

Capstone Project-1

EDA On Hotel Booking Analysis

BY

Samim Ali, Mohd. Izhar, Sarath
(Cohort - Florence)



- For this project we will be analyzing Hotel Booking data. This data set contains booking information for a city hotels and a resort hotels, 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.



➤ 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 analysis:** 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.

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:

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

```
#checking for Null Values
```

```
df1.isna().sum().sort_values(ascending=False)[:6]
```

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



```
[19] df1['agent'].fillna(0,inplace=True)
      df1['company'].fillna(0,inplace=True)
      df1['country'].fillna('others',inplace=True)
      df1['children'].fillna(0,inplace=True)
```

```
[20] # Done with missing values
```

```
df1.isna().sum().sort_values(ascending=False)[:6]
```

```
hotel      0
is_canceled 0
reservation_status    0
total_of_special_requests    0
required_car_parking_spaces    0
adr      0
dtype: int64
```

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

```
[13] # checking for the duplicate rows
      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.

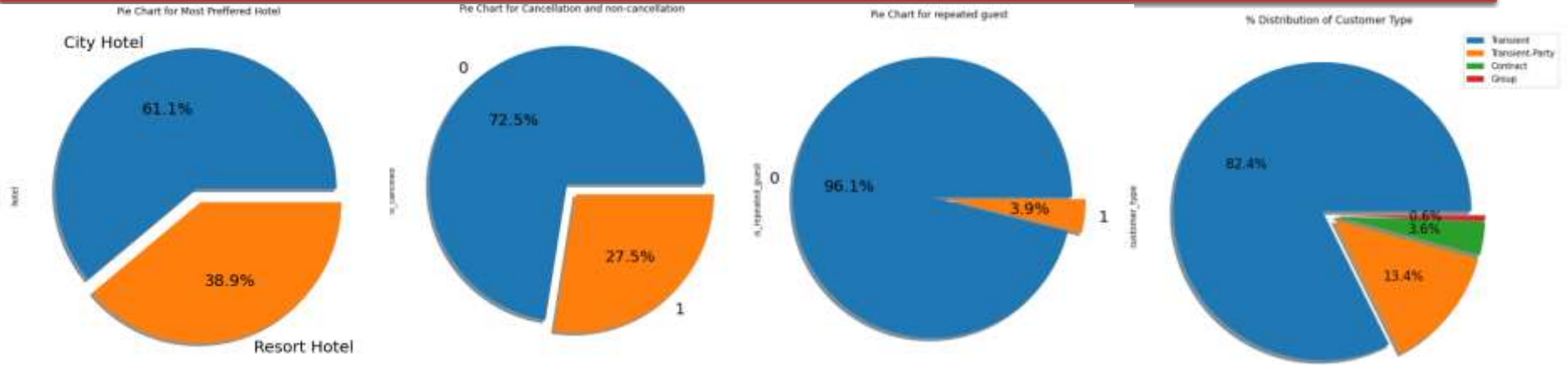
2) 'Total_stay' = From weekend nights and weekdays night

```
[25] # lets add some new columns
```

```
df1['total_people'] = df1['adults'] + df1['babies'] + df1['children']
df1['total_stay'] = df1['stays in weekend nights'] + df1['stays in week nights']
```

❖ Exploratory Data Analysis (EDA) :

AI



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

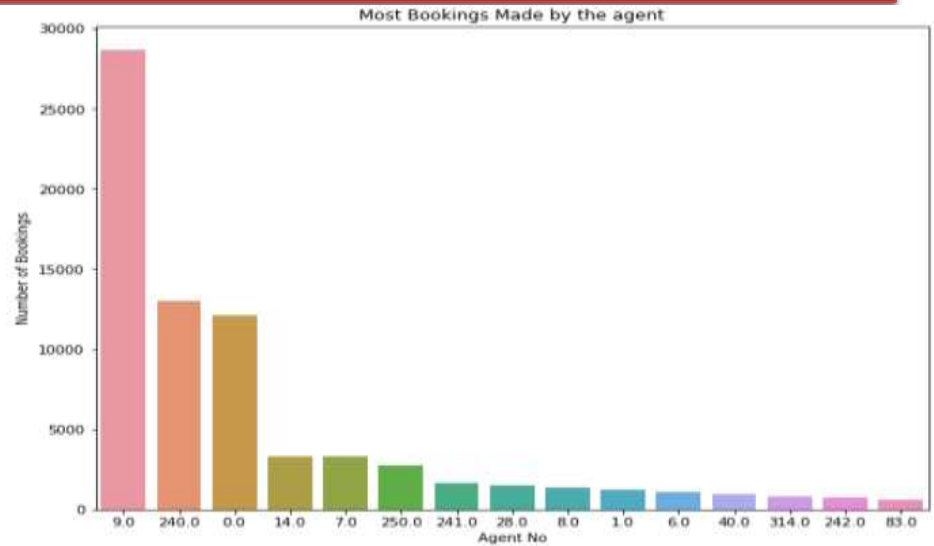
Group -when the booking is associated to a group

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) :

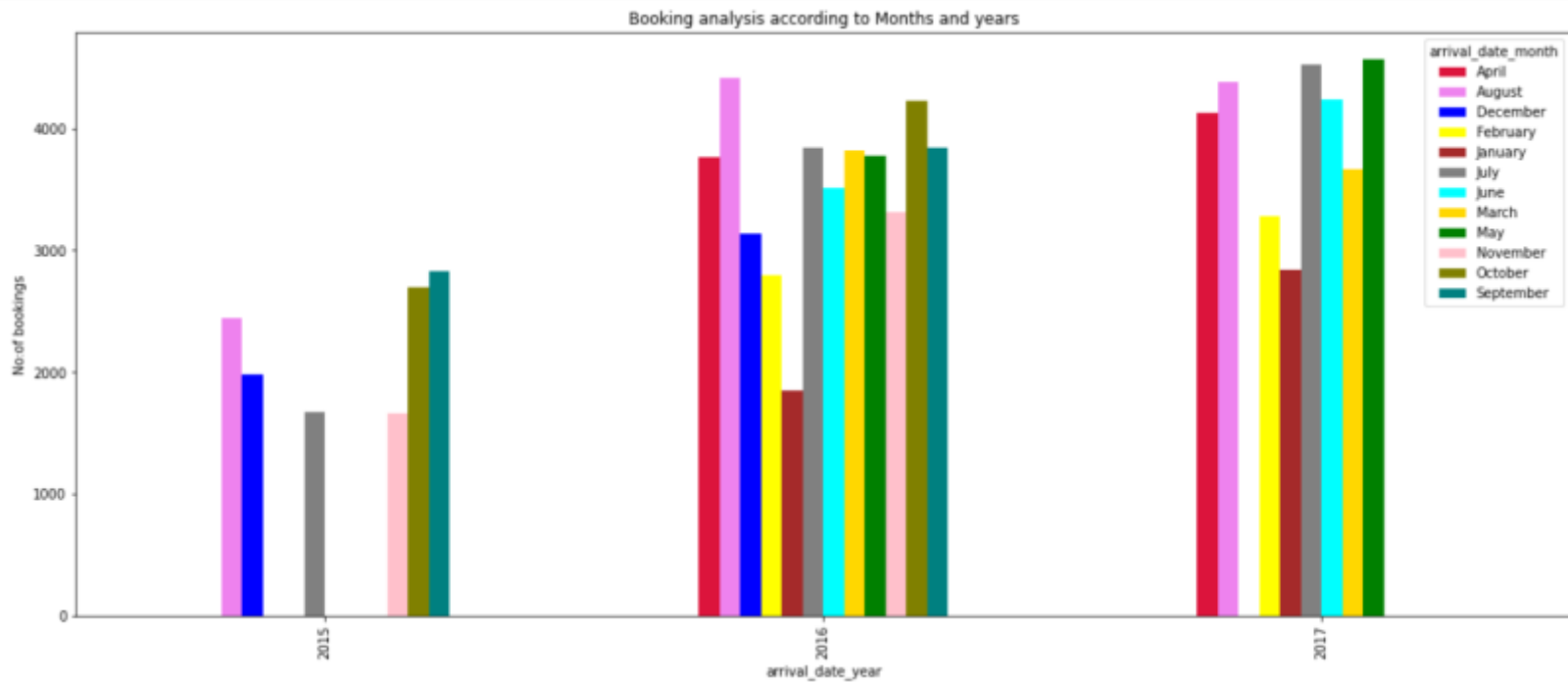
AI



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.

❖ Exploratory Data Analysis (EDA):

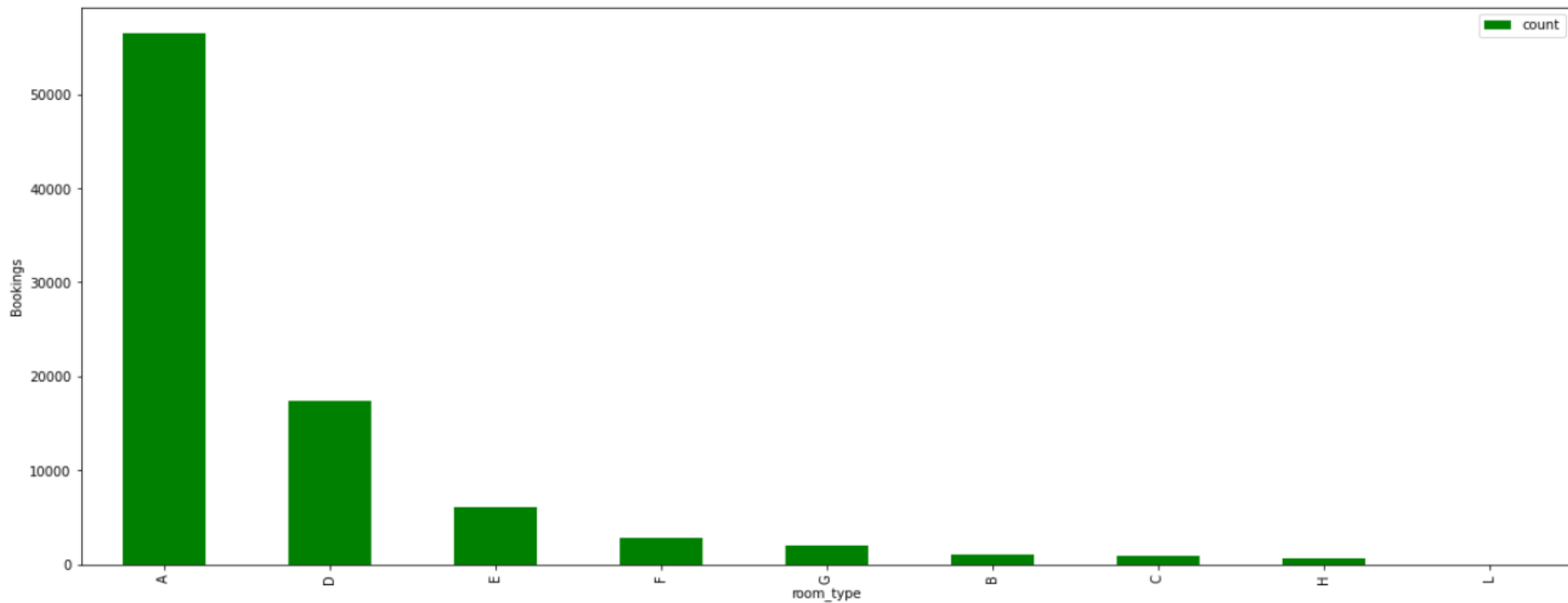


Conclusions:

- 1. August and July are the only two months in the given three years where there always has been bookings.
- 2. 2016 saw the largest number of hotel bookings among the given three years.

❖ Exploratory Data Analysis (EDA) :

AI



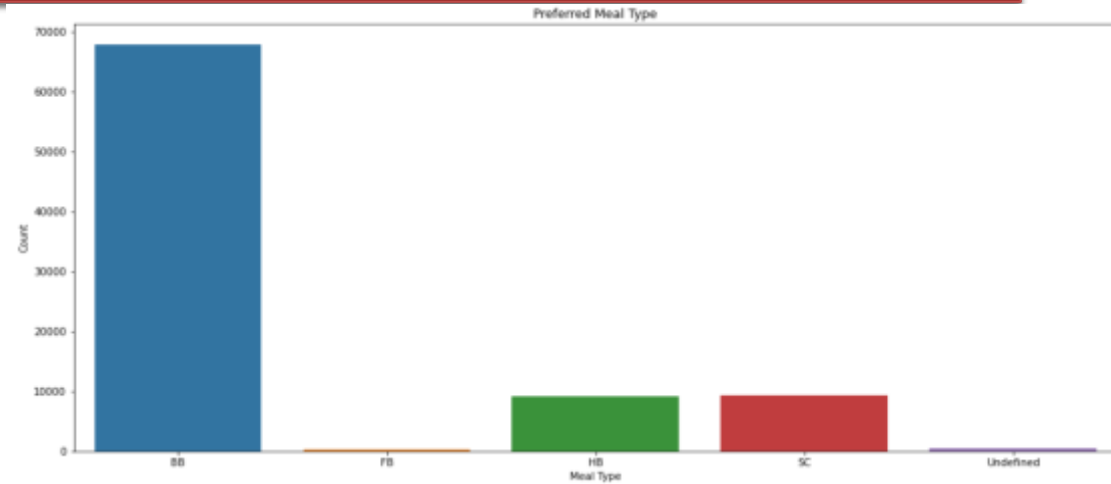
Conclusions:

➤ Room type 'A' is most preferred by the guests second most preferred is 'D'.

❖ Exploratory Data Analysis (EDA) : AI

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.

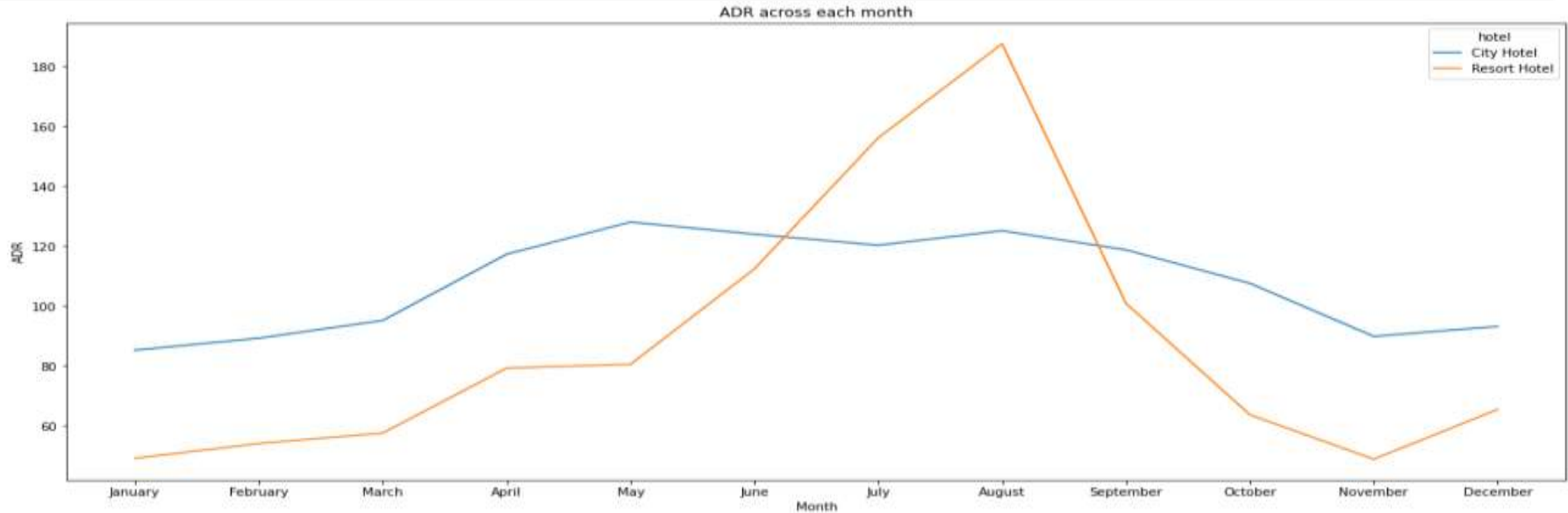


- As we can see in the line chart, from June to September most of the bookings happened. In July and August had the most Bookings because of Summer vacations.



❖ Exploratory Data Analysis (EDA) :

AI

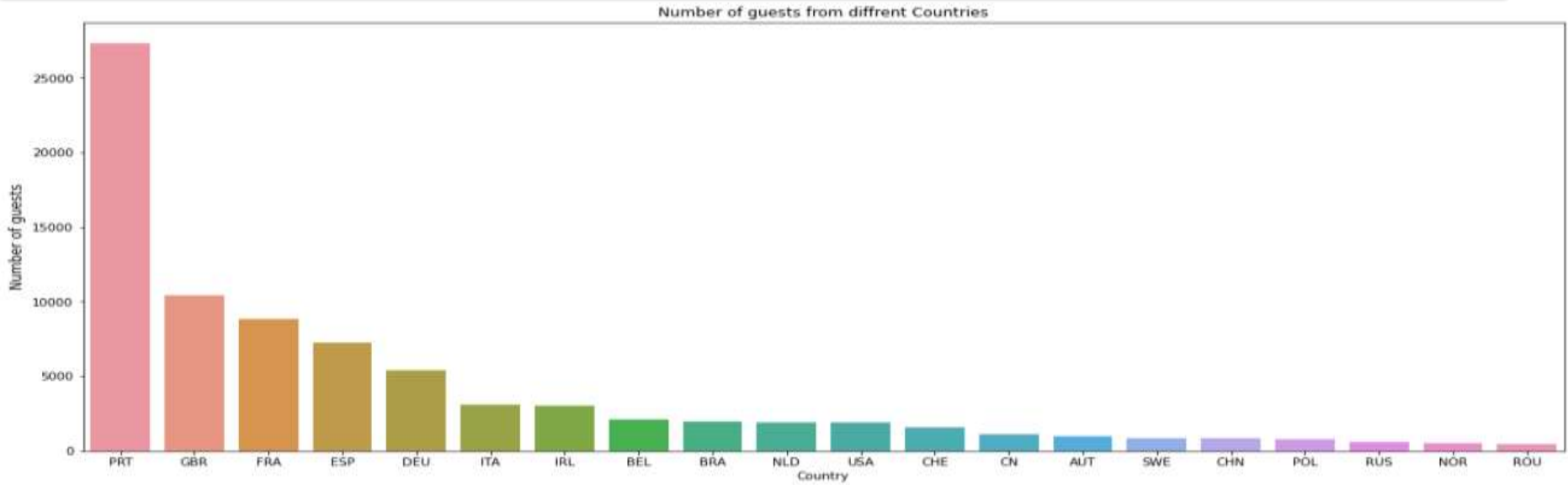


Conclusions:

- 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) :

AI

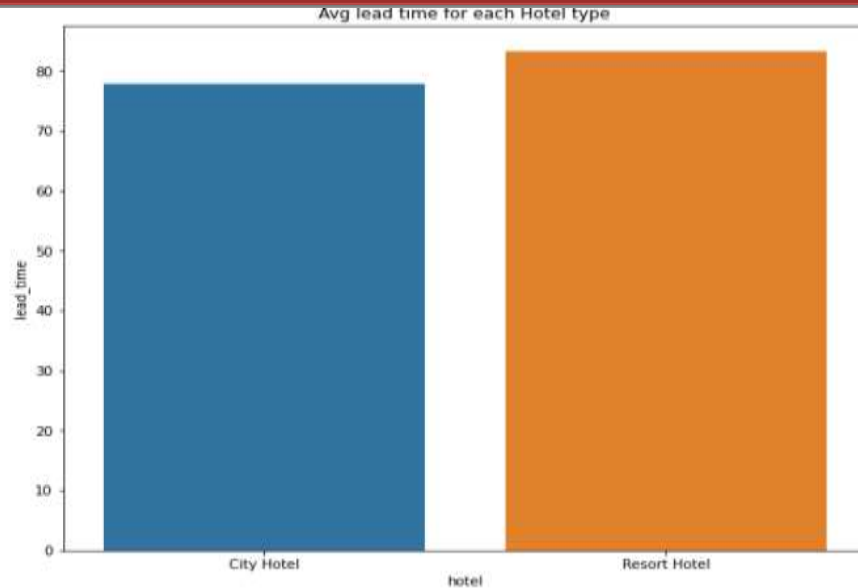
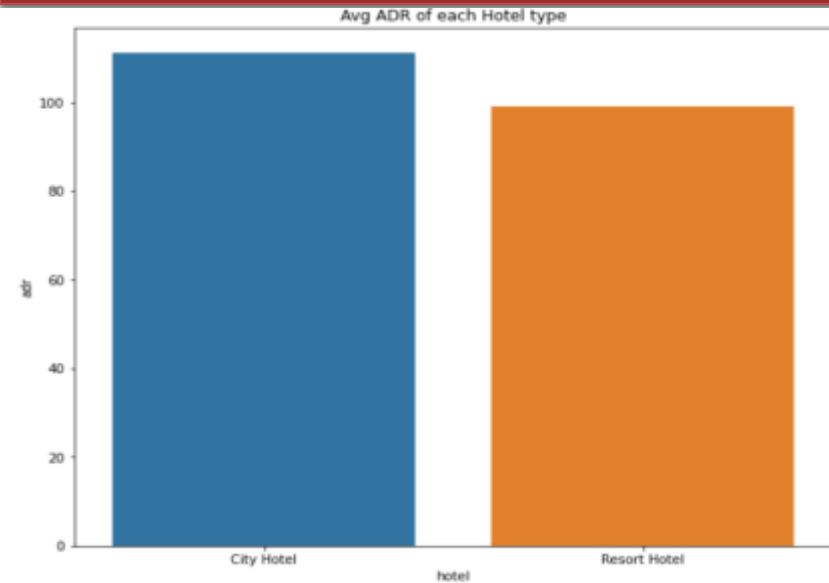


Conclusions:

- Maximum number of guests were from Portugal.
i.e. more than 25000 guests.
- After Portugal, GBR(Great Britain), France and Spain are the countries from where most of the guests came.

❖ Exploratory Data Analysis (EDA) :

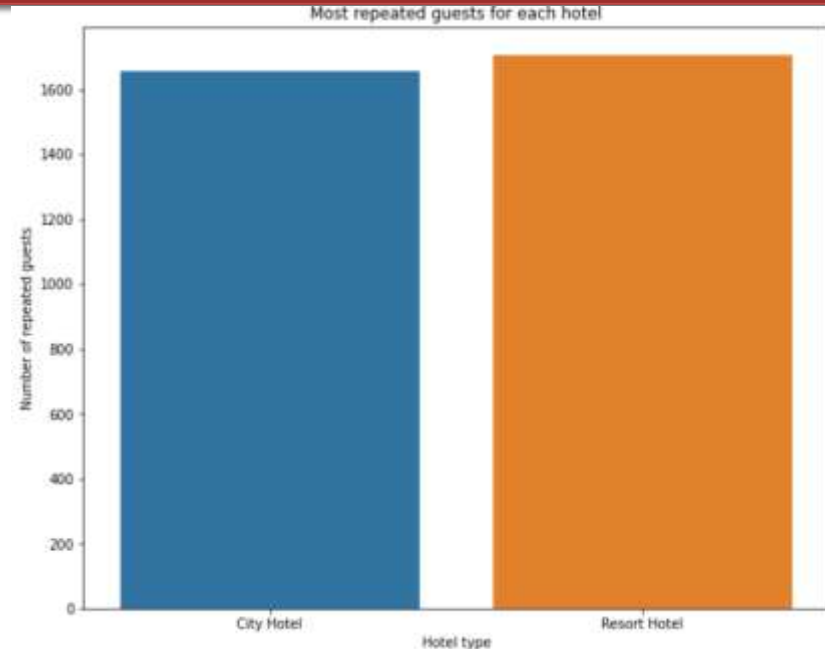
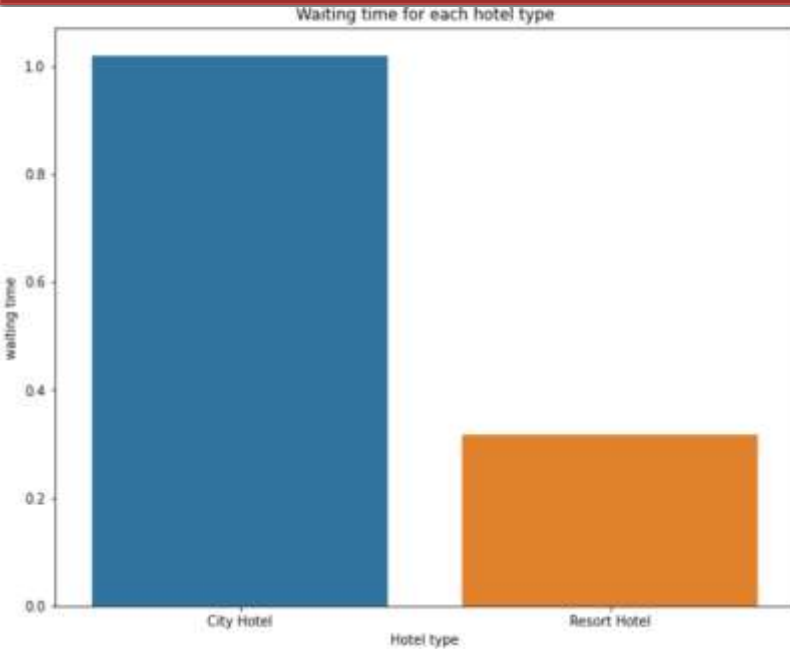
AI



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

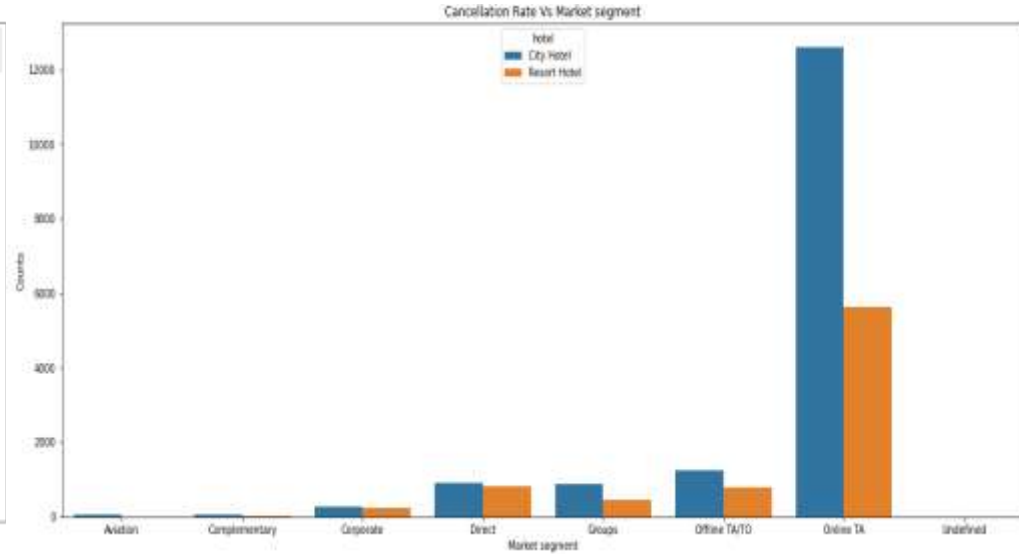
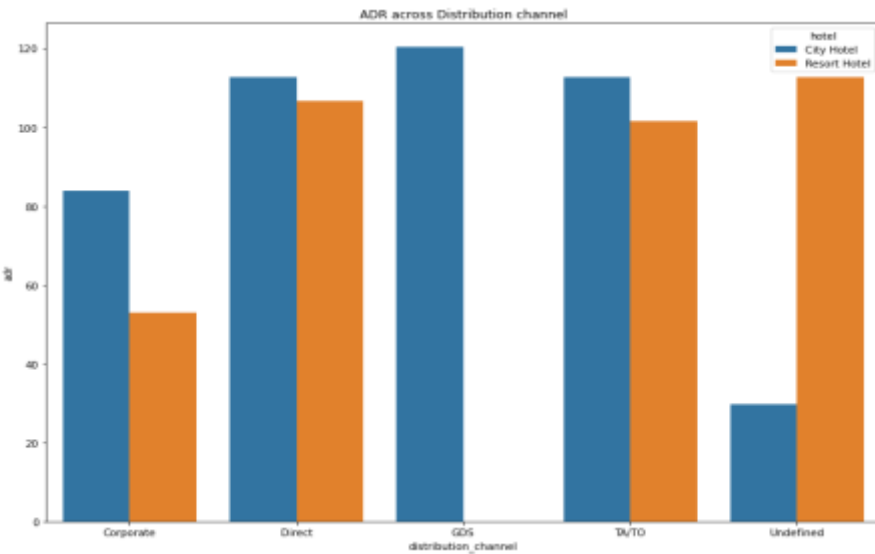
❖ Exploratory Data Analysis (EDA) : AI



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) : AI



Conclusions:

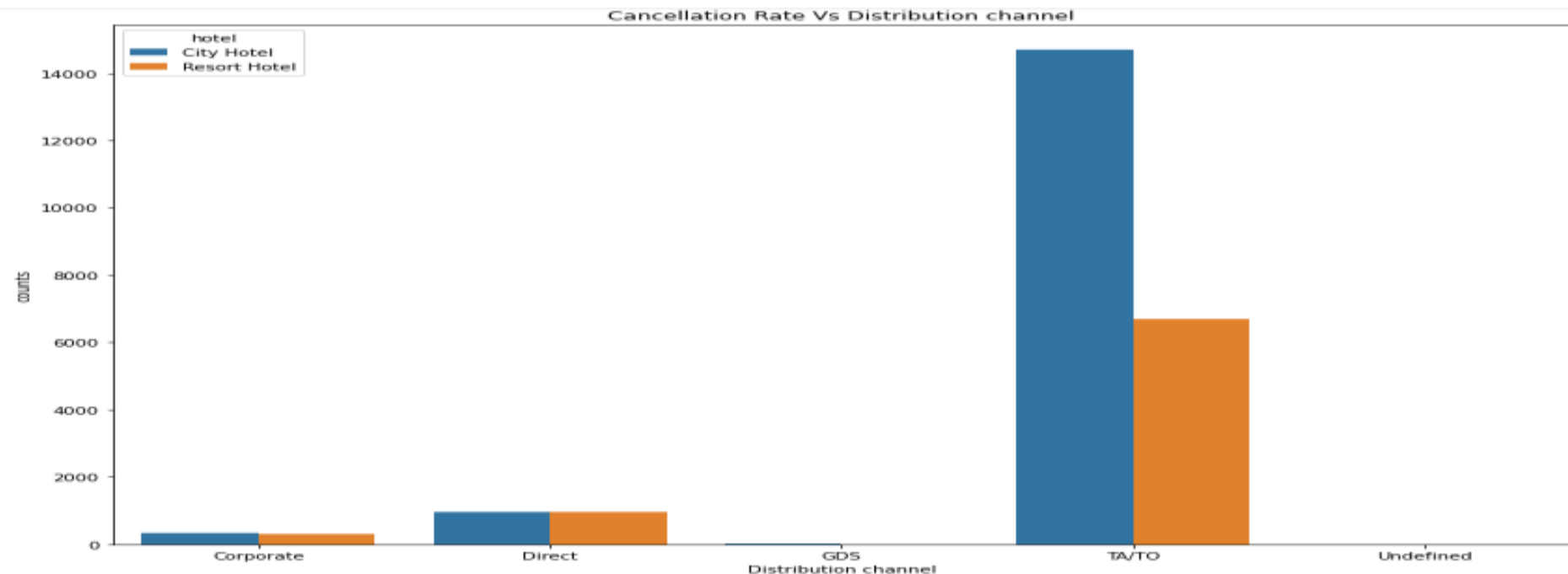
Distribution channel:

- '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:

- Online T/A' has the highest cancellation in both type of Hotels
- In order to reduce the booking cancellations, hotels need to set the refundable/non refundable and deposit policies.

❖ Exploratory Data Analysis (EDA) : AI



Conclusions:

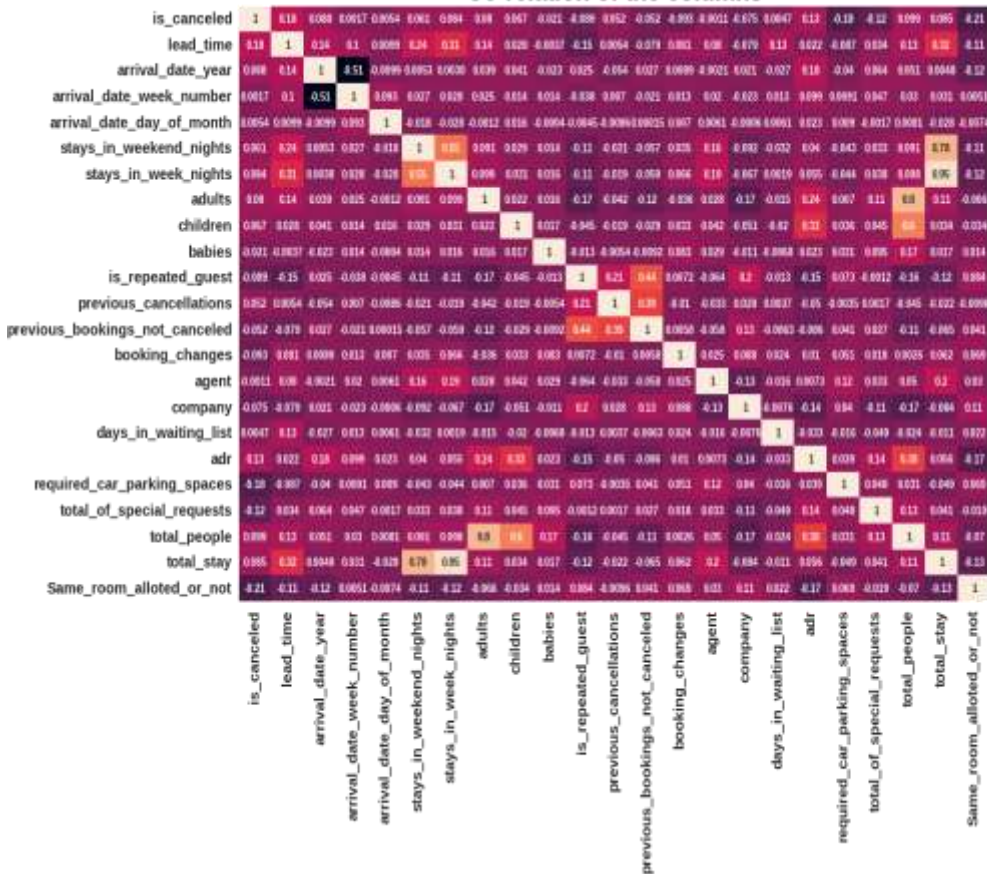
Distribution channel:

➤ '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.

❖ Exploratory Data Analysis (EDA) :

AI

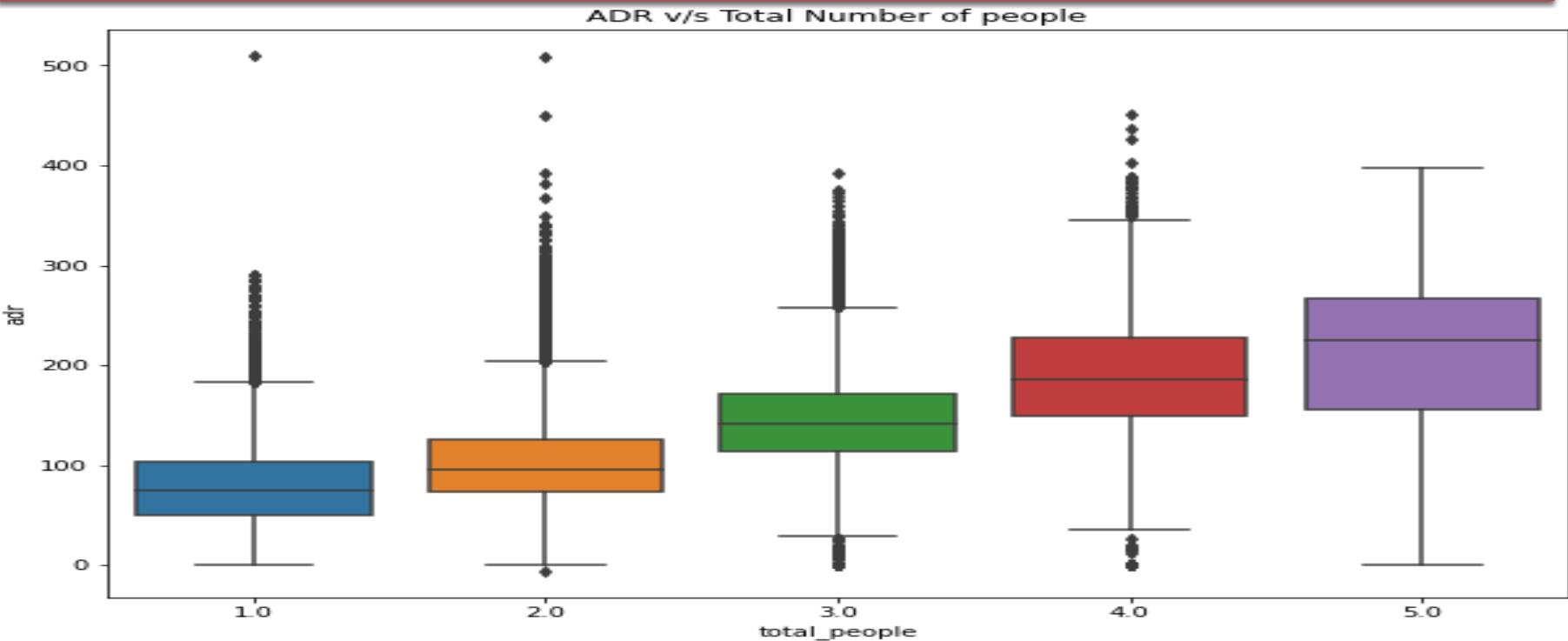
Co-relation of the columns



Conclusions:

- 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) : AI

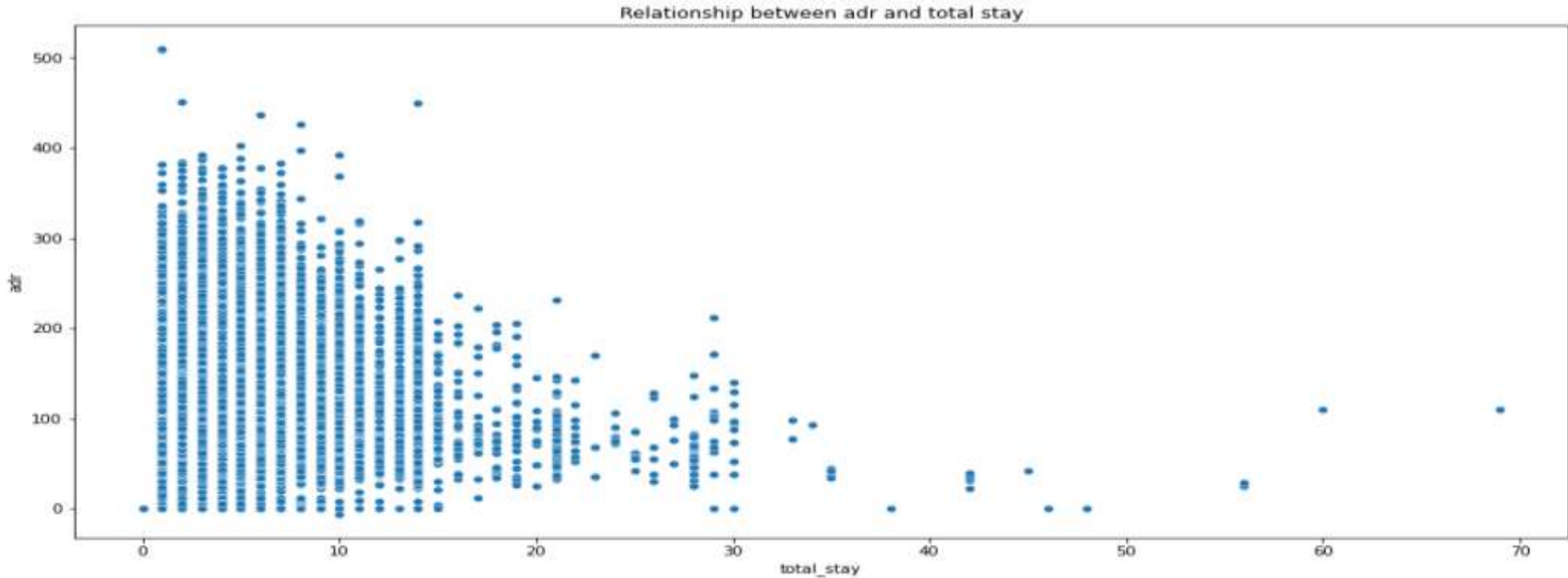


Conclusions:

- 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.

❖ Exploratory Data Analysis (EDA) :

AI



Conclusions:

- We can see that as length of total stay or duration of stay increases the ADR decreases. This means for longer stay, customer has a chance of finalizing a better deal.



THANK YOU