People' = from the Children, adults, babies.
                         2) 'Total_stay' = From weekend nights and weekdays night

```
# lets add some new columns

Hotel_df1['total_people'] = Hotel_df1['adults'] + Hotel_df1['babies'] + Hotel_df1['children']
Hotel_df1['total_stay'] = Hotel_df1['stays_in_weekend_nights'] + Hotel_df1['stays_in_week_nights']
```

➢ Exploratory Data Analysis (EDA) :

AI

Pie Chart for Most Preffered Hotel

City Hotel

61.1%

38.9%

Resort Hotel

hotel

Cancellation and non Cancellation

0

72.5%

27.5%

1

is_canceled

Percentgae (%) of repeated guests

0

96.1%

3.9%

1

is_repeated_guest

% Distribution of Customer Type

- Transient
- Transient-Party
- Contract
- Group

82.4%

0.6%
3.6%

13.4%

customer_type

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 got 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) :

AI
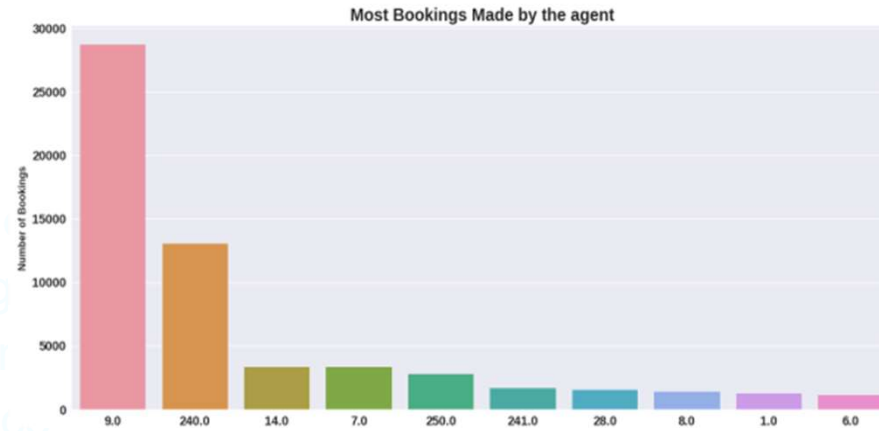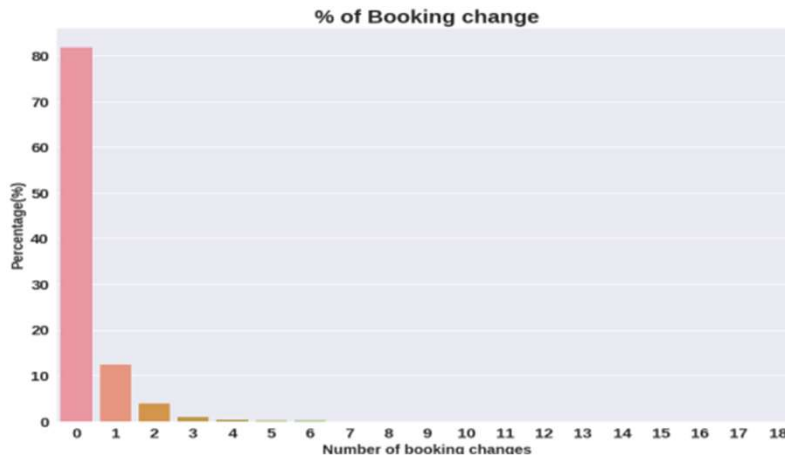
Conclusions:

❖ 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.
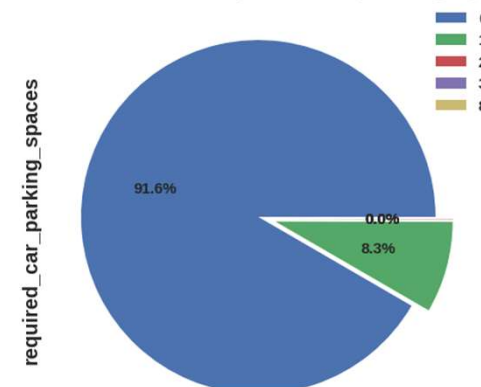
➢ # Exploratory Data Analysis (EDA) :

Mostly Used Distribution Channel for Hotel Bookings

- TA/TO, 79.1%
- Direct, 14.9%
- Corporate, 5.8%
- GDS, 0.2%
- Undefined, 0.0%
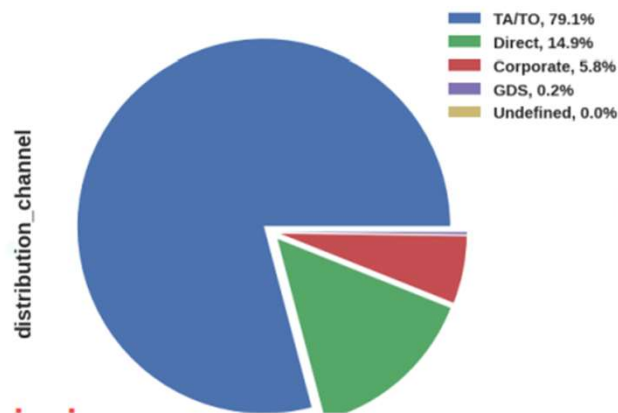
Most preferred Room type

% Distribution of deposit type

- No Deposit
- Non Refund
- Refundable
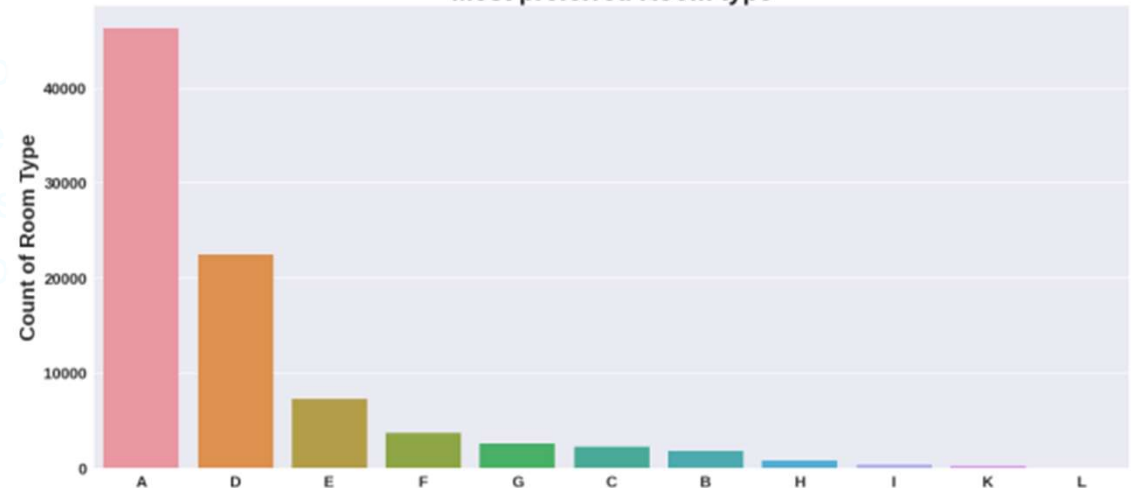
Conclusions:

❖ 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.
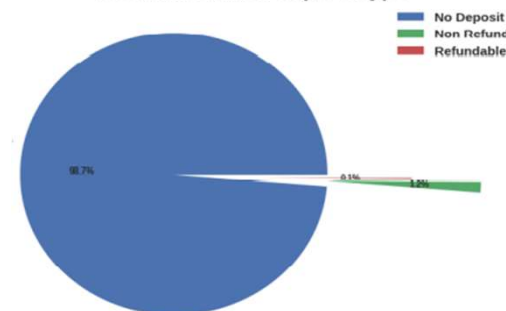
➢ Exploratory Data Analysis (EDA) :

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 preferred.

❖ 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.


Preferred Meal Type

Number of bookings across each month

➢ Exploratory Data Analysis (EDA) :

ADR across each month
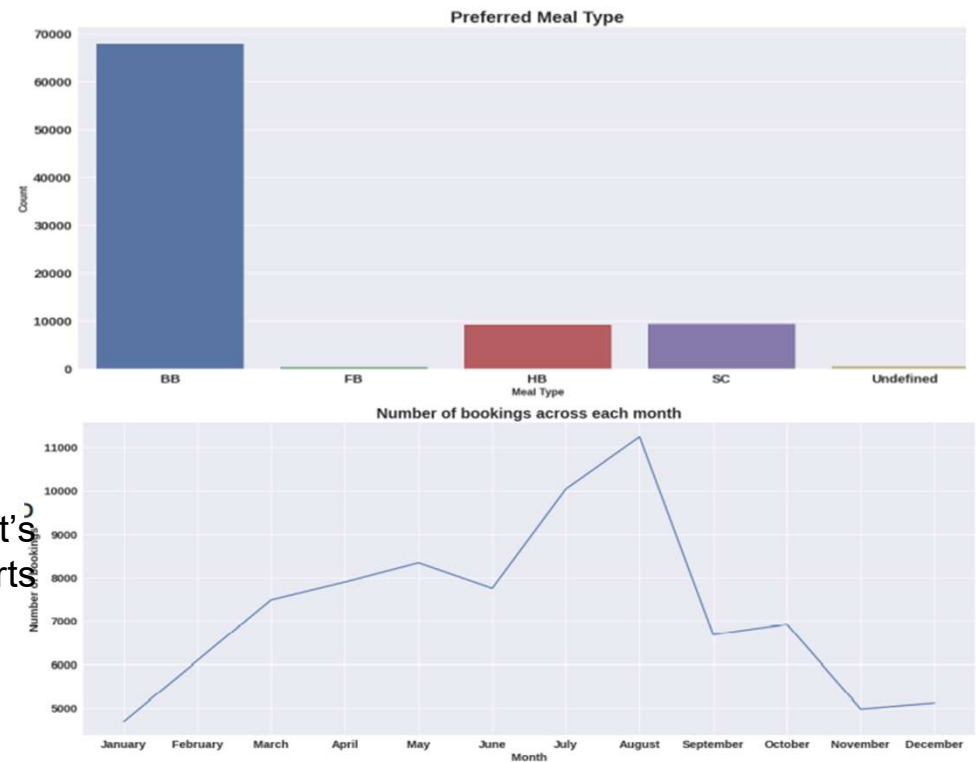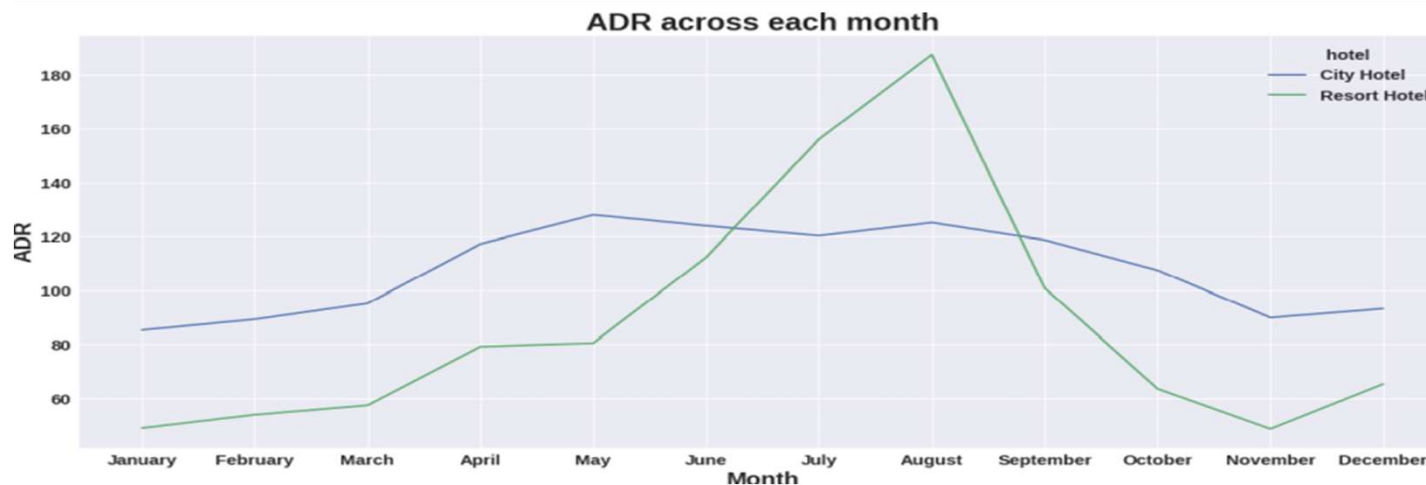
Conclusions:

❖ 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) :
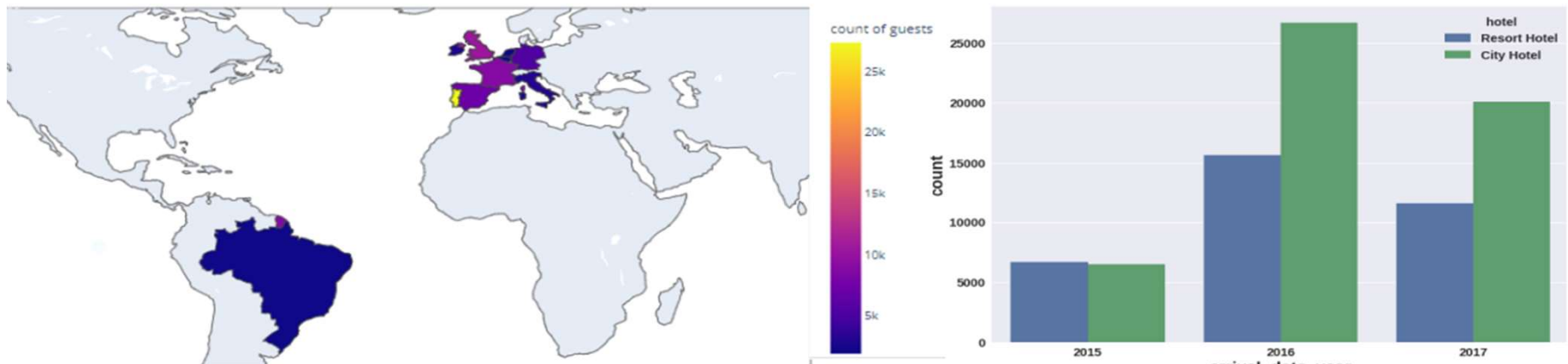
Conclusions:

☐ 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.
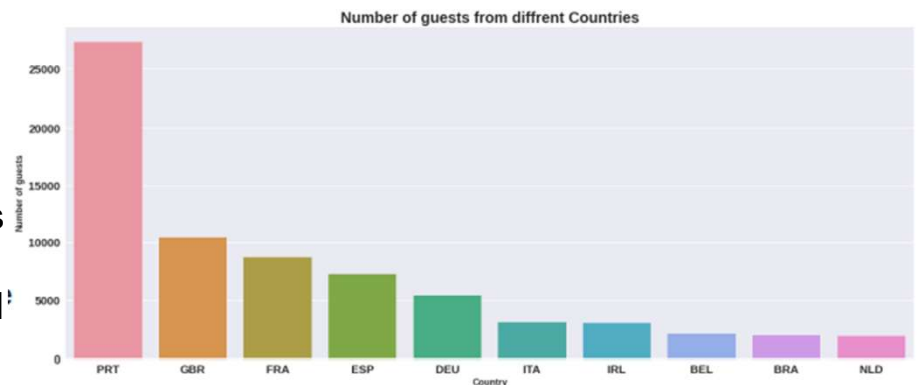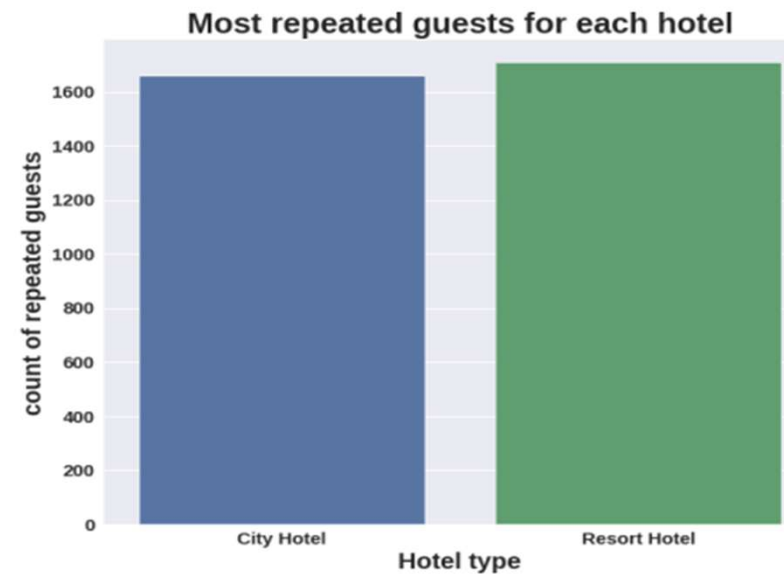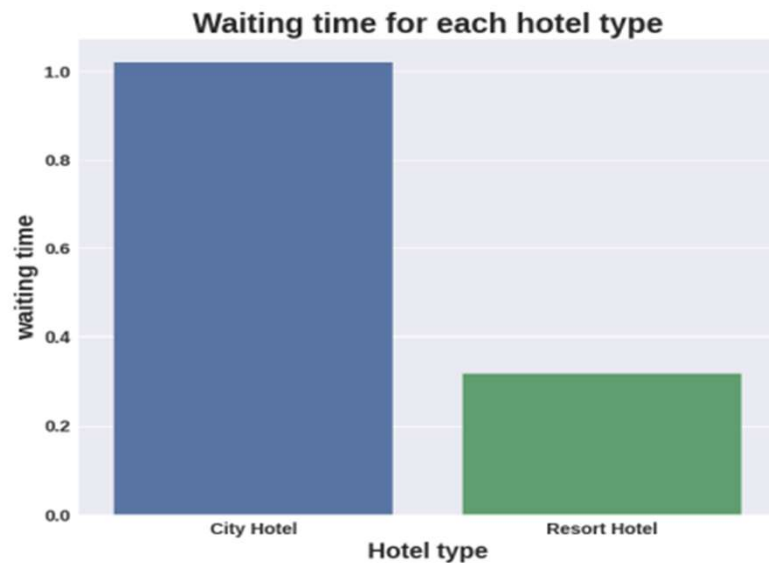
➤ Exploratory Data Analysis (EDA) :
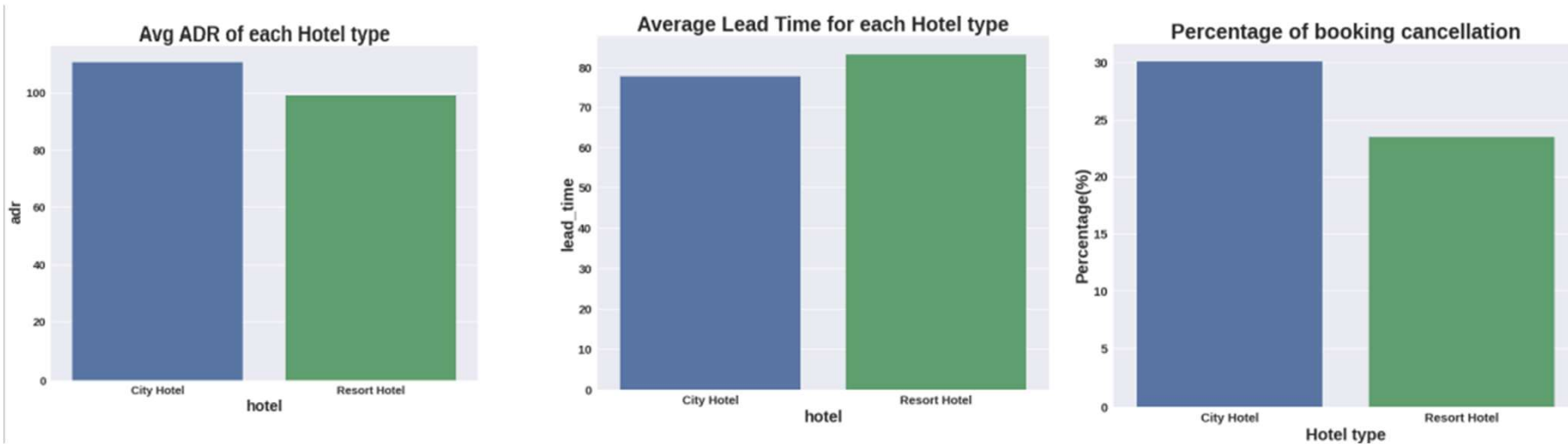
Conclusions:
* Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
* Average lead time for resort hotel is high. It means people plan their trip too early. Usually people prefer resort hotels for longer stays. That's why people plan early
* Booking cancellation rate is high for City hotels which almost 30 %.

➢ Exploratory Data Analysis (EDA) :

AI

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) :
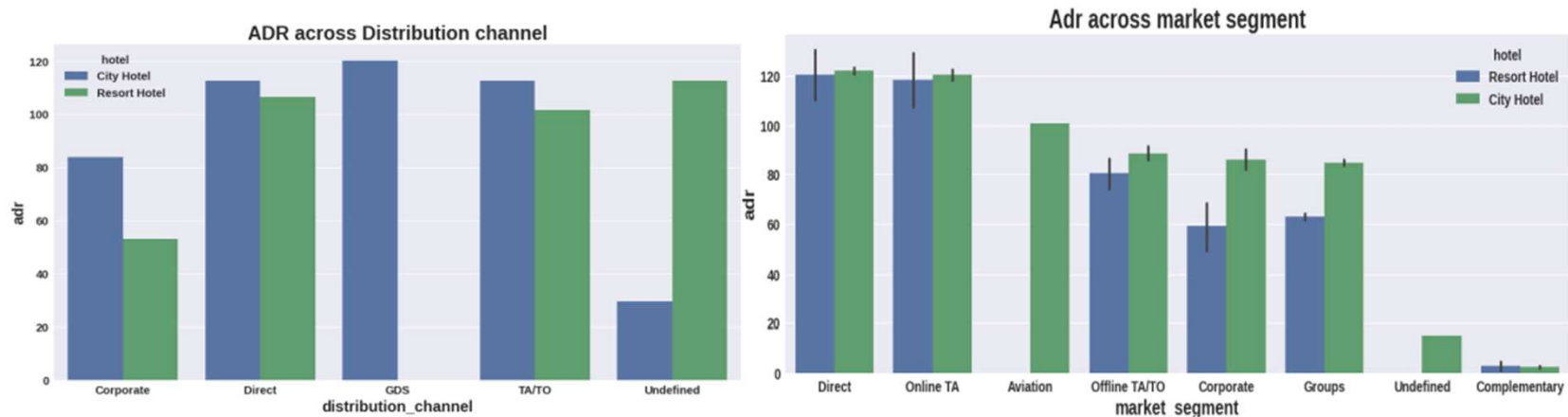
Conclusions:

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) :
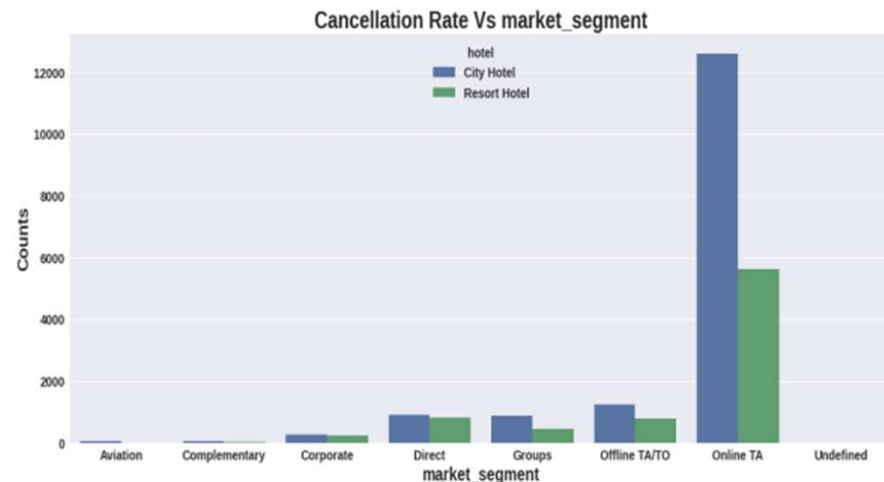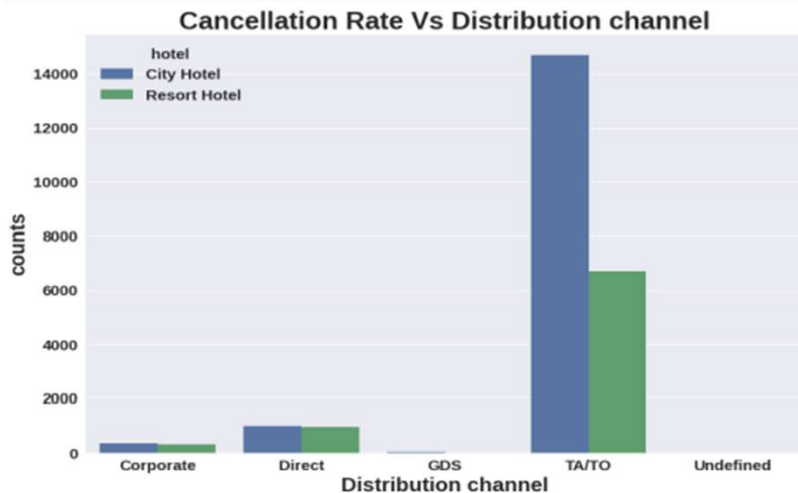
Conclusions:

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.
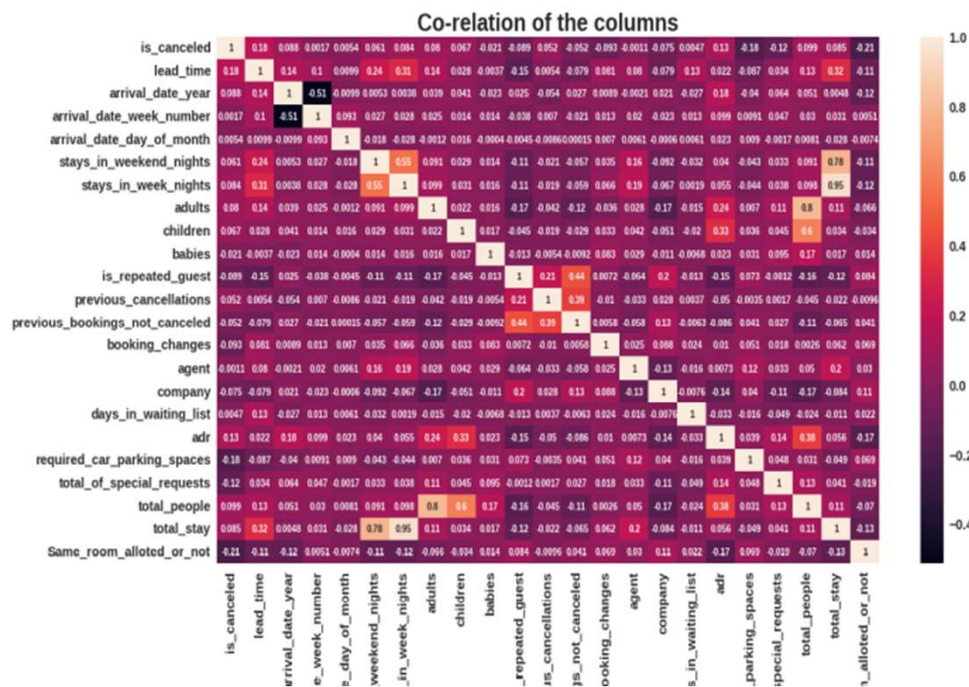
Exploratory Data Analysis (EDA) :

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) :

Co-relation of the columns

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 more is the stay of customer more will be the lead time.

❖ ADR and total people are highly correlated. That means more the people more will be 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)

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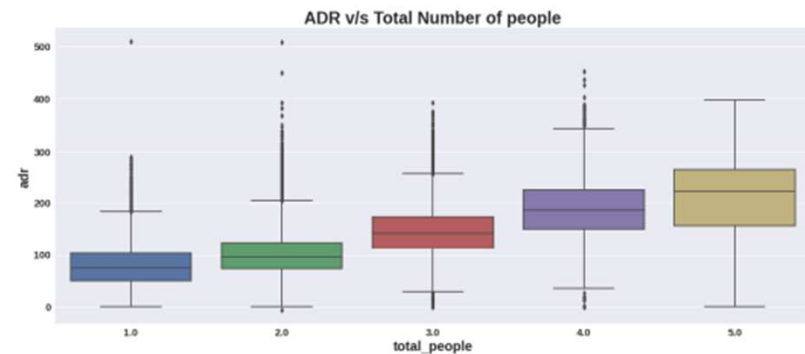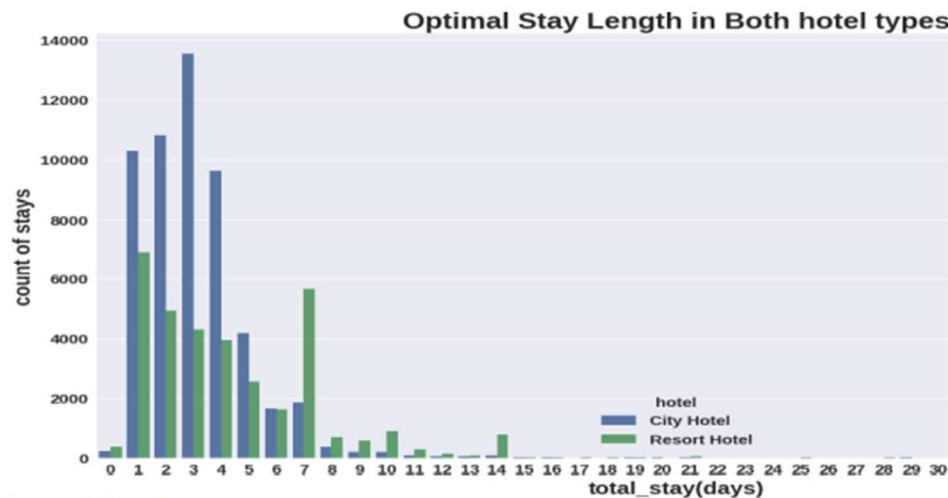
Conclusions:
- ❖ Optimal stay in both the type hotel is less than 7 days. Usually people stays for a week.
- ❖ For stay more than 7 days people likes to stay in Resort hotels. As we can see after 7 days City Hotel Bookings are very less as compared to Resort hotels.

- ❖ 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.

➤ **Challenges**

- (1) There was a lot of duplicate data.
- (2) Data was present in wrong datatype format.
- (3) Choosing appropriate visualization techniques to use was difficult.
- (4) A lot of null values were there in the dataset.

THANK YOU