

Capstone Project - 1

EDA ON HOTEL BOOKING ANALYSIS BY

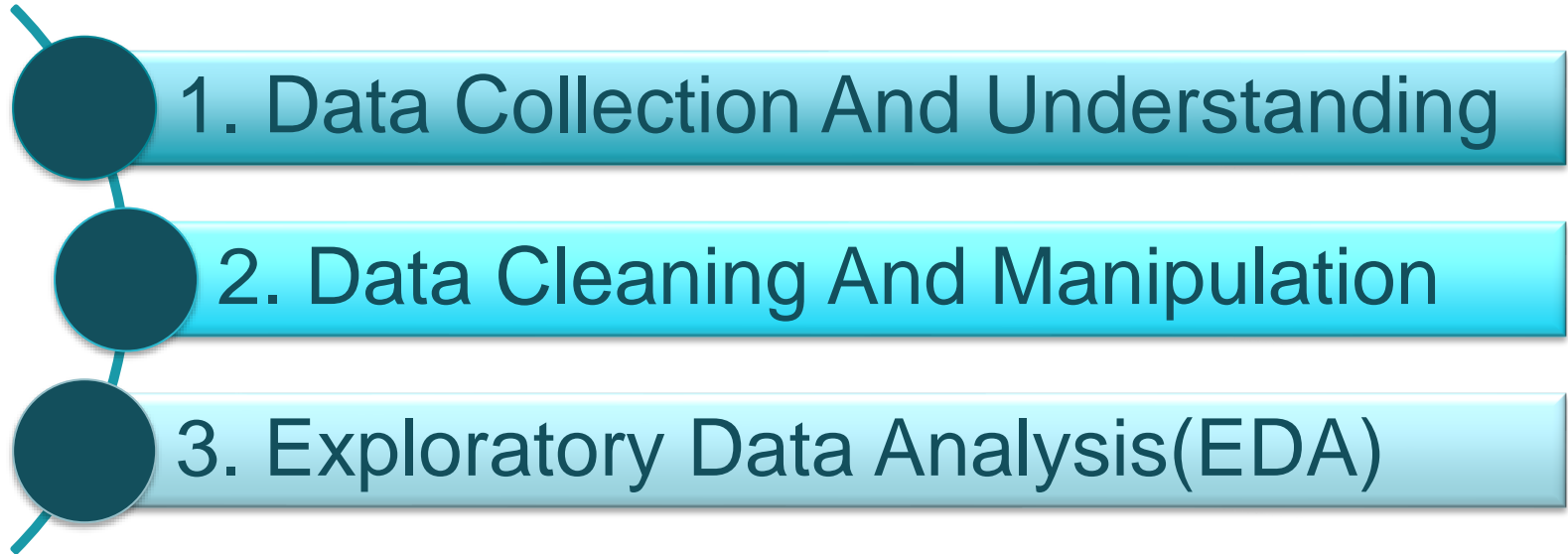
- ***Omkar Desai***
- ***Sumit Berde***
- ***Ankit Khetan***
- ***Saiprasad Bodul***

Problem Statement

- In this project we have done EDA on Hotel Bookings to find useful insights and tried to identify customer behavior which can be used by hotels to increase their profits and accordingly develop suitable business strategies.
- To understand and compare the average daily rates (adr) of both the hotels.
- To identify patterns behind booking cancellations and accordingly take suitable actions to reduce losses due to booking cancellations.

Workflow Of Analysis

The steps involved are as follows :-

- 
1. Data Collection And Understanding
 2. Data Cleaning And Manipulation
 3. Exploratory Data Analysis(EDA)

1. Data Collection And Understanding

It is very important to understand the dataset for any analysis. We had **119390** rows and **32** columns. Lets explore the 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.

reservation_status_date : Date at which reservation status was updated

2. Data Cleaning And Manipulation

Dealing with Nan Values : Nan values were present in 4 columns namely **company**, **agent**, **country** and **children**

- Filling Nan values in company and agent with '0'
- Filling Nan values in country with 'other'
- Filling Nan values in children with '0' . Assuming no children in family that visited the hotel

```
# Dealing with null values
df1.isnull().sum().sort_values(ascending=False)
```

company	82137
agent	12193
country	452
children	4

```
# Filling nan values with '0'
columns = ['company','agent','children']
for i in columns:
    df1[i].fillna(0,inplace=True)

# Filling nan values with 'other'
df1['country'].fillna('other',inplace=True)
```

Handling Duplicate Values : Dataset had **31994** duplicate values, so we have dropped these duplicate values

```
[ ] # Checking for duplicate values  
df1.duplicated().value_counts()    # True means duplicated rows
```

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

▼ We have 31994 duplicate rows

```
# Dropping duplicate rows  
df1 = df1.drop_duplicates()
```

Datetime Object : We have changed datatype of 'reservation status date' column to 'Datetime' object which was earlier 'String' object

```
# Converting str data type to datetime data type  
df1['reservation_status_date']=pd.to_datetime(df1['reservation_status_date'],format='%Y-%m-%d')
```


Handling incorrect values : Some wrong entries were made in the dataset as total people leaving in a room were recorded to be 0. So we have dropped these **166** rows

	hotel	is_canceled	lead_time	reservation_status	reservation_status_date	total_people	total_stays
2224	Resort Hotel	0	1	Check-Out	2015-10-06	0.0	3
2409	Resort Hotel	0	0	Check-Out	2015-10-12	0.0	0
3181	Resort Hotel	0	36	Check-Out	2015-11-23	0.0	3
3684	Resort Hotel	0	165	Check-Out	2016-01-04	0.0	5
3708	Resort Hotel	0	165	Check-Out	2016-01-05	0.0	6

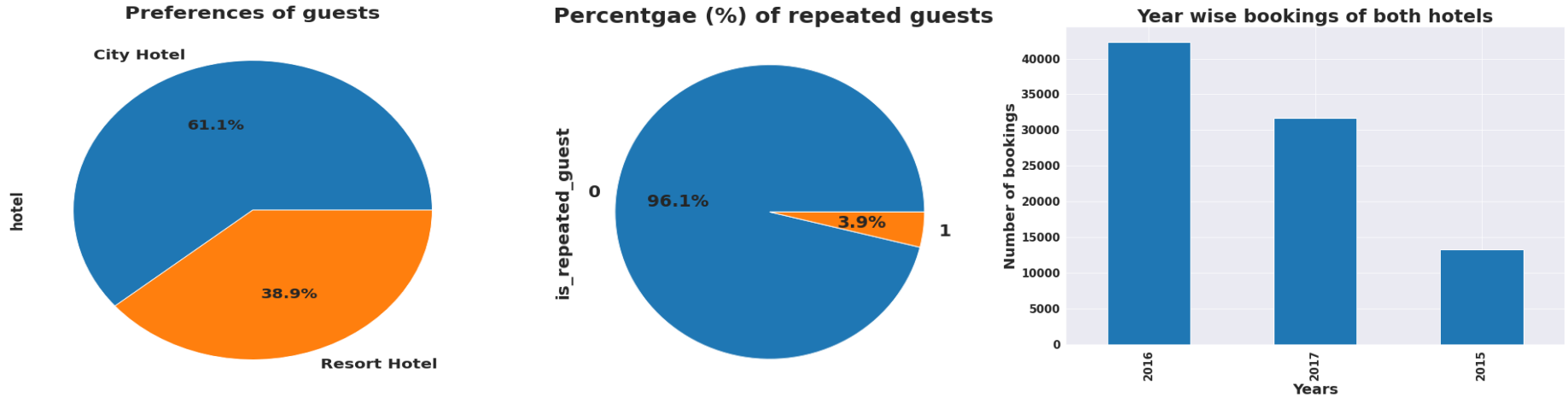
```
# Deleting rows where total_people=0  
df1.drop(df1[df1['total_people']==0].index,inplace=True)
```

```
# Final rows and columns  
df1.shape  
  
(87230, 34)
```

So after doing all the cleaning and manipulations we were left with **87230** rows and **34** columns

3. Exploratory Data Analysis (EDA)

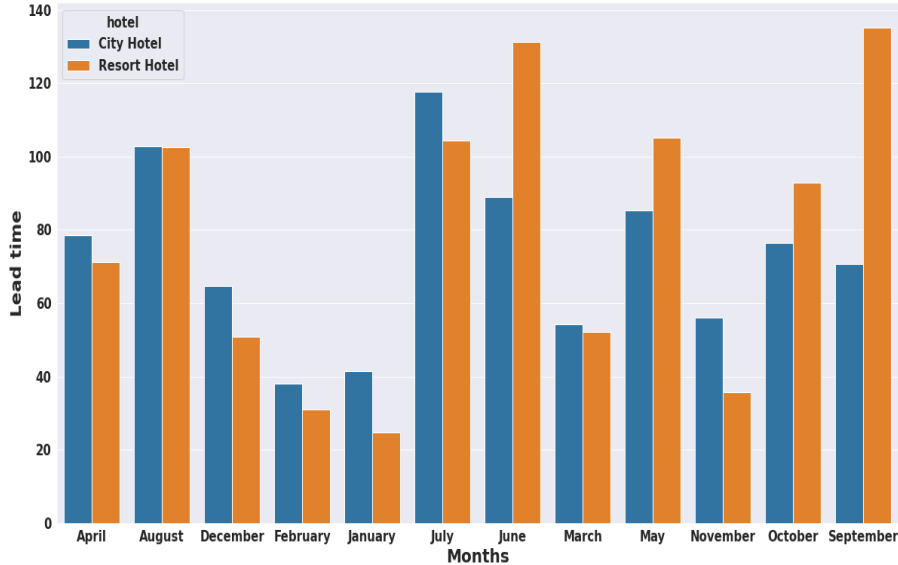
Lets start with the basic level analysis of the dataset



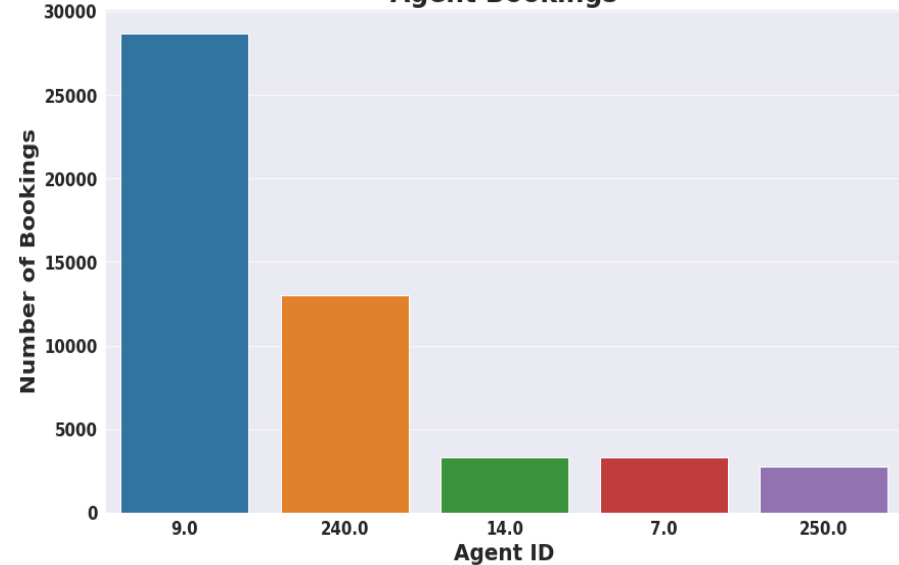
Observations :

- 61.1% of guests preferred City hotel and 38.9% of guests preferred Resort hotel. Hence most of the guests preferred City hotel.
- 96.1% of guests were new guests. Only 3.9% guests revisited the hotel. Thus retention rate is very low for both the hotels.
- In 2016 maximum number of guests visited the hotel and 2015 was the worst performing year for the hotels.

Variation of lead time across months

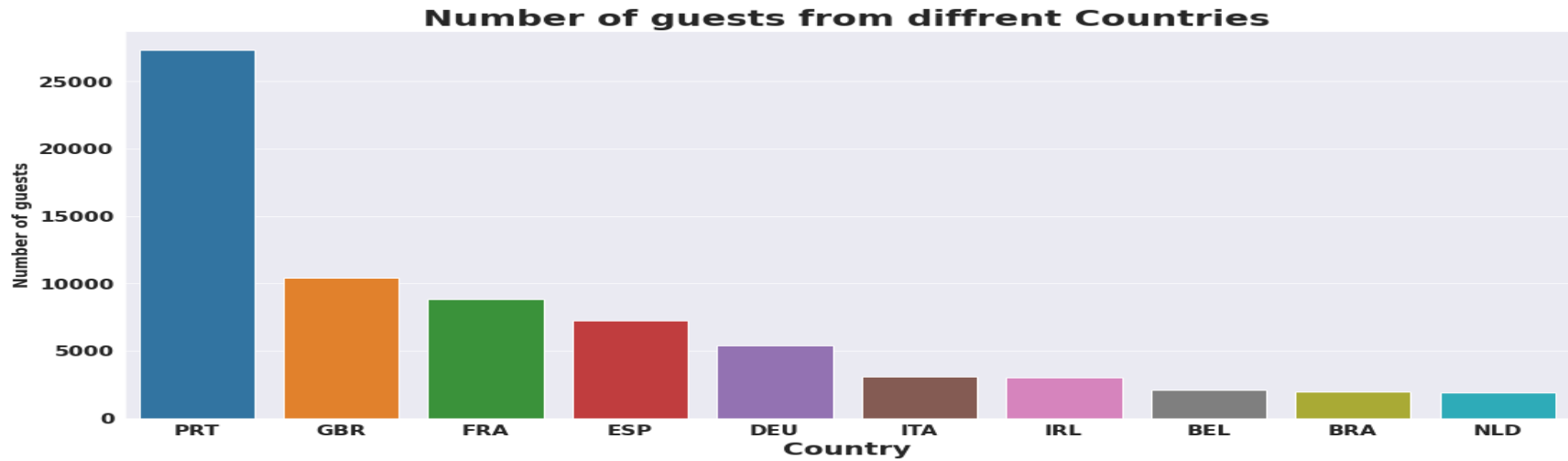


Agent Bookings



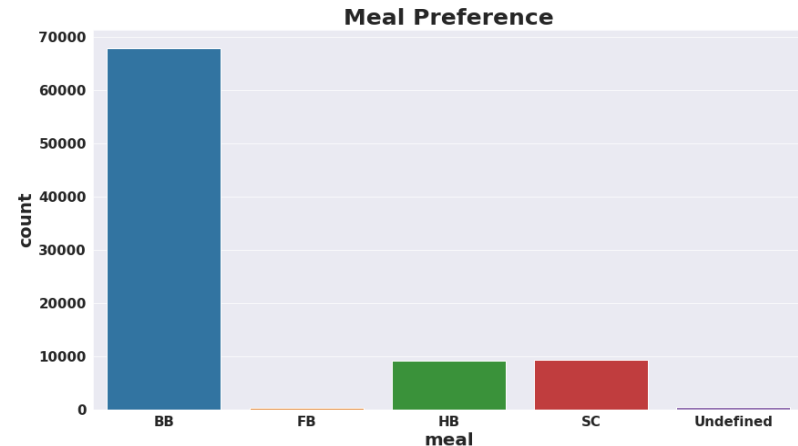
Observations :

- City hotel has high lead time in the months of July and August while Resort hotel has high lead time in the months of June and September.
- Agent ID 9 has done most number of bookings which is more than 28700.



Observations :

- Maximum number of guests were from Portugal that is more than 25000.
- After Portugal the most number of guests were from Great Britain, France and Spain.
- Bed & Breakfast (BB) is the most preferred type of meal by the guests.



Lets deep dive into our dataset and try to answer some complex questions

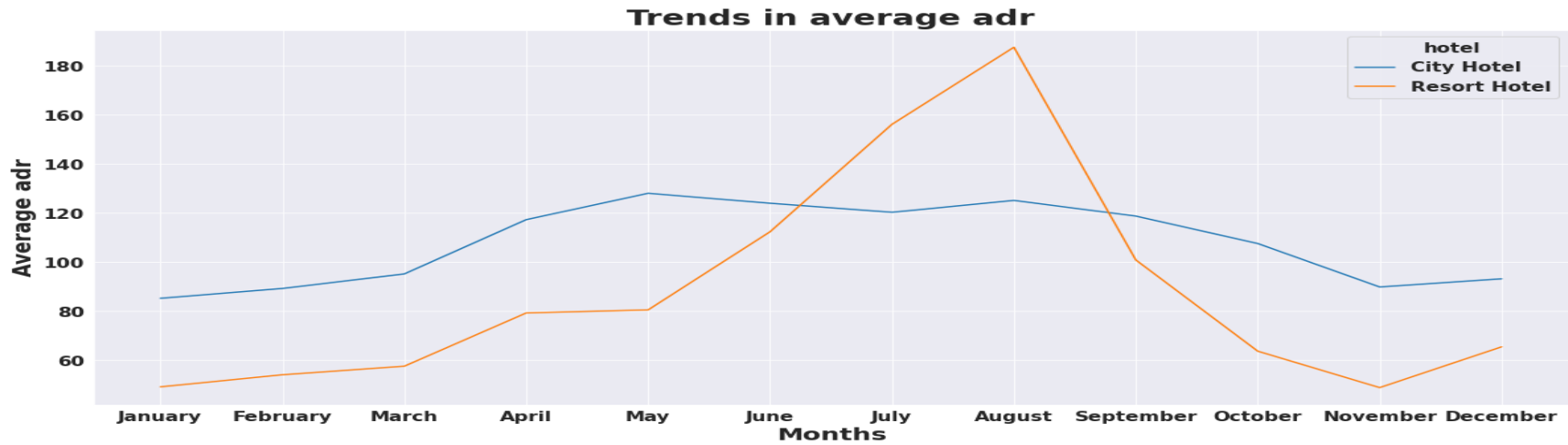


Lets understand the trends and main factors affecting the average daily rates (adr) of hotels



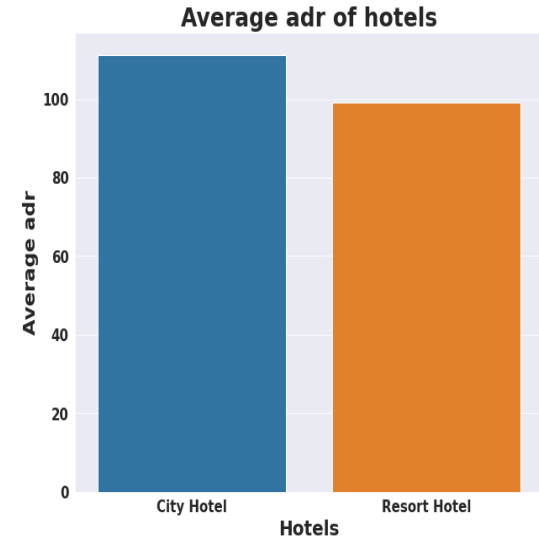
Observations :

- We can see that the number of bookings steadily increases reaches maximum value and then drastically decreases across months.
- July and August were the months where bookings reached their peak values.

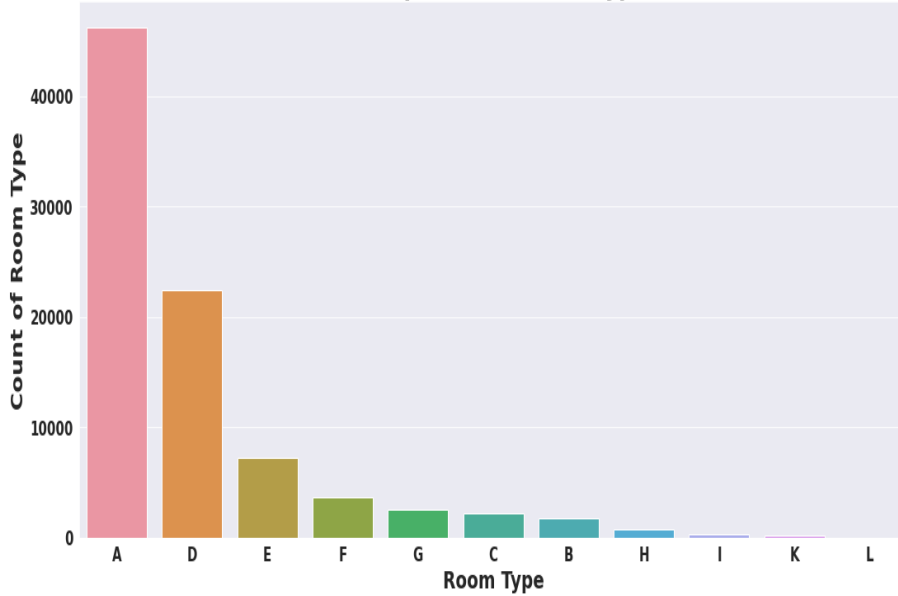


Observations :

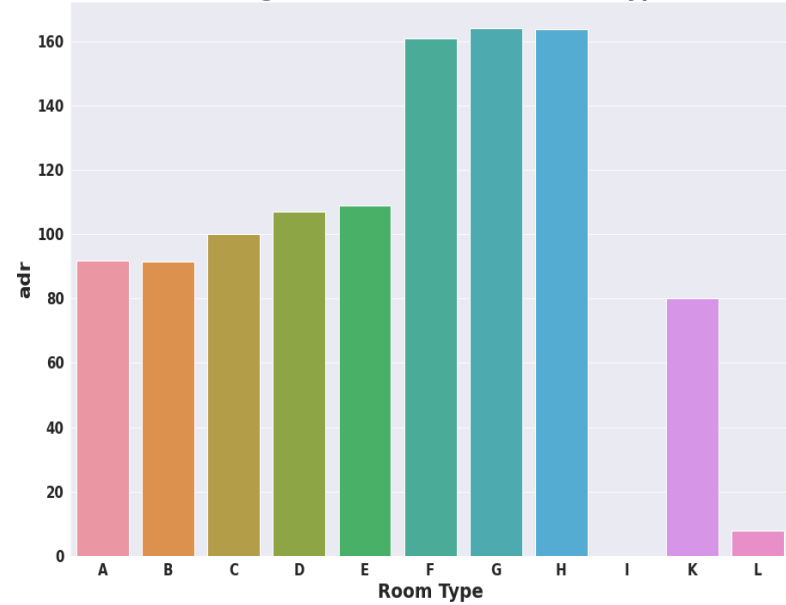
- City hotel have more or less same average adr across months showing a more consistent performance while Resort hotel show large variation in average adr.
- City hotel has the highest average adr in the month of May(128.05) while Resort hotel in the month of August(187.57)
- City hotel have high average adr than Resort hotel. Hence City hotel is more profitable as compared to Resort hotel.



Most preferred Room type

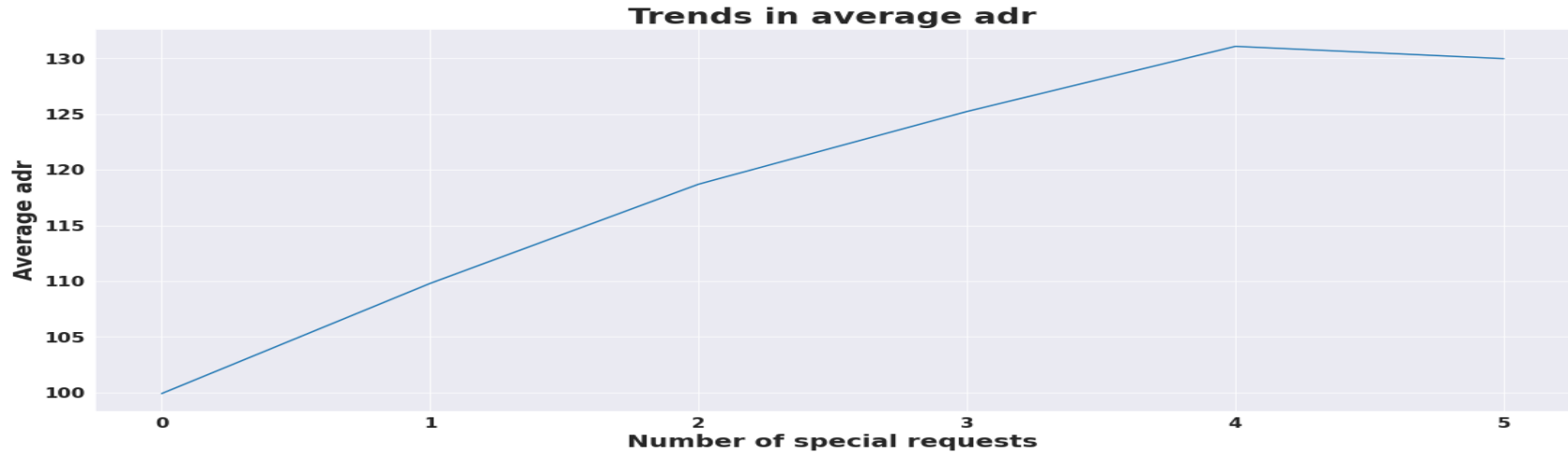


Average adr across different room type



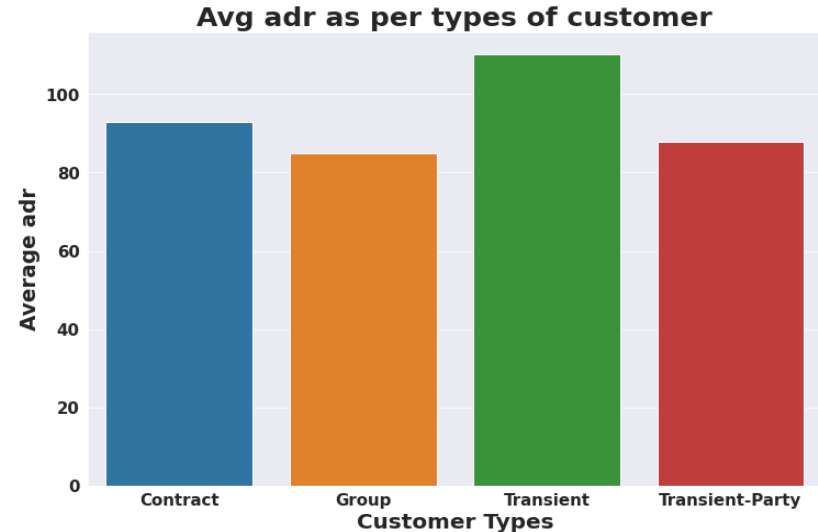
Observations:

- A,D,E are most preferred room types by guests.
- F,G,H are the rooms with high average adr.
- Hence we can say that even though F,G,H are less preferred they are more profitable for hotels. They are like luxury rooms.



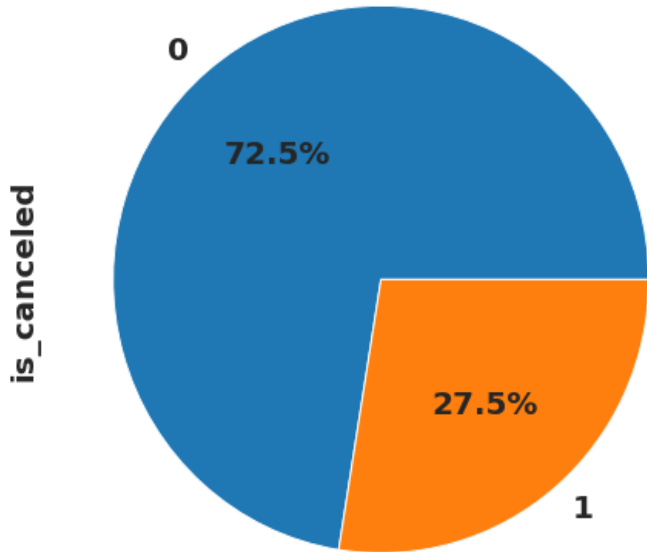
Observations :

- We can see that as the number of special requests increases the average adr also increases.
- Transient types of guests are spending the most followed by contract type.



Now let's understand the various patterns behind booking cancellations and its effect on profitability of hotels

Cancellation and non Cancellation



Observation : We can see that out of total bookings, 27.5% of bookings were cancelled.

	hotels	total_bookings	total_cancellations	percentage_cancellations	avg_loss_in_adr
0	City Hotel	53274	16035	30.10	117.379025
1	Resort Hotel	33956	7974	23.48	118.799984

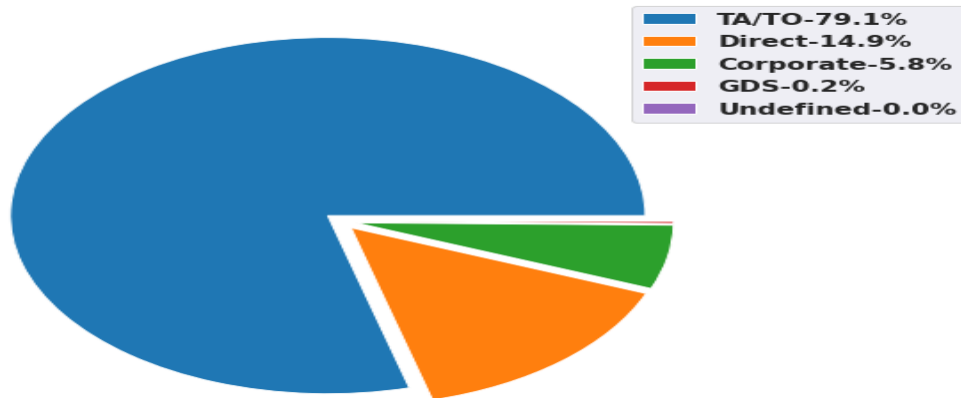
Observations :

- City hotel has higher percentage of booking cancellations as compared to Resort hotel.
- Average loss in adr due to cancellations for Resort hotel(118.79) is slightly higher than City hotel(117.37).



Percentage Distribution Channel

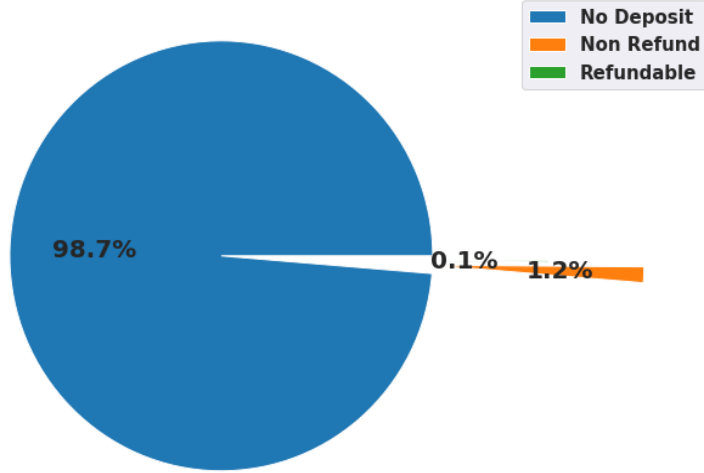
distribution_channel

**Observations :**

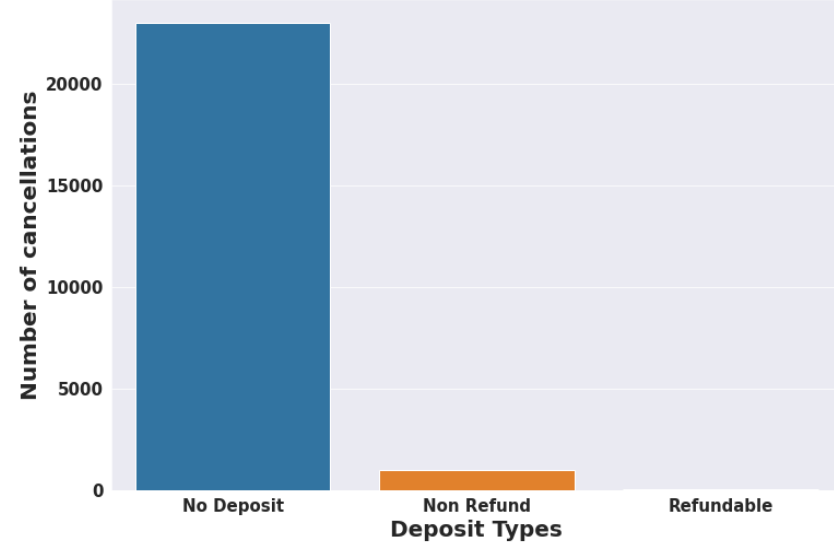
- 79.1% of total bookings were made through TA/TO followed by Direct bookings which is 14.9%
- As excepted out of 16035 cancellations for City hotel, 14649 cancellations took place through TA/TO.
- Similarly out of 7974 cancellations for Resort hotel 6706 cancellations took place through TA/TO.

distribution_channel	Corporate	Direct	GDS	TA/TO	Undefined	total_cancellations
hotel						
City Hotel	330.0	971.0	36.0	14694.0	4.0	16035
Resort Hotel	316.0	952.0	NaN	6706.0	NaN	7974
total_cancellations	646.0	1923.0	36.0	21400.0	4.0	24009

Percentage Deposit Type

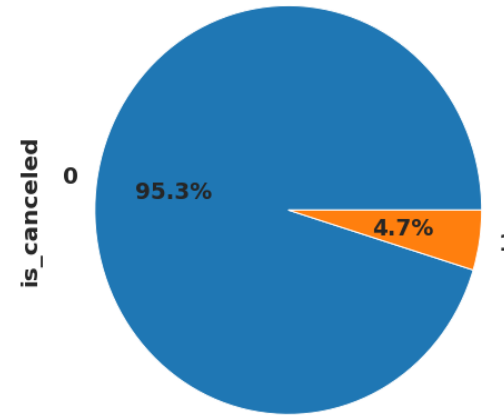
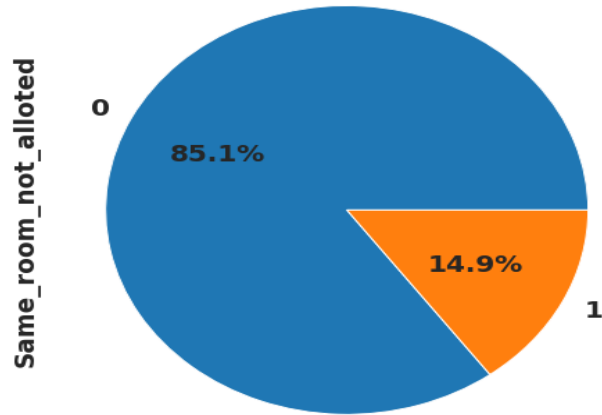


Cancellations as per deposit types



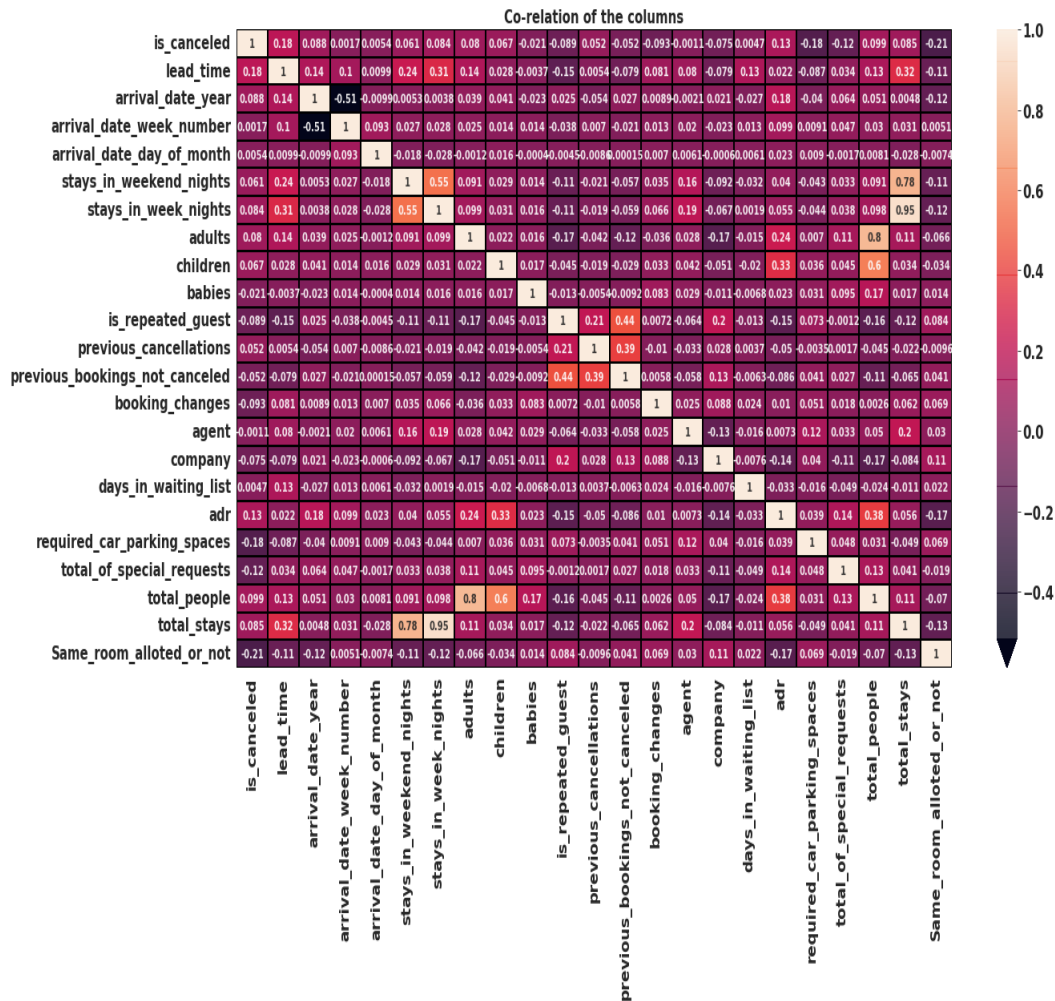
Observations :

- We can see that most of the bookings were made through 'No Deposit' mode almost 98.7%
- Also most of the cancellations were made in 'No Deposit' mode.
- We can now make an important conclusion that if hotels would have taken some charges for bookings, the cancellations would have decreased drastically.



Observations:

- We can see that out of total bookings, there were 14.9% bookings where guests were not assigned the same room for which they had made their reservations.
- Out of these 14.9% bookings only 4.7% of bookings were cancelled.
- Here we can make an important conclusion that not getting the same room as per reserved room is not the reason for booking cancellations.



Observations :

- lead_time and total_stay is positively correlated hence more the stay more will be lead time.
- is_repeated guest and previous_bookings_not_cancelled has strong correlation, may be repeated guests are not more likely to cancel their bookings.
- total_people are positively correlated to adr hence more people more will be adr.

Conclusions:

- As the retention rate is low, to increase it hotels can follow different ways such as:-
 - i. Dealing with guests issues instantly to increase customer satisfaction.
 - ii. Excite guests with exclusive packages and deals so they are tempted to visit again.
- F, G, H are most profitable room types, so encourage guests to book these rooms by offering special complementary services exclusive to only these rooms.
- Hotels can run discount offers and promotions to encourage guests to stay longer, as most customers are transient type and will be scouting for good deals.
- Cancellation rate is more since there are no cancellation charges, so having cancellation charges policy may lead to dip in cancellation of bookings in both hotels.
- Hotels can minimize TA/TO cancellations by having a lower commission for TA/TO with higher cancellations.

THANK YOU