Capstone Project-1

EDA On Hotel Booking Analysis

BY

Tushar Kuril

Index :

AI

- Understanding the Problem
- Flow of Project
- Data Summary
- Data Wrangling
- EDA (Exploratory Data Analysis)

Understanding the Problem

AI

Item 1

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.

Item 2

Hotel industry is a very volatile industry and the bookings depends on different factors.

Item 3

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.

- Flow

AI


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

Data Summary

Given data set has different columns of variables crucial for hotel bookings. Some of them are :

- **hotel**: The category of hotels, which are two resort hotel and city hotel.
- **is_cancelled** : The value of column show the cancellation type. If the booking was cancelled or not. Values[0,1], where 0 indicates not cancelled.
- **lead_time** : The time between reservation and actual arrival.
- **stayed_in_weekend_nights**: The number of weekend nights stay per reservation
- **meal**: Meal preferences per reservation.[BB,FB,HB,SC,Undefined]
- **Country**: The origin country of guest.

Data Summary

- **market_segment**: This column show how reservation was made and what is the purpose of reservation. Eg, corporate means corporate trip, TA for travel agency.
- **distribution_channel**: The medium through booking was made.[Direct,Corporate,TA/TO,undefined,GDS.]
- **Is_repeated_guest**: Shows if the guest is who has arrived earlier or not.Values[0,1]-->0 indicates no and 1 indicated yes person is repeated guest.
- **days_in_waiting_list**: Number of days between actual booking and transact.
- **customer_type**: Type of customers(Transient, group, etc.)

• Data Wrangling :

AI

• Handling Missing Values :

➔ There were 4 columns company, agent, country and children with missing values.
➔ We had filled this null values with zeros.

	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

	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
5	adr	0

• Data Wrangling :

AI

• Handling Duplicates :

➔ Data contains 31994 duplicate values, so we had dropped it.

```
#Checking for duplicates in row
df.duplicated().value_counts()                    #True means duplicaated rows
```

```
False    87396
True     31994
dtype: int64
```

• Featured Engineering :

➔ We created 2 new columns,

1. ‘Total_People’ = from the Children, adults, babies.
2. ‘Total_stay’ = From weekend nights and weekdays night

• Exploratory Data Analysis (EDA) :

AI

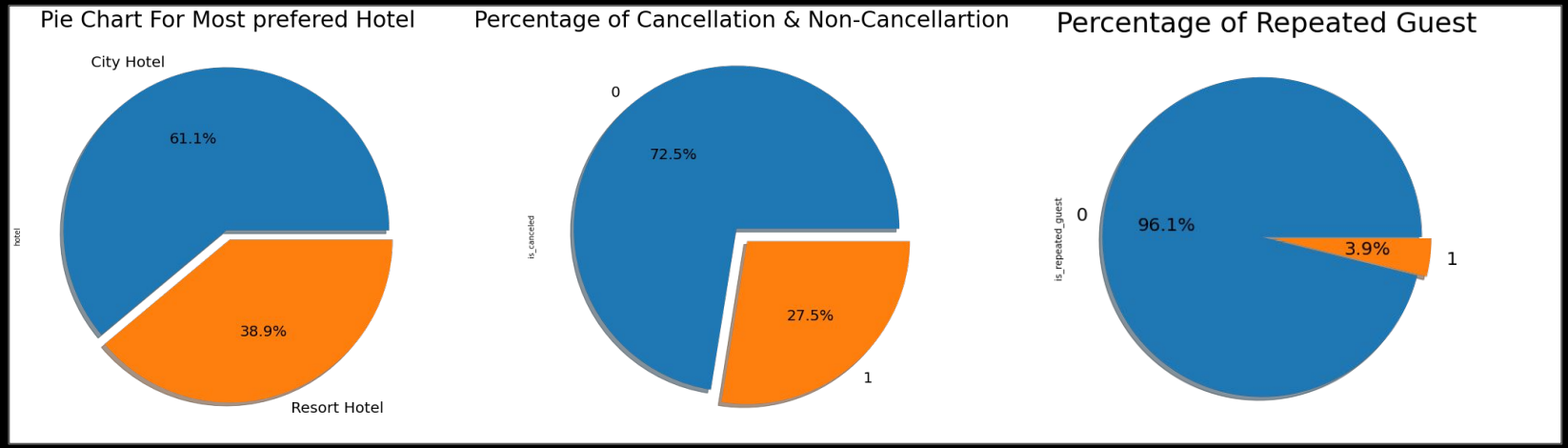
EDA will be divided into following 3 analysis.

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.

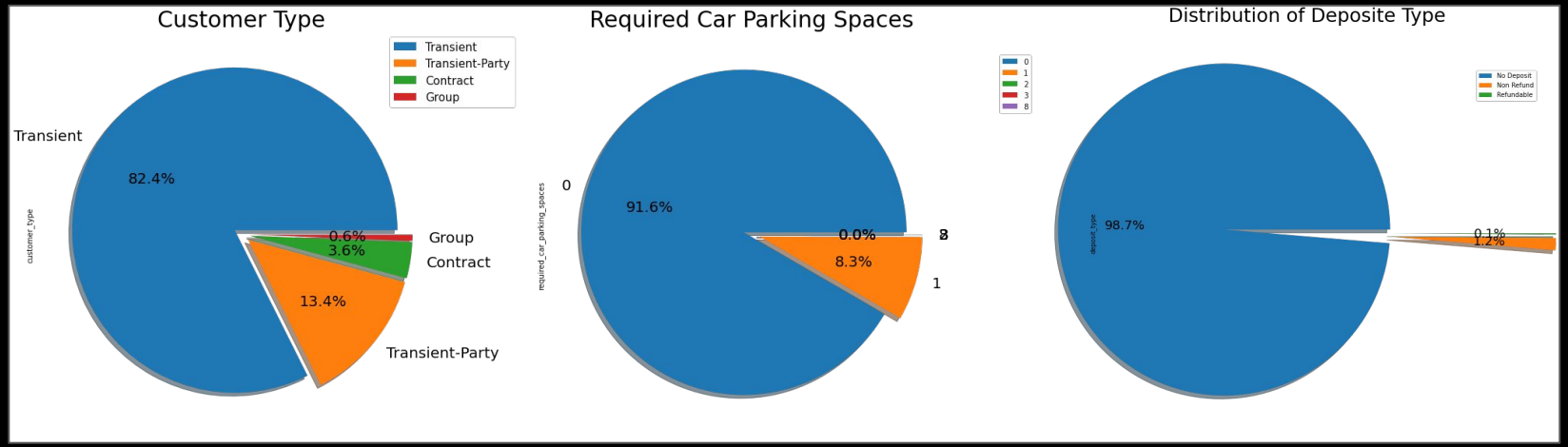
• Univariate Analysis:



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.

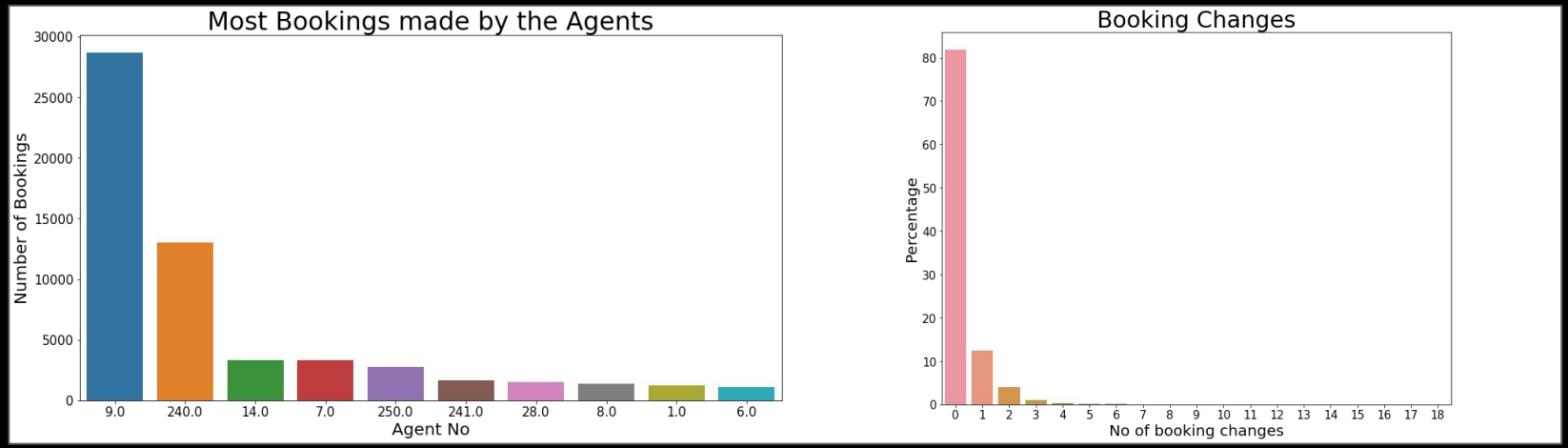
• Univariate Analysis:



Conclusions:

- ➔ 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.
- ➔ Most of the customers(91.6%) do not require car parking spaces. Only 8.3 % people required only 1 car parking space.
- ➔ Almost 98.7% of the guests prefer ‘No deposit’ type of criterion while booking hotels.

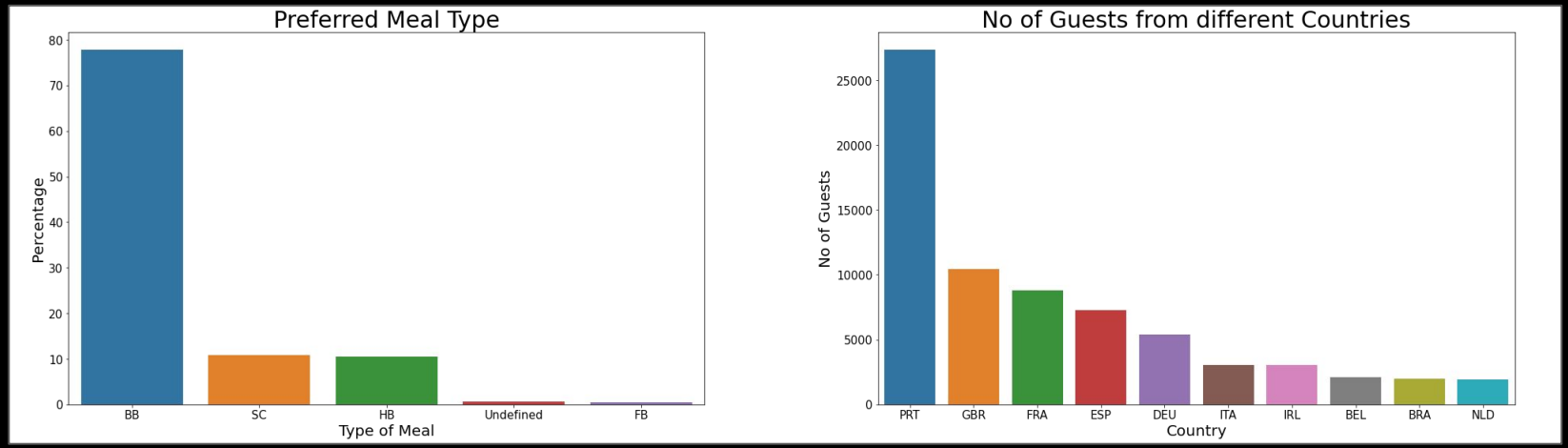
• Univariate Analysis:



Conclusions:

- ➔ Agent Id no -9 made the highest bookings which is more than 28721.
- ➔ The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.

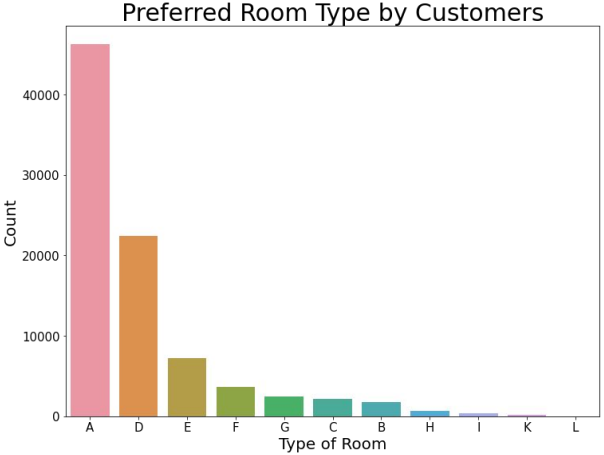
• Univariate Analysis:



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.
- ➔ 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

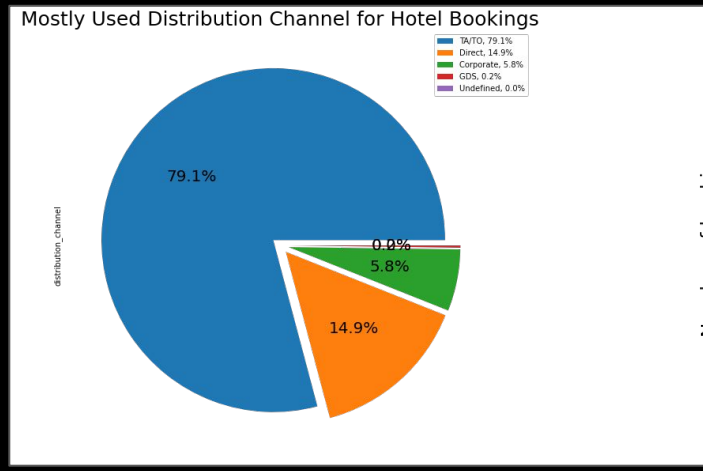
• Univariate Analysis:



Conclusions:

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

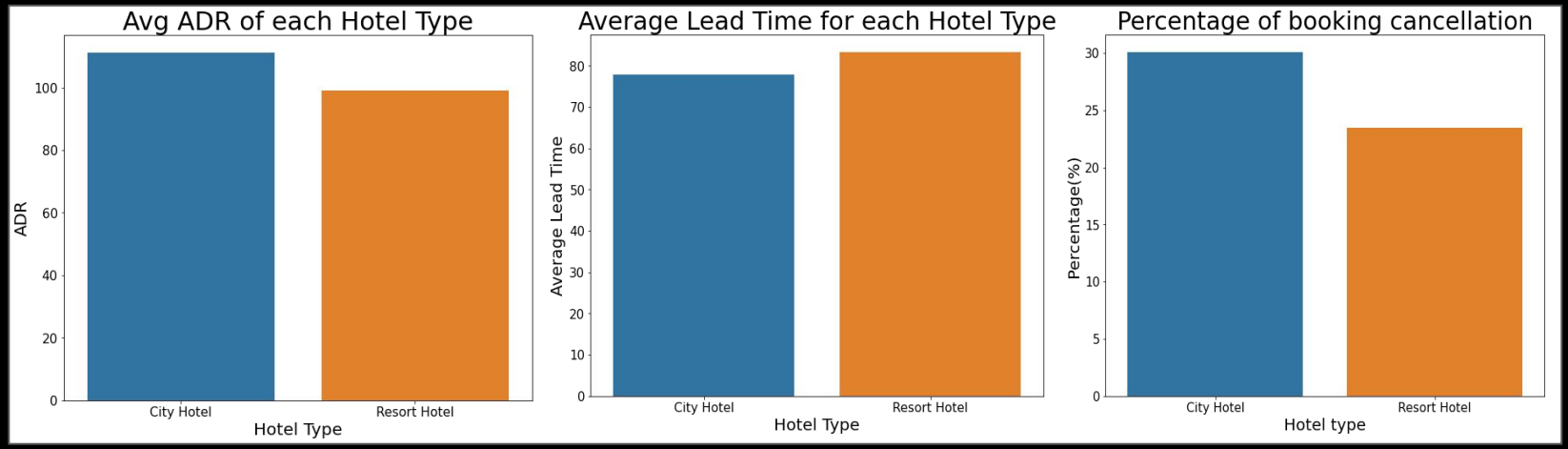
• Univariate Analysis:



Conclusions:

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

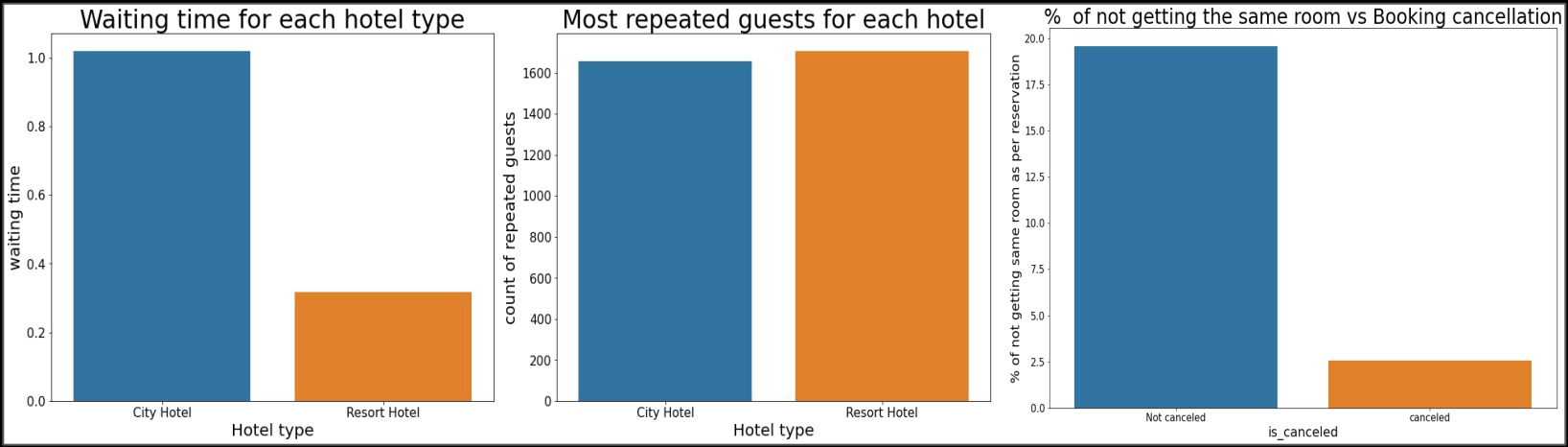
• Bivariate Analysis:



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 %.

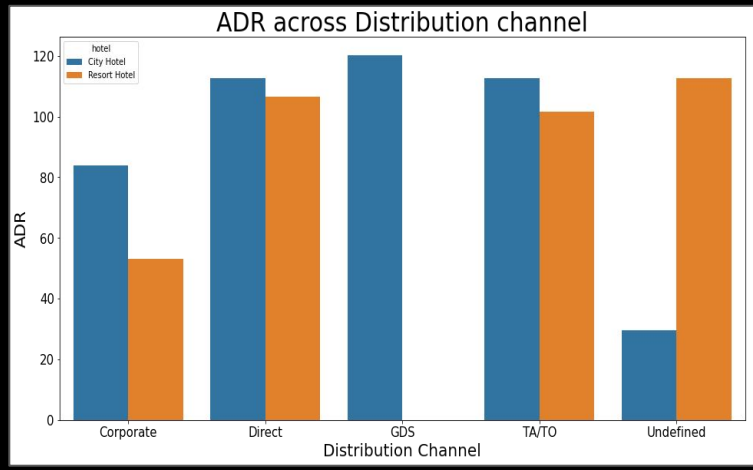
• Bivariate Analysis:

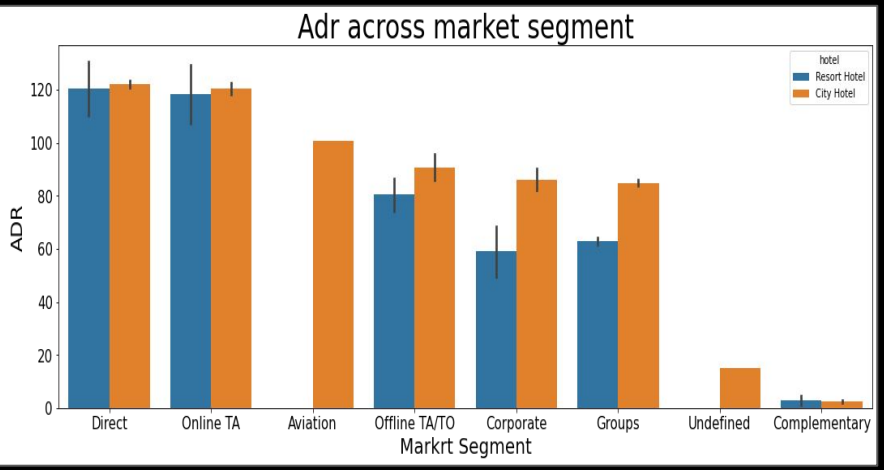


Conclusions:

- ➔ 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.
- ➔ 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

• Bivariate Analysis:



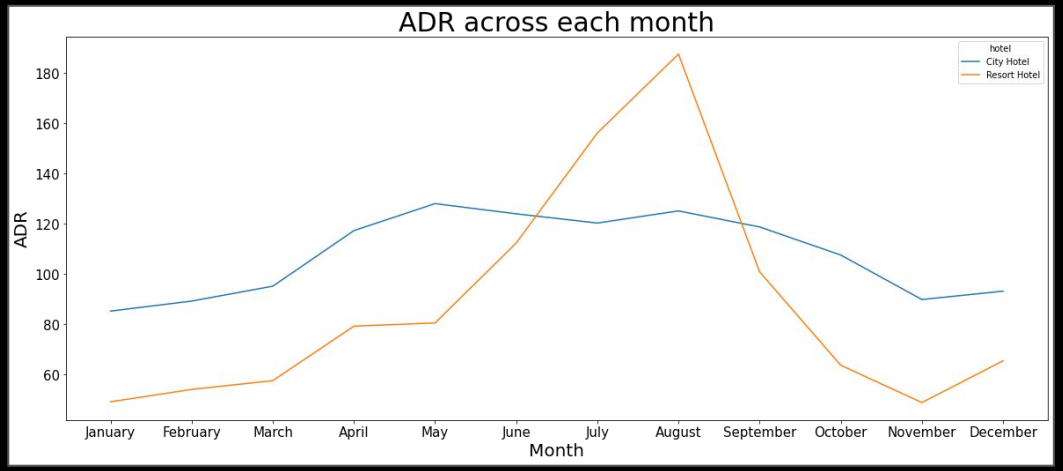


Conclusions:

- ➔ **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.

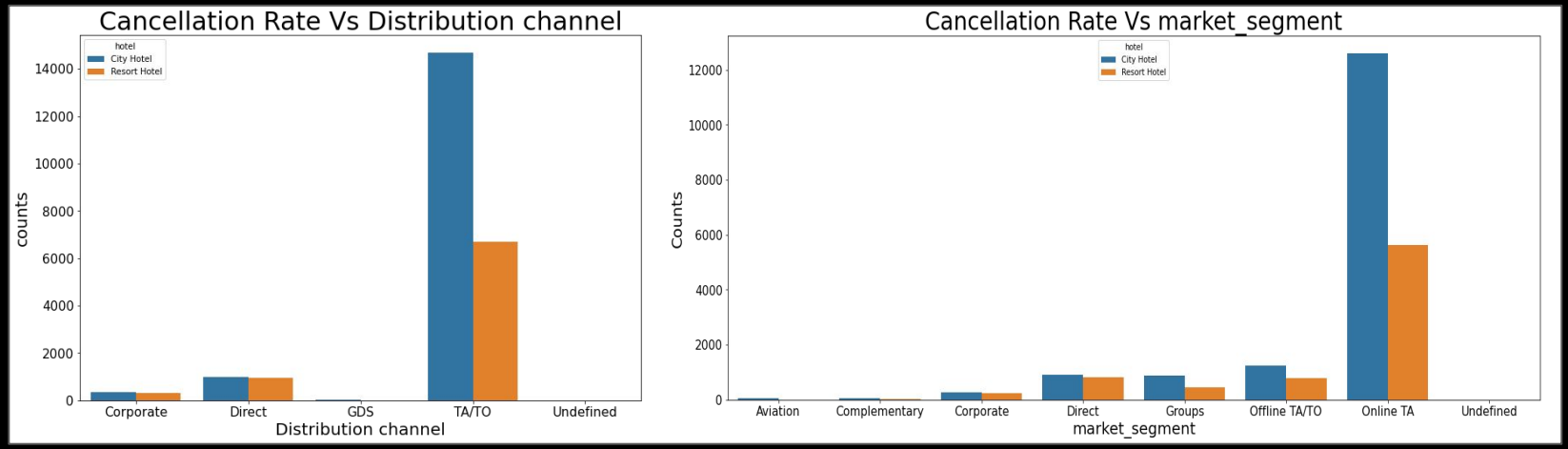
• Bivariate Analysis:



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

• Bivariate Analysis:

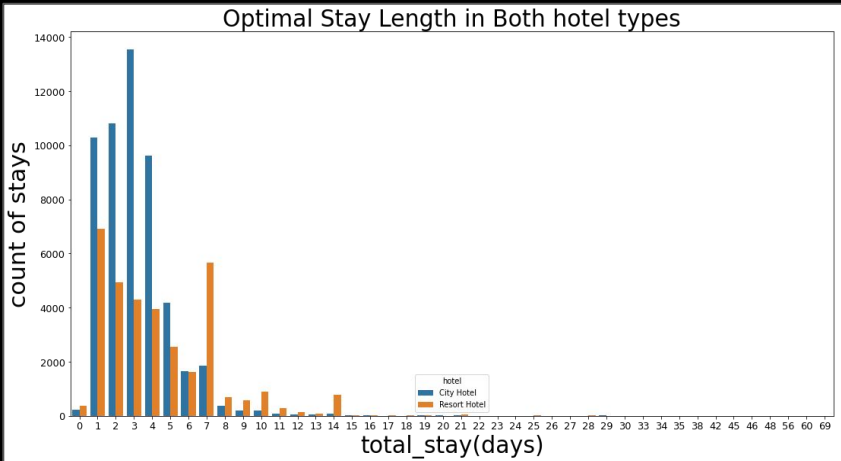


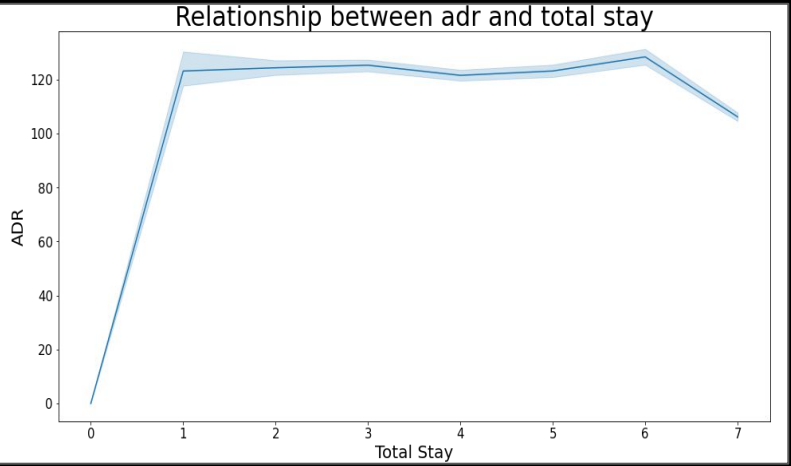
Conclusions:

- ➔ **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.

• Bivariate Analysis:

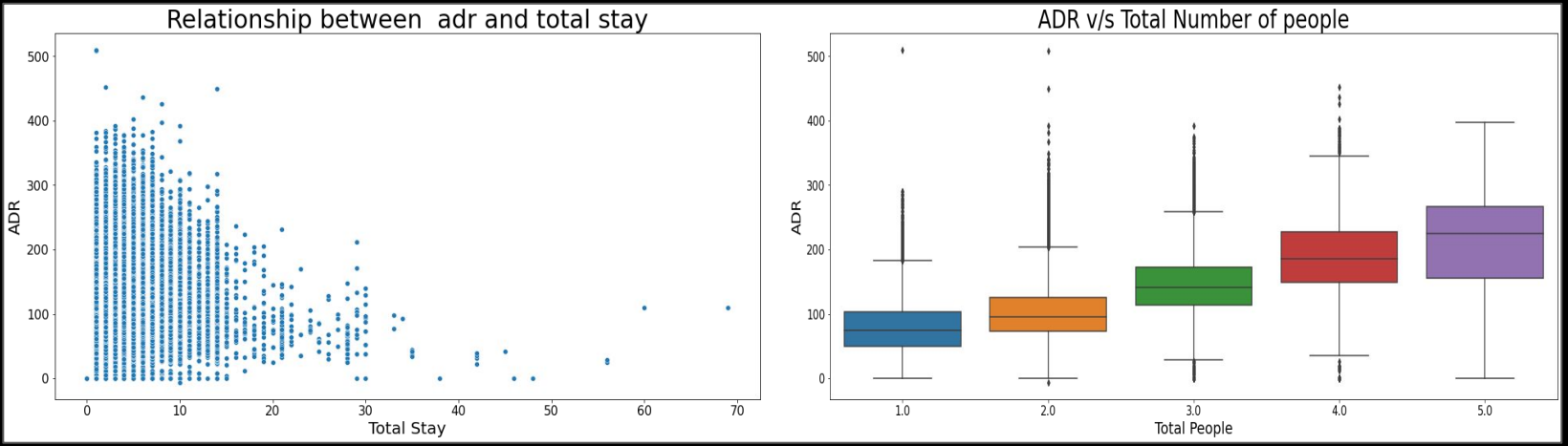




Conclusions:

- ➔ 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 the total stay increases the adr also increases.

• Bivariate Analysis:

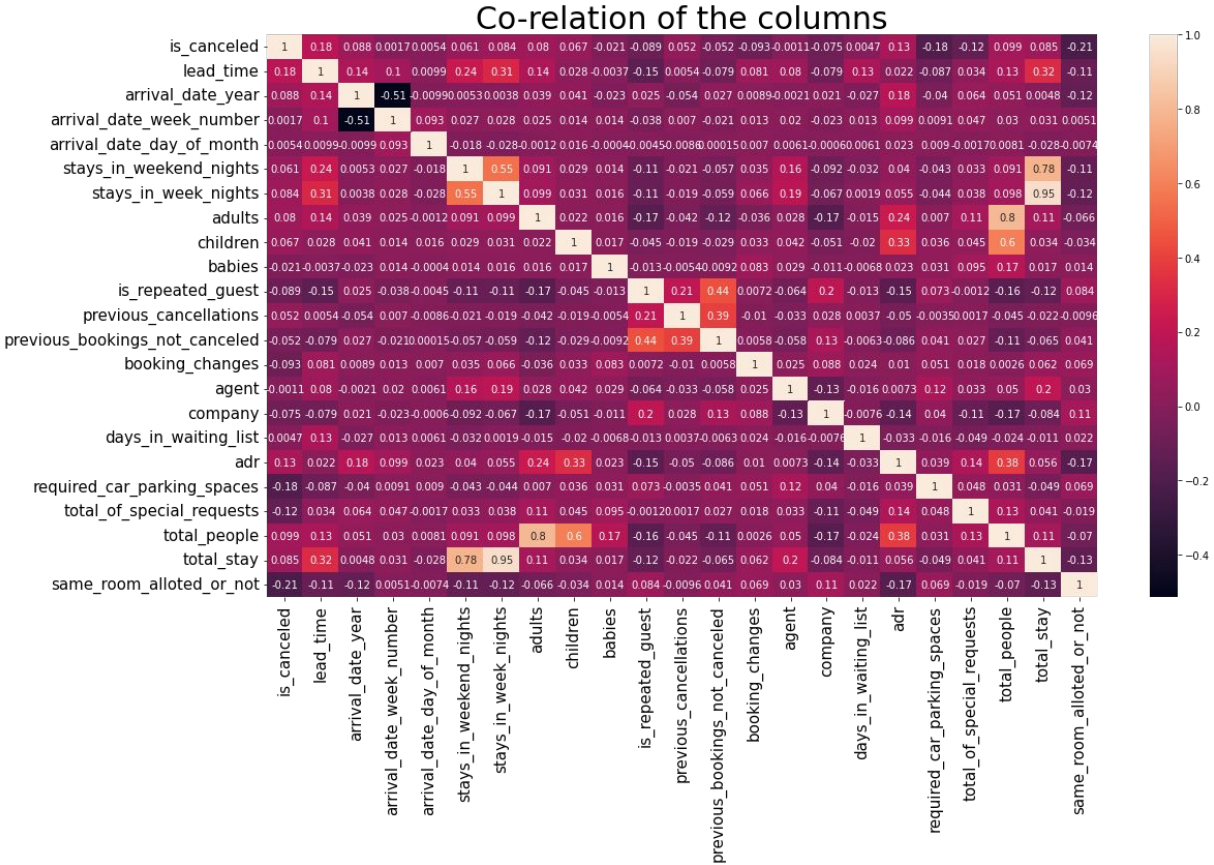


Conclusions:

- ➔ From above scatter we can say that as the stay increases adr is decreasing. Thus for longer stays customer can get good adr.

- ➔ As the total number of people increases adr also increases. Thus adr and total people are directly proportional to each other.

• Heatmap:


Co-relation of the columns
is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
agent
company
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
total_people
total_stay
same_room_alloted_or_not
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

• Heatmap :

Conclusions Heatmap:

- ➔ 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.
- ➔ 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.

Thank You..!!