Capstone Project-1

Exploratory data analysis of Hotel Booking

By:- Praveen kumar sharma

List of content

Problem Statement

- We were given a dataset that contains booking information for a city hotel and a resort hotel. It includes information such as booking time, length of stay, number of adults, children/babies, number of available parking spaces, etc.

- As we know, hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more.

- Our main objective is perform EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

EDA process flow chart

Data types
- Float = 4
- Int = 16
- Object = 12

Data Collection and Understanding
- Data collection
- Understood the dataset in term of information given
- Understood the datatypes of each columns

Data Cleaning and Manipulation
- Removed duplicate rows.
- Handled missing values.
- Converted columns to appropriate datatypes.
- Added important columns.

Exploratory Data Analysis (EDA)
- Univariate Analysis
- Hotel wise Analysis
- Distribution channel wise Analysis
- Booking cancellation Analysis
- Customer Centric Analysis
- Special Request Analysis

Data Collection and Understanding:

We have a Dataset of hotel booking analysis from years 2015 to 2017 and having 32 columns. Our aims to find the relevant insights from this dataset.

Data Description:

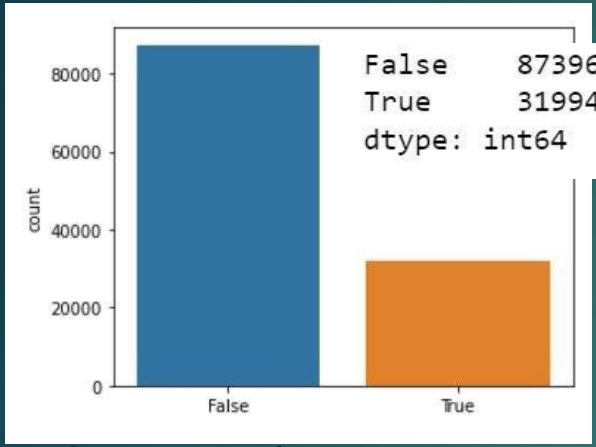
hotel :Resort Hotel or City Hotel
is_canceled :Value indicating if the booking was canceled (1) or not (0)
lead_time :Number of days that elapsed between the entering date of the booking and the arrival date
arrival_date_year :Year of arrival date
arrival_date_month :Month of arrival date **arrival_date_week_number** :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.

Cont.....

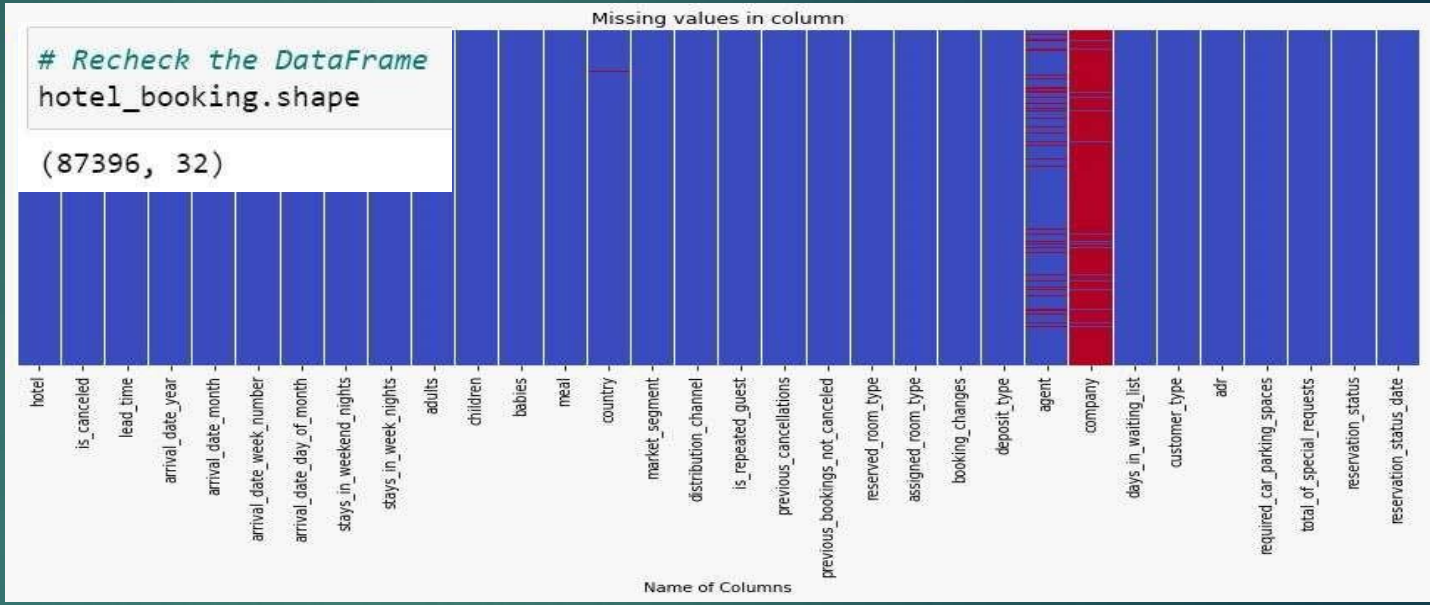
market_segment : Market segment designation. (TA/TO)
distribution_channel : Booking distribution channel.(T/A/TO)
is_repeated_guest : is a repeated guest (1) or not (0)
previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled : 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

1- Remove duplicate rows.



2- Handling missing values.



3- Convert columns to appropriate datatypes.

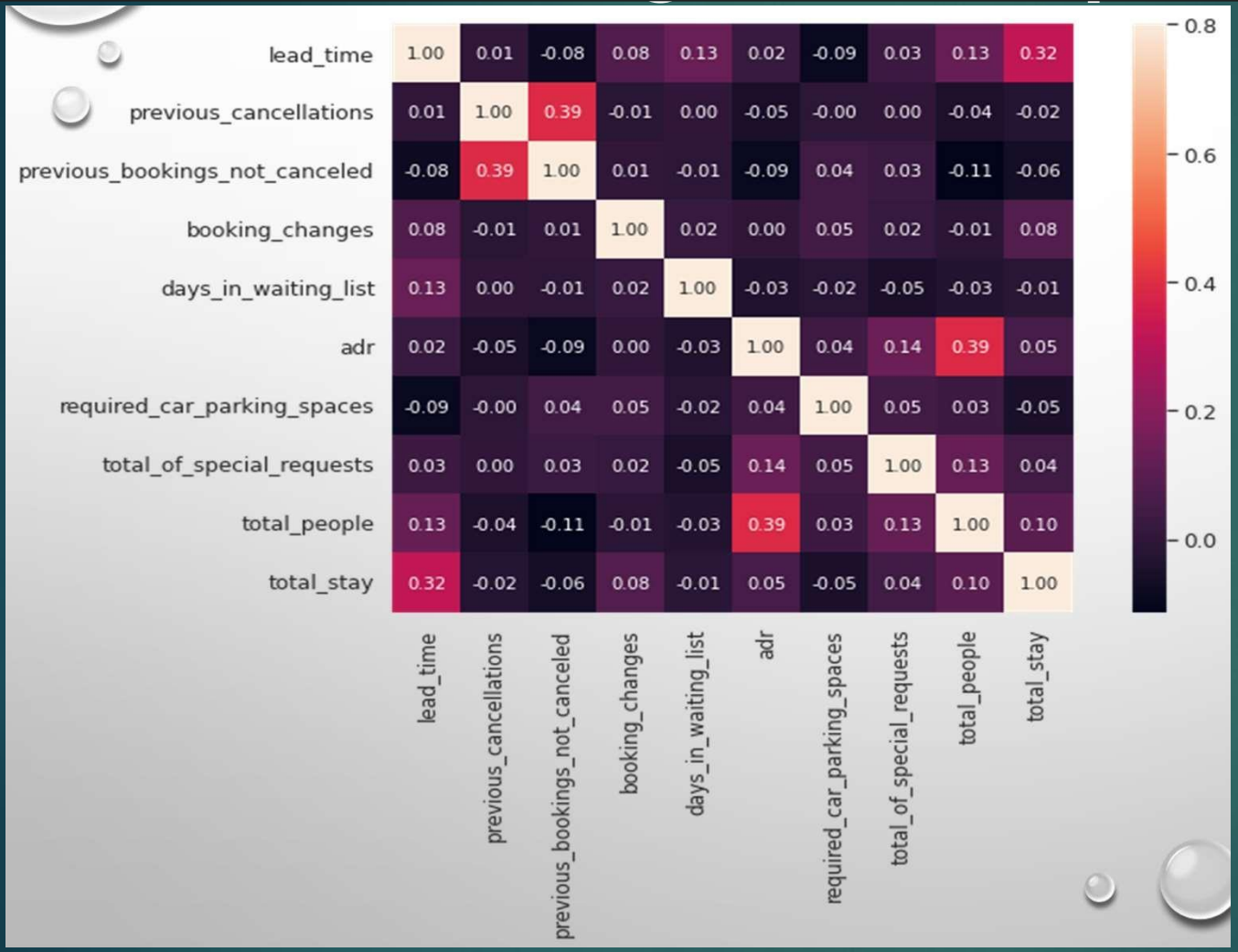
```
dtypes: bool(3), float64(1), int64(16), object(12)
memory usage: 20.3+ MB
```

4- Adding important columns.

```
# Calculating the total_stay by adding `stays_in_weekend_nights` and `stays_in_week_nights`

hotel_booking['total_stay'] = hotel_booking['stays_in_weekend_nights']+hotel_booking['stays_in_week_nights']

# Calculating total_people by adding (numbers of adults + children + babies)

hotel_booking['total_people'] = hotel_booking['adults']+hotel_booking['children']+hotel_booking['babies']
```

- Now our dataset is ready for Analysis (we have successfully replaced duplicates, NaN values, and convert inappropriate datatype to appropriate datatype)

Correlation Heat map



- Total stay length and lead time are slightly correlated. This may means that for longer hotel stays, people generally plan little before the actual arrival.
- adr is slightly correlated with total_people, which makes sense as more no. of people means more service to deliver, therefore more adr.

Exploratory Data Analysis (EDA)

Overview of EDA Analysis

Univariate Analysis:

- What is the most preferred meal by customers?
- What is the percentage distribution of required car parking spaces?
- What is the percentage of booking changes made by the customer?
- What is Percentage distribution of Deposit type ?
- Which is the most preferred room type by the customers?

Hotel wise Analysis:

- Which type of hotel is mostly preferred by the guests?
- What is most preferred stay length in each hotel?
- Which hotel has higher lead time?
- Which hotel makes more revenue?
- Which hotel has the higher customer retention rate?
- For which hotel, does people have to wait longer to get a booking confirmed?
- From which country most guest come?

Distribution channel wise Analysis:

- Which Distribution Channel is contributing in most of the hotel bookings?
- Which channel is contributing most for early booking of the hotel?
- Which distribution channel brings better revenue generating deals for hotels?
- Which is the most favorable Channel for Customers to book hotel?

Booking Cancellation Analysis:

- Which hotel has higher booking cancellation rate?
- How many bookings were cancelled ?
- Which significant distribution channel has highest cancellation percentage?
- What is Yearwise Percentage of Cancellations?
- Which period of year has longer lead time analysis?
- What is Average Daily Rate Yearwise and Monthwise?

Customer centered analysis:

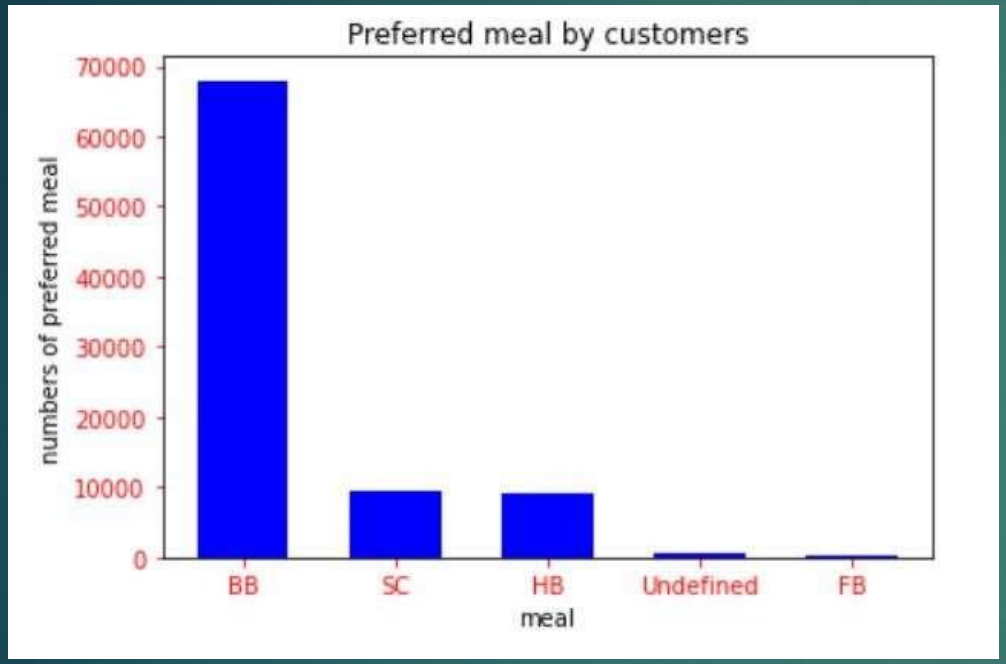
- Which type of hotel is better choice for families?
- How not getting same room as reserved affects adr?
- Overview of arrival period
- Whether Stay is over a weekend or weekday?

Special Request Analysis

- Prediction of whether or not a hotel was likely to receive a disproportionately high number of special requests?

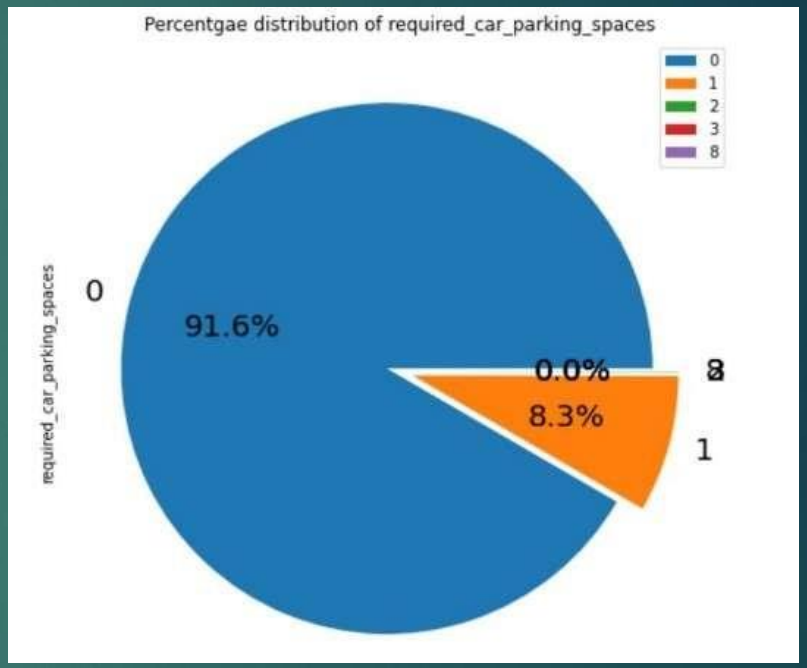
Univariate Analysis

Q1-What is the most preferred meal by customers?



- Most preferred meal is BB – Bed & Breakfast
- HB-Half Board and SC-Self Catering are equally preferred

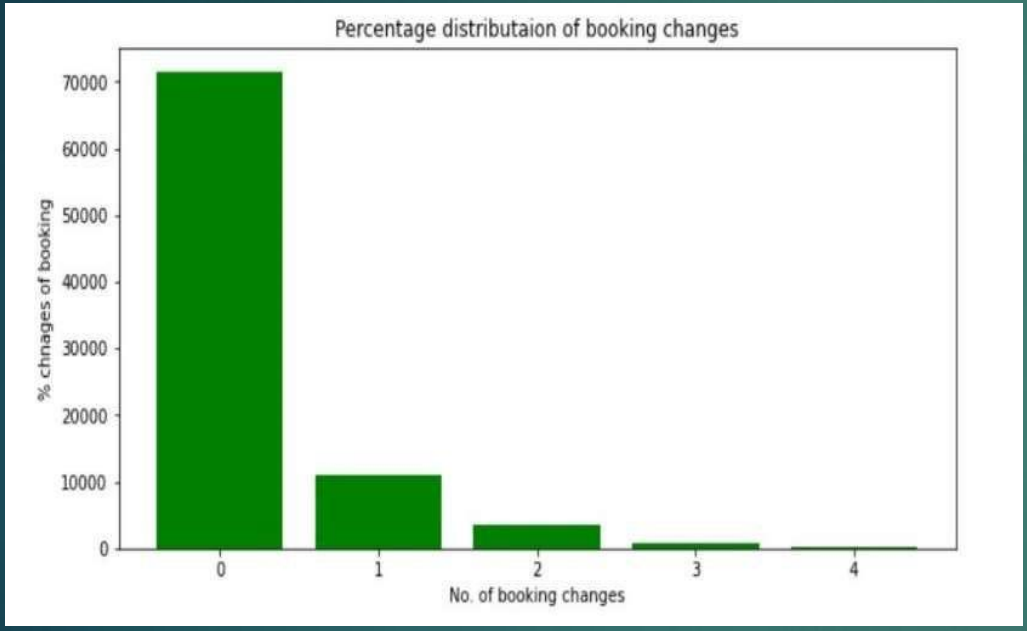
Q2-What is the percentage distribution of required car parking spaces?



- 91.6 % guests did not required the parking space. only 8.3 % guests required only 1 parking space.

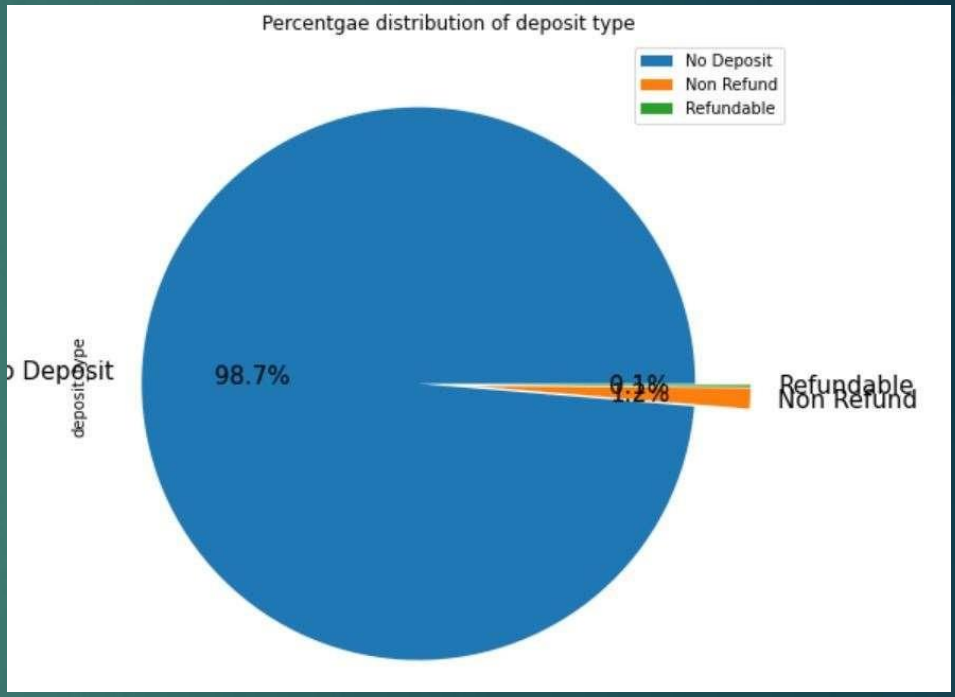
Univariate Analysis

Q3-What is the percentage of booking changes made by the customer?



- 0= 0 changes made in the booking
- 1= 1 changes made in the booking

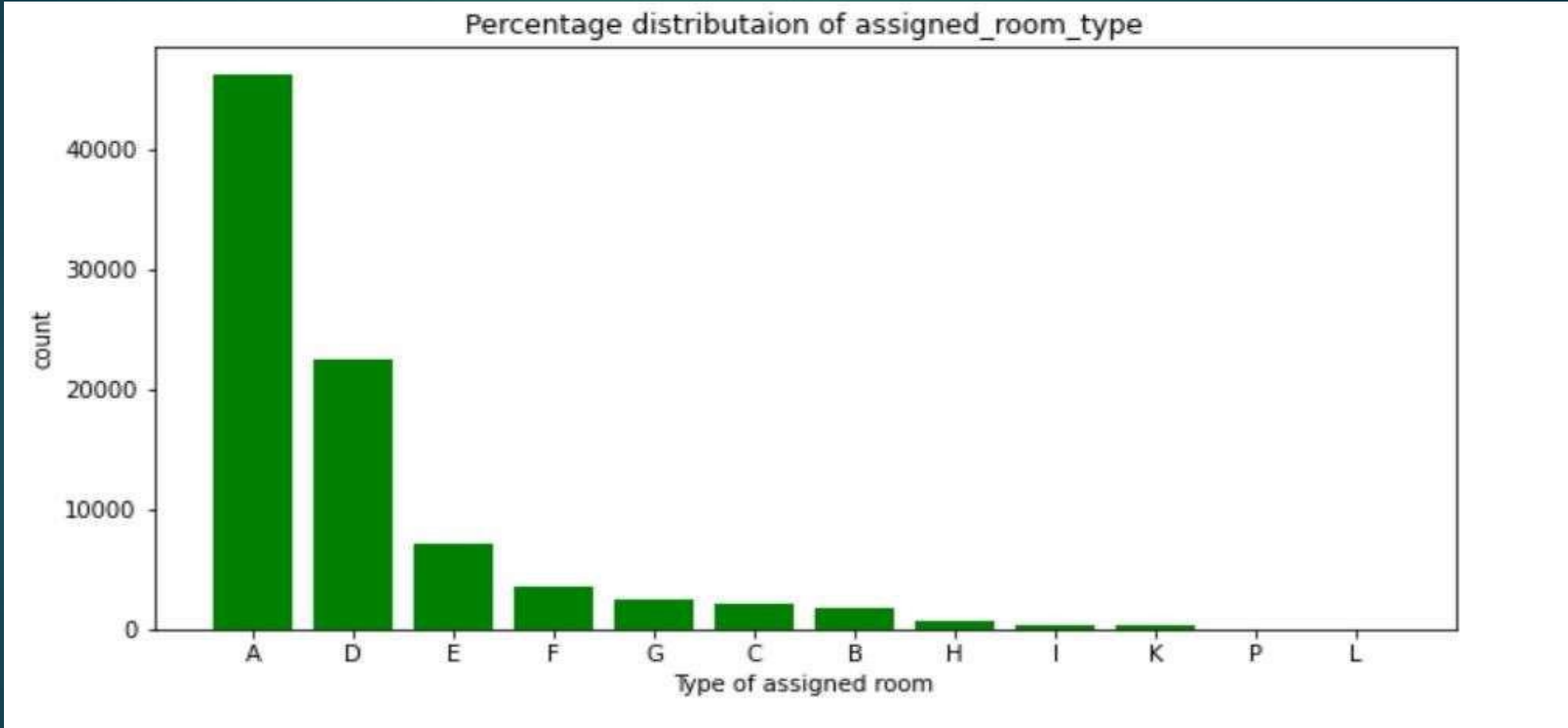
Q4-What is Percentage distribution of Deposit type ?



- 98.7 % of the guests prefer No deposit type of stay.

Univariate Analysis

Q5-Which is the most preferred room type by the customers?

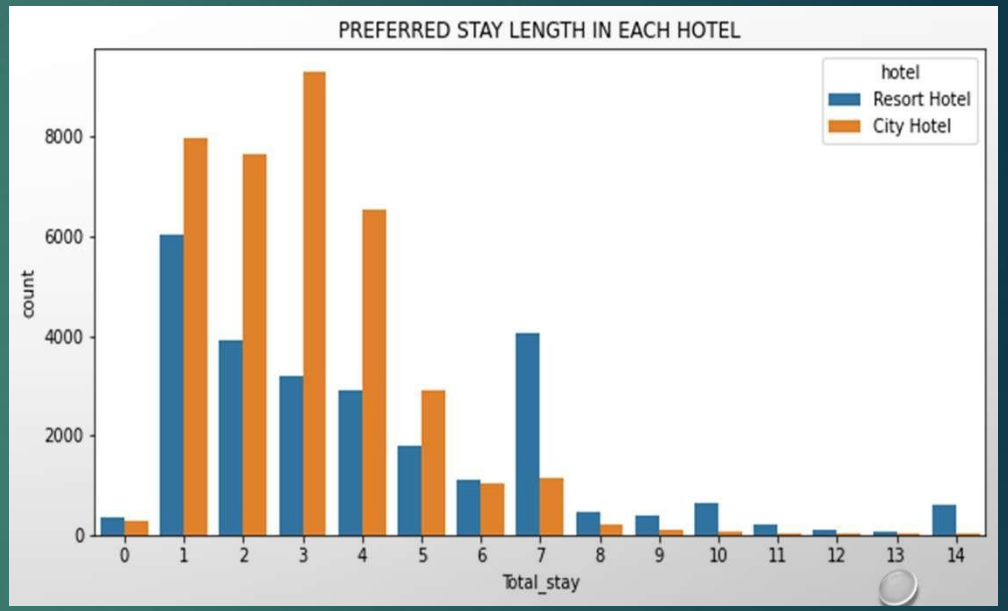


Hotel wise Analysis

Q 1- Which type of hotel is mostly preferred by the guests?

- City Hotel is most preferred by guests and thus city hotel has got maximum bookings.

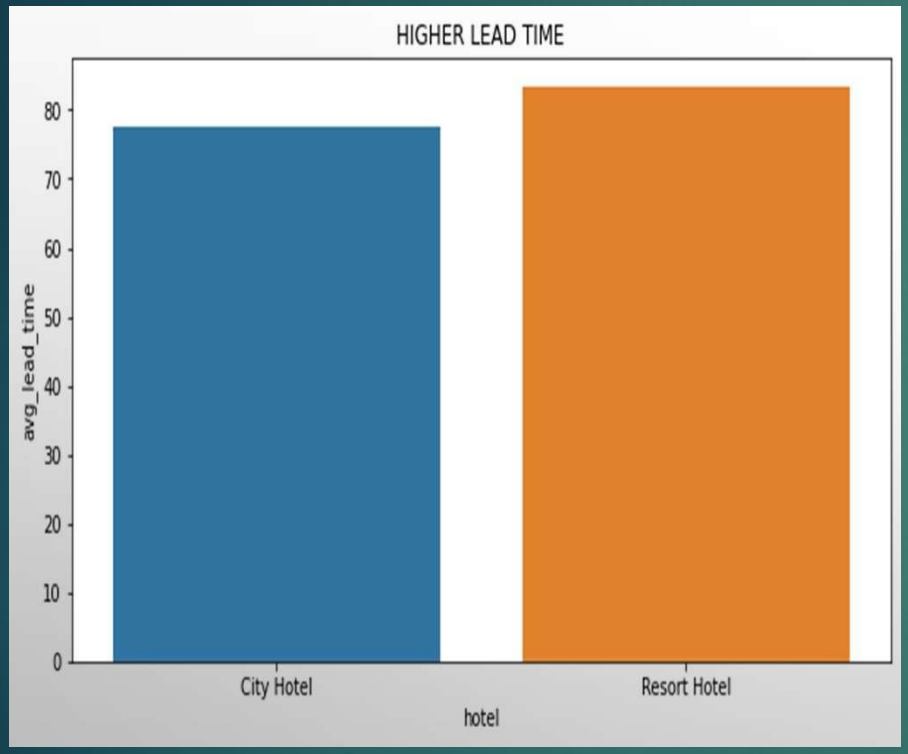
Q 2- What is most preferred stay length in each hotel?



- Most common stay length is less than 4 days and generally people prefer city hotel for shorter stay ,but for longer stay resort hotel is preferred.

Hotel wise Analysis

Q3- Which hotel has higher lead time?



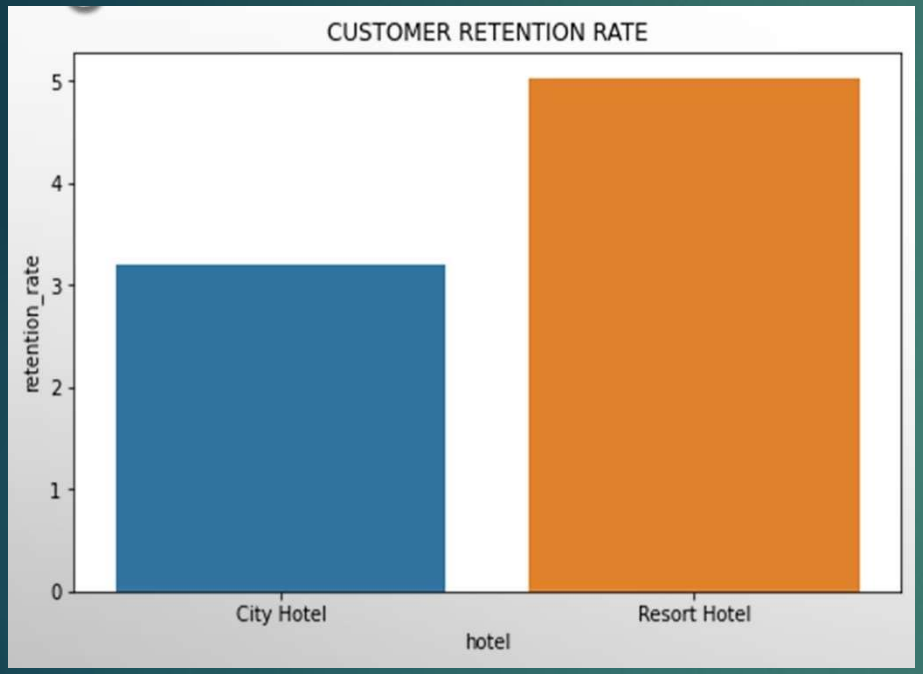
- Resort hotel has slightly high avg lead time.That means customers plan their trips very early

Q4- Which hotel makes more revenue?

- City hotels has slightly more revenue then resort hotel.

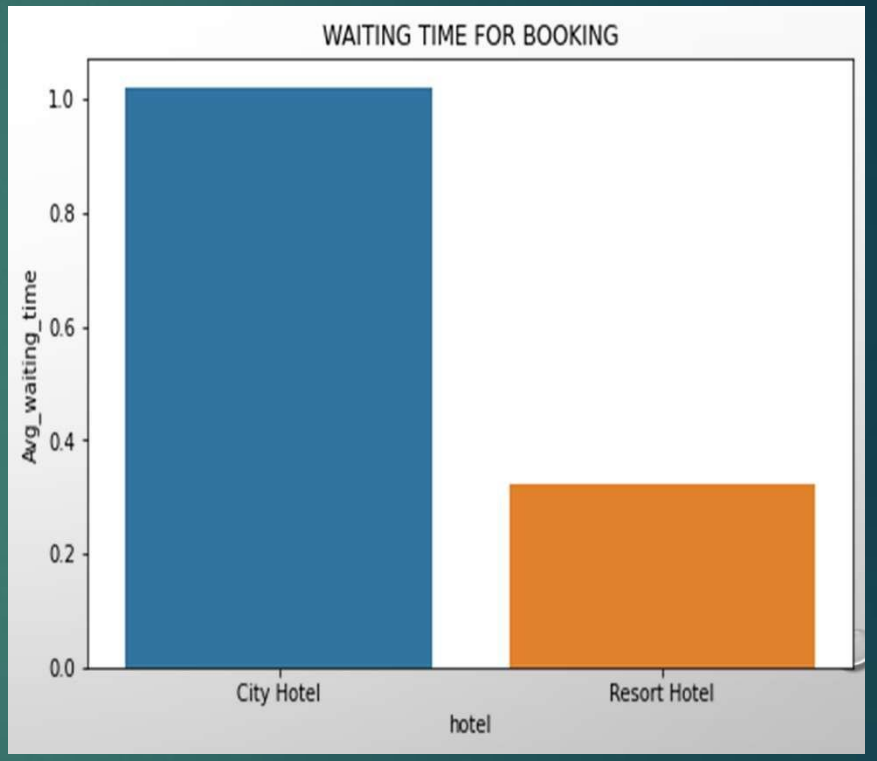
Hotel wise Analysis

Q5- Which hotel has the higher customer retention rate?



- Resort hotel has higher retention rate compare to city hotel that means customers are willing to stay again in resort hotel.
- But retention rate for city hotel 3.20% and for resort hotel is 5.03% which is very less.

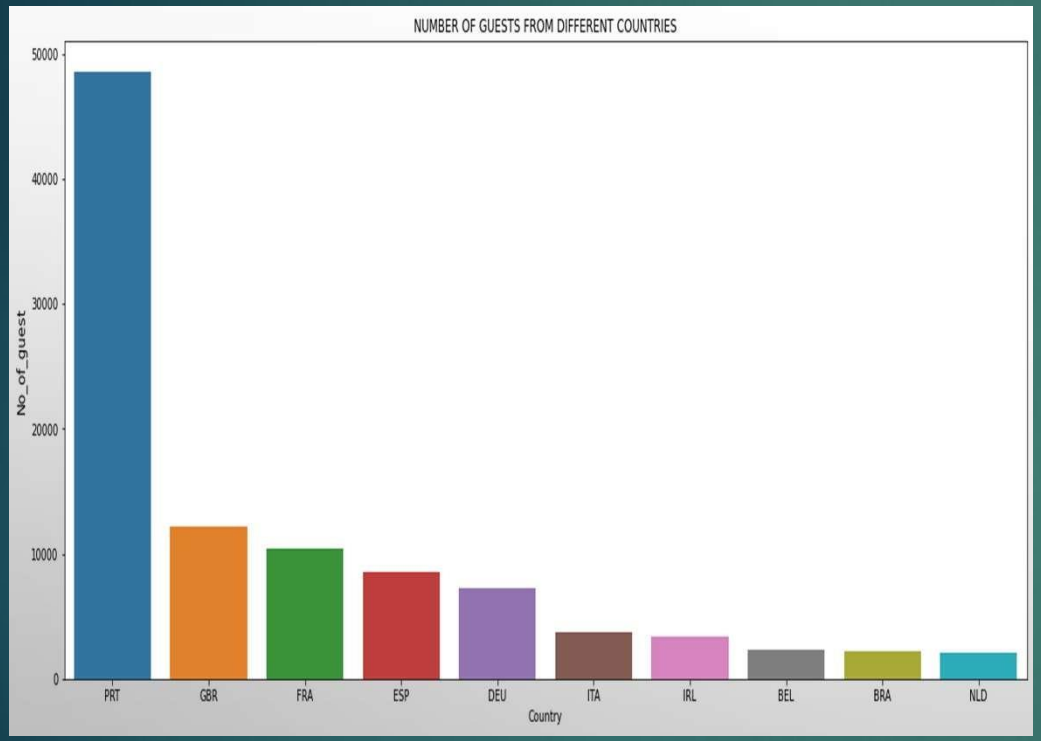
Q6- For which hotel, does people have to wait longer to get a booking confirmed?

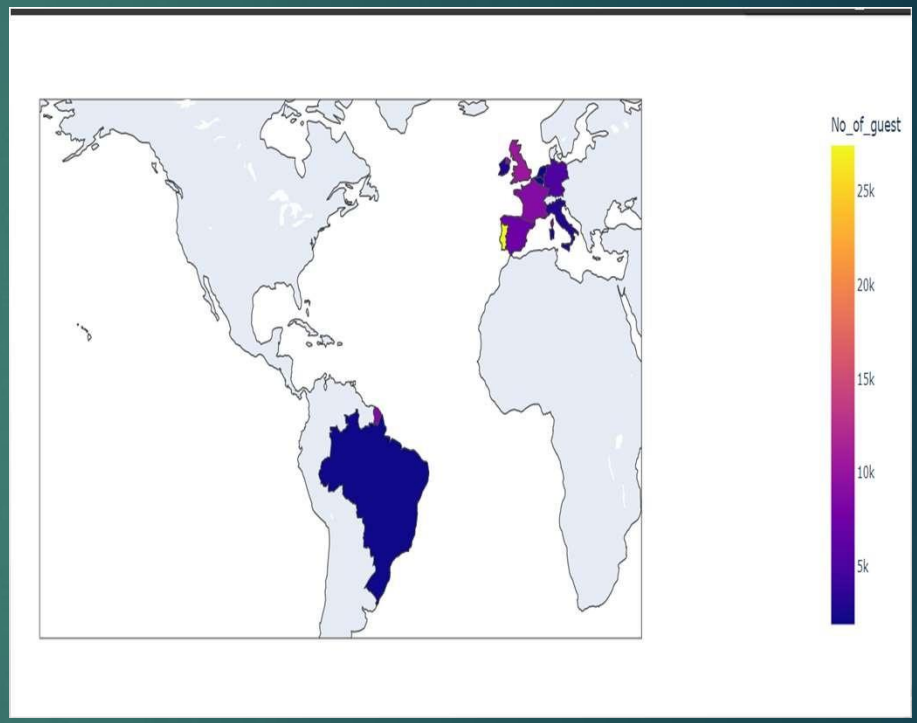


- City hotel has significantly longer waiting time then resort hotel hence City Hotel is much busier than Resort Hotel.

Hotel wise Analysis

Q7- From which country most guest come?

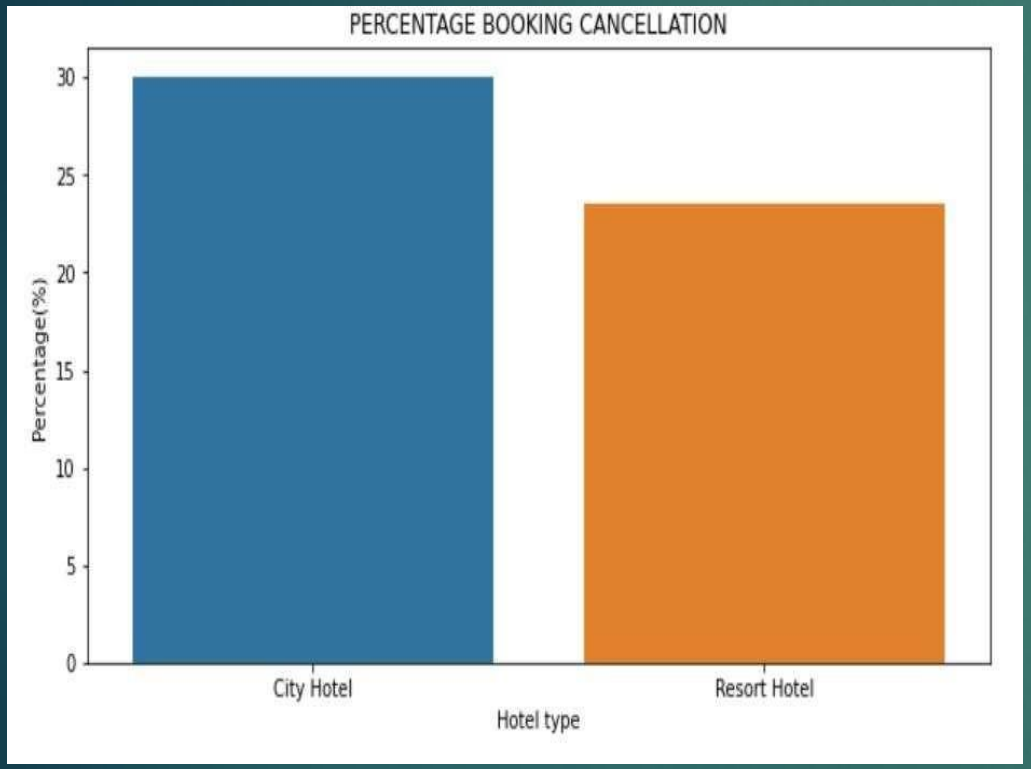




- We have a huge number of visitors from western Europe, namely Portugal, UK and France being the highest.
- We can instruct the marketing team to target people of this region

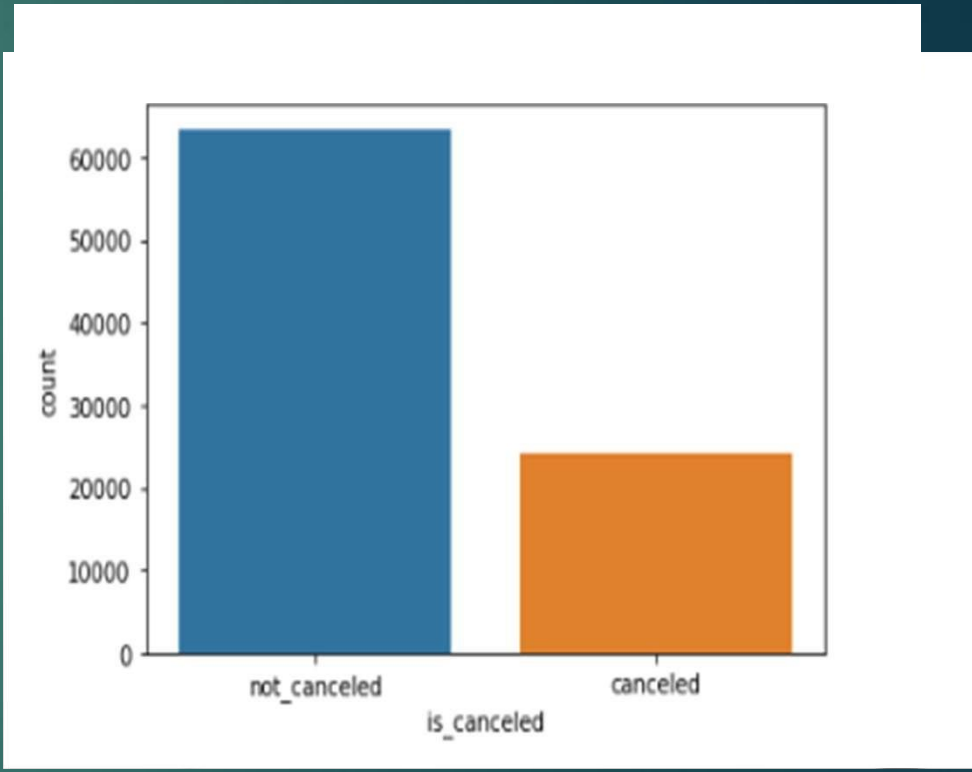
Booking Cancellation Analysis

Q1- Which hotel has higher booking cancellation rate?



- City hotel has higher booking cancelation rate

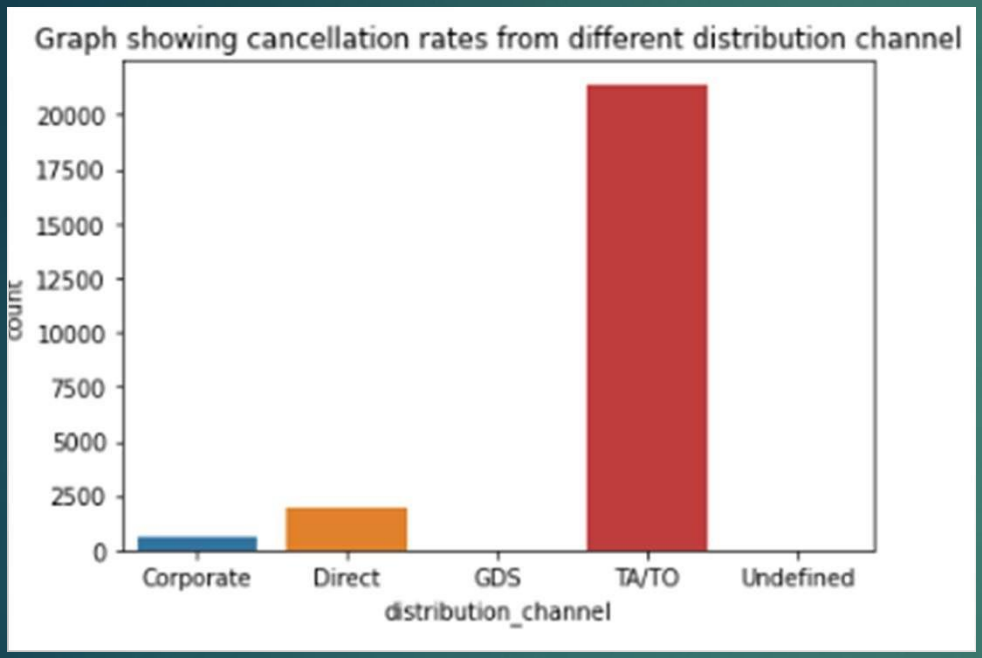
Q2- How many booking were cancelled?



- approximately 25% of booking were cancelled.

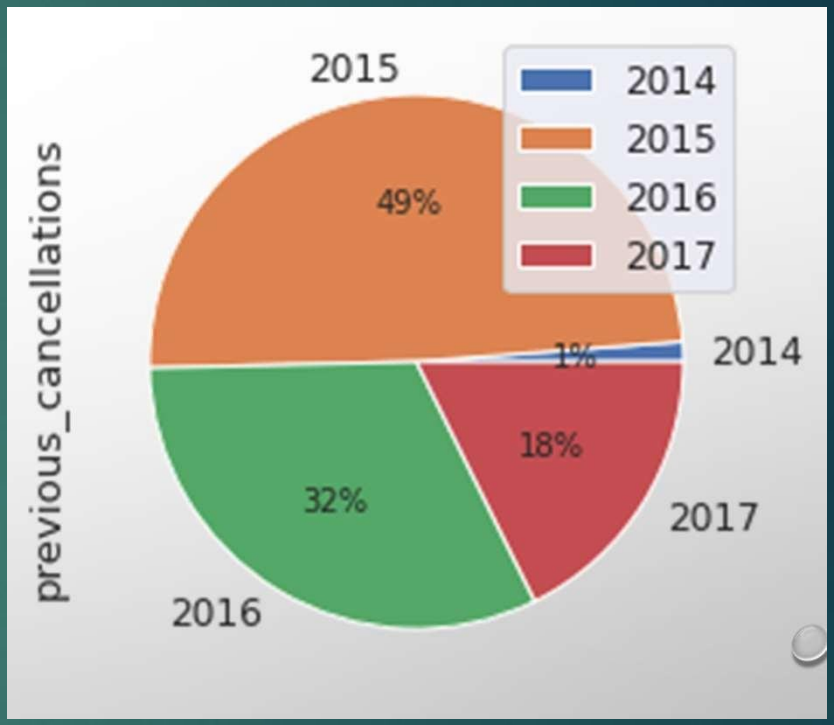
Booking Cancellation Analysis

Q 3-Which significant distribution channel has highest Cancellations? cancellation percentage?



- Maximum cancellation have been observed by Travel Agent.

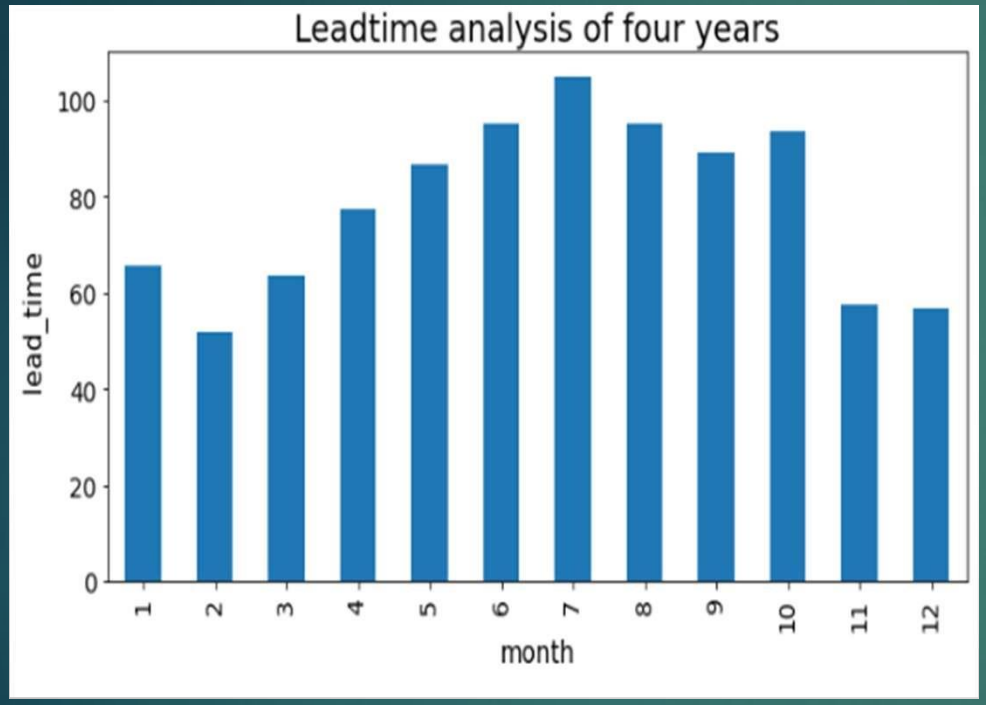
Q 4- What is Yearwise Percentage of



- Maximum cancellation has been done in year 2015.

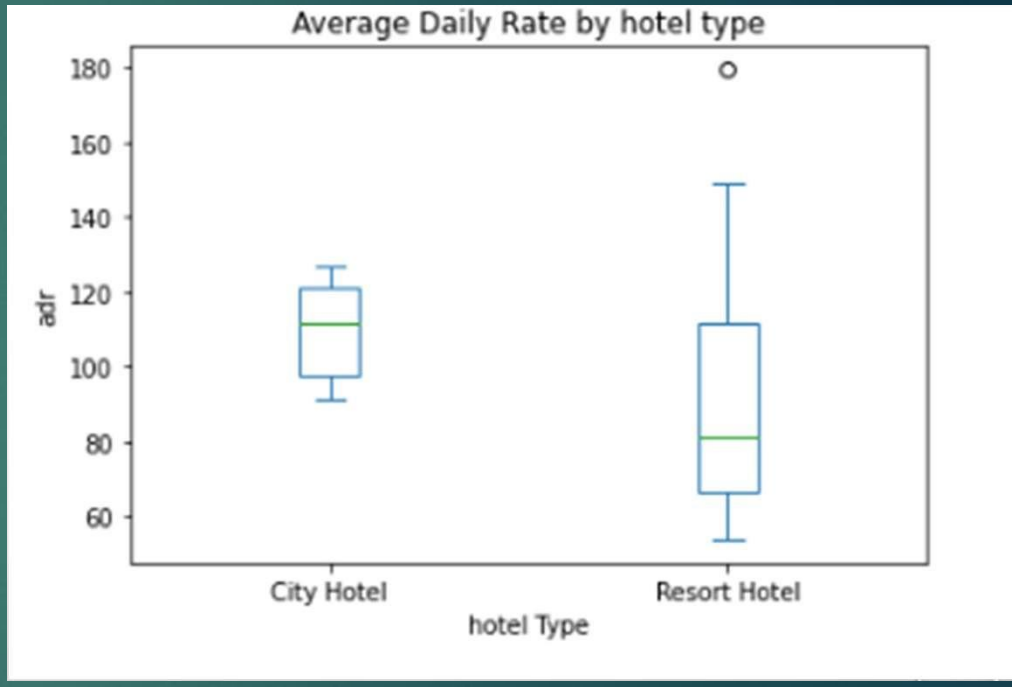
Booking Cancellation Analysis

Q 5- Which period of year has longer lead time analysis?



- In month of August, longer lead time can be seen.

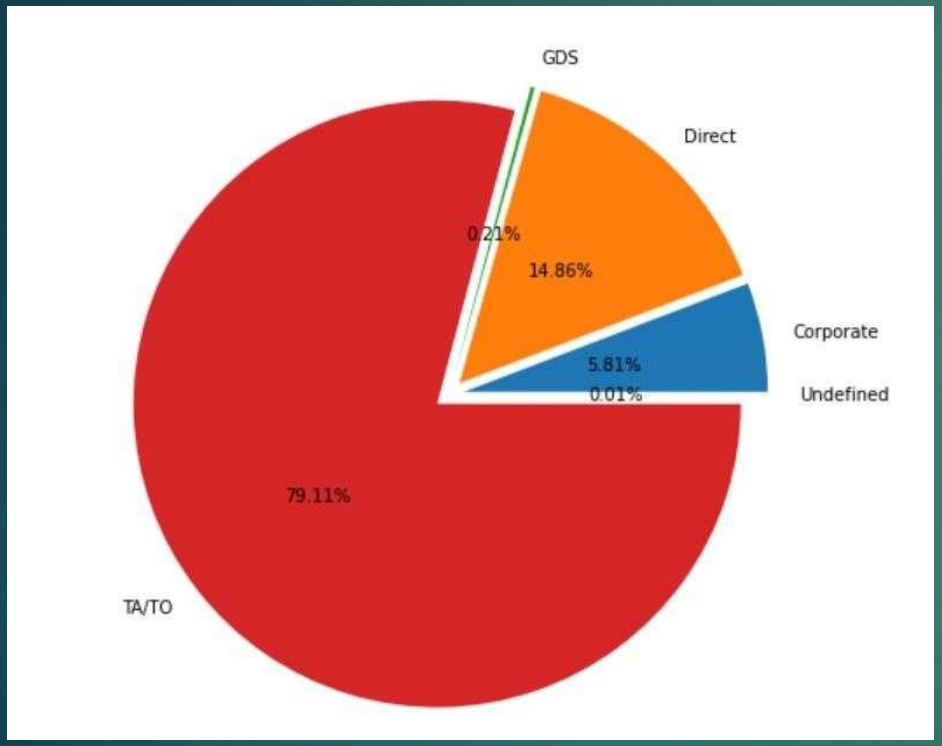
Q 6- What is Average Daily Rate Yearwise and Monthwise?



- Average daily rate is more in Resort hotel than City Hotel.

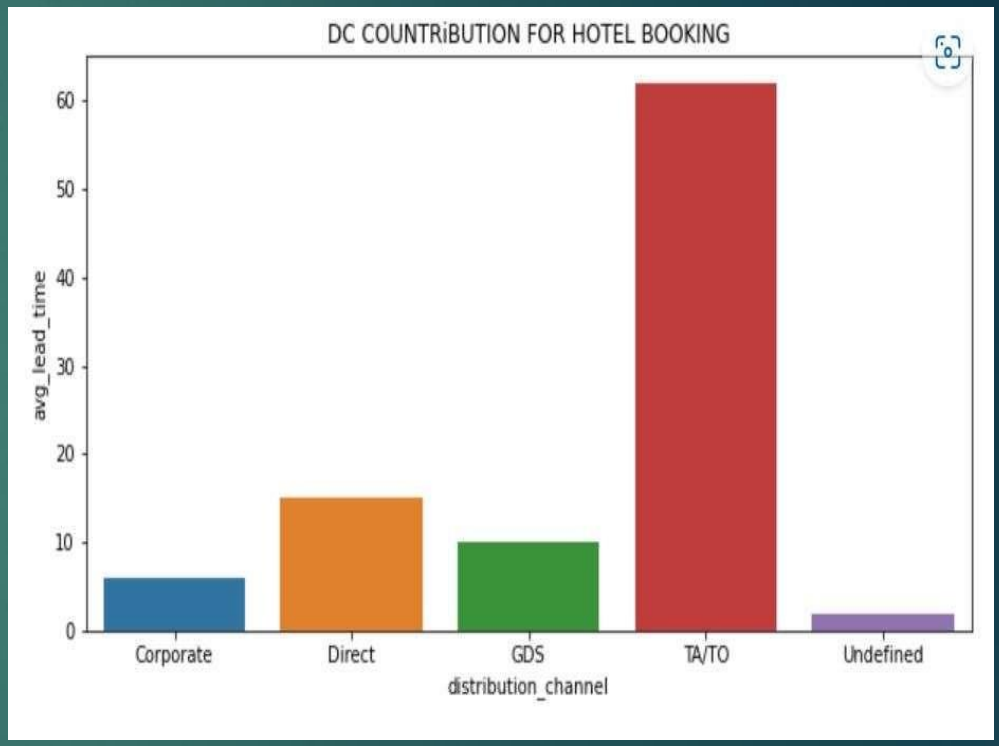
Distribution channel wise Analysis:

Q1- Which Distribution Channel is contributing in most of the hotel bookings?



- Highest Booking received by the hotels are through TA/OT so they are one of the most trusted booking provider.

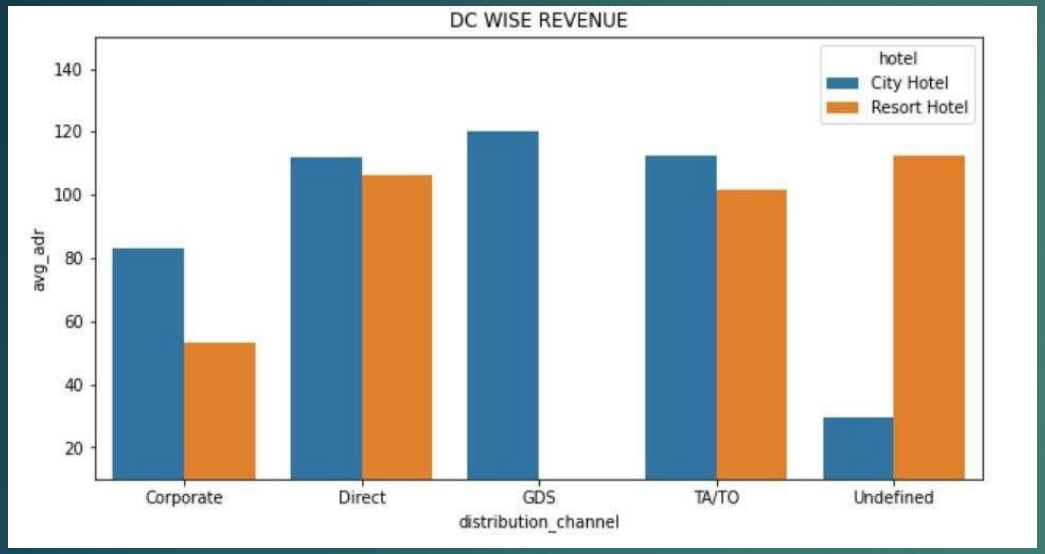
Q2- Which channel is contributing most for early booking of the hotel?



- Most of the bookings we have received from TA/TO.

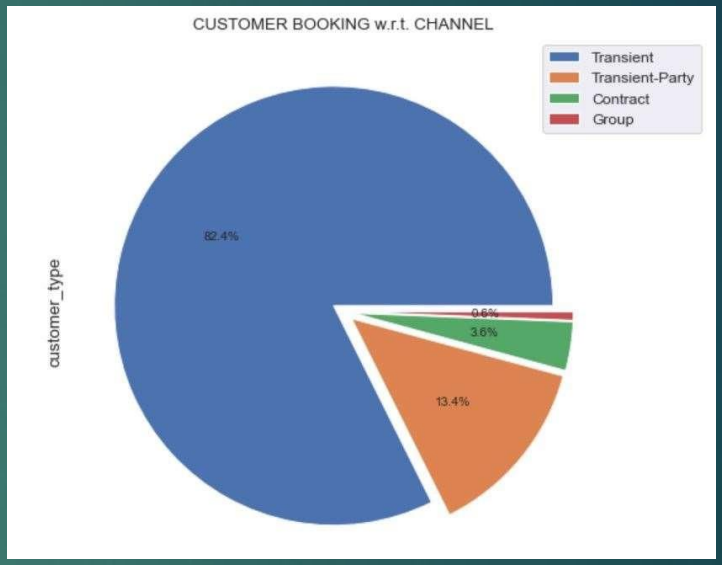
D istribution channel wise A nalysis:

Q3- Which distribution channel brings better revenue generating deals for hotels?



- GDS is the most revenue generating Channel but its only for City hotel. For Resort Hotel its contribution is negligible as compared to other channels distribution.
- Undefined can be associated to multiple channel distribution channels whose data is not provided so after undefined bookings TA/TO are generating most revenue for the Resort Hotel.

Q4- Which is the most favorable Channel for Customers to book hotel?



- The majority of booking channel is from Transient and Transient Party having 82.4% and 13.4% contribution respectively.
- Transient parties are somewhere linked to Transient Group.

Customer centered analysis

Q1- **Which type of hotel is better choice for families?**

```
[ ]  hotel_df['adults'].groupby(hotel_df['hotel']).describe()
```

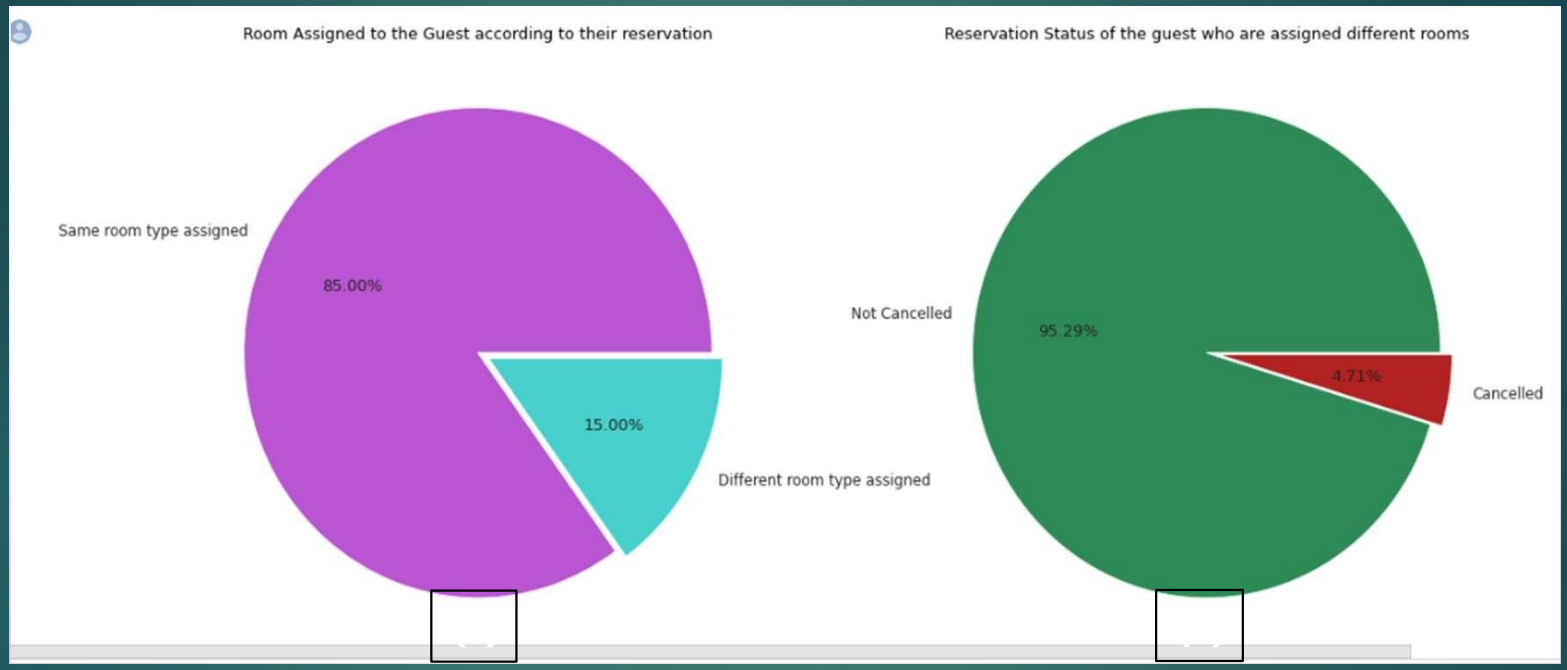
hotel	count	mean	std	min	25%	50%	75%	max
City Hotel	79330.0	1.850977	0.509292	0.0	2.0	2.0	2.0	4.0
Resort Hotel	40060.0	1.867149	0.697285	0.0	2.0	2.0	2.0	55.0

```
hotel_df['children'].groupby(hotel_df['hotel']).describe()
```

hotel	count	mean	std	min	25%	50%	75%	max
City Hotel	79326.0	0.091370	0.372177	0.0	0.0	0.0	0.0	3.0
Resort Hotel	40060.0	0.128682	0.445195	0.0	0.0	0.0	0.0	10.0

Customer centered analysis

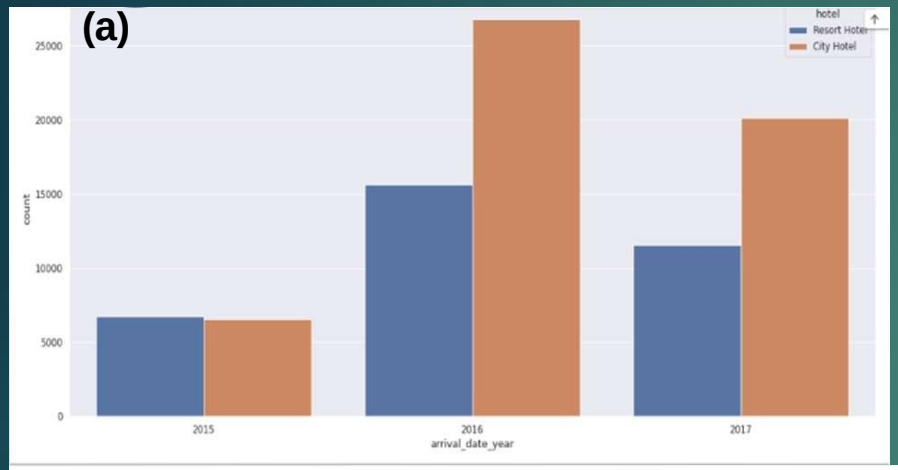
Q2 - How not getting same room as reserved affects adr?

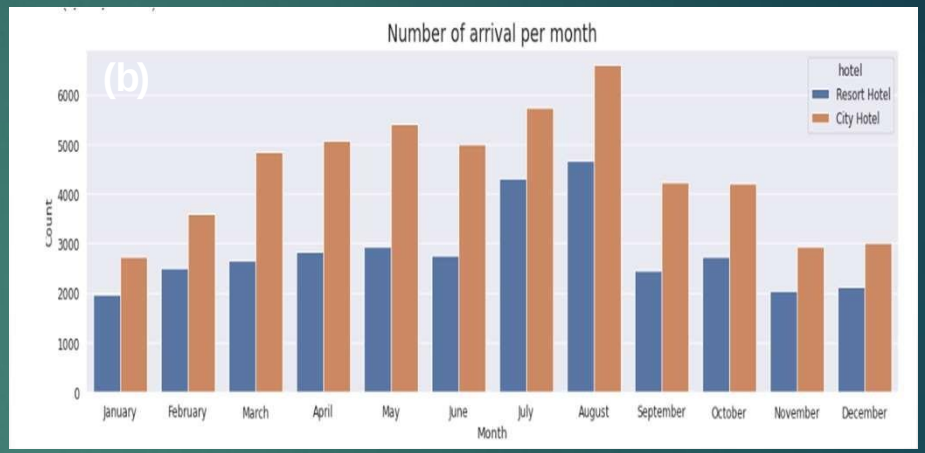


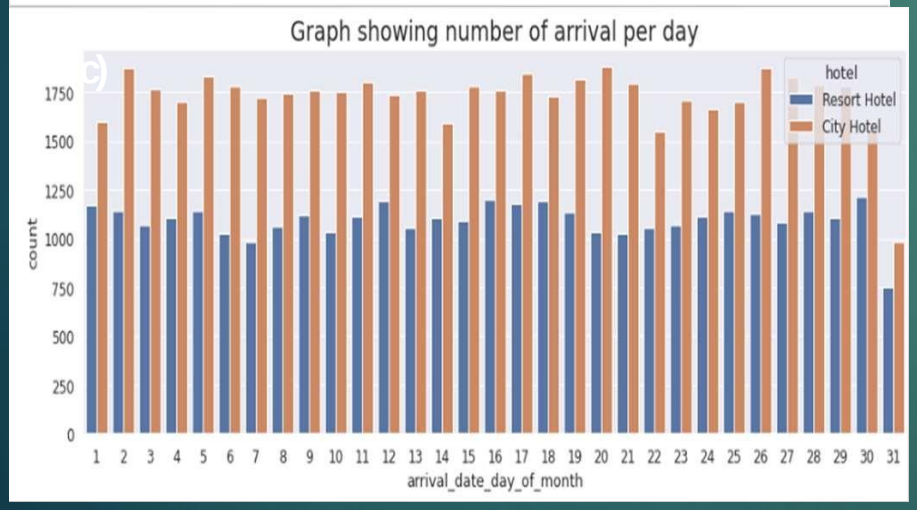
- Figure (a) depicts the distribution of room assigned to the guest according to their reservation and result shows that 85% people got same room type which is assigned and 15% customers got different room. Figure (b) shows reservation status of the guest who are assigned different room and result shows that only 4.71% customers were cancelled their reservation.

Customer centered analysis

Q3- Overview of arrival period



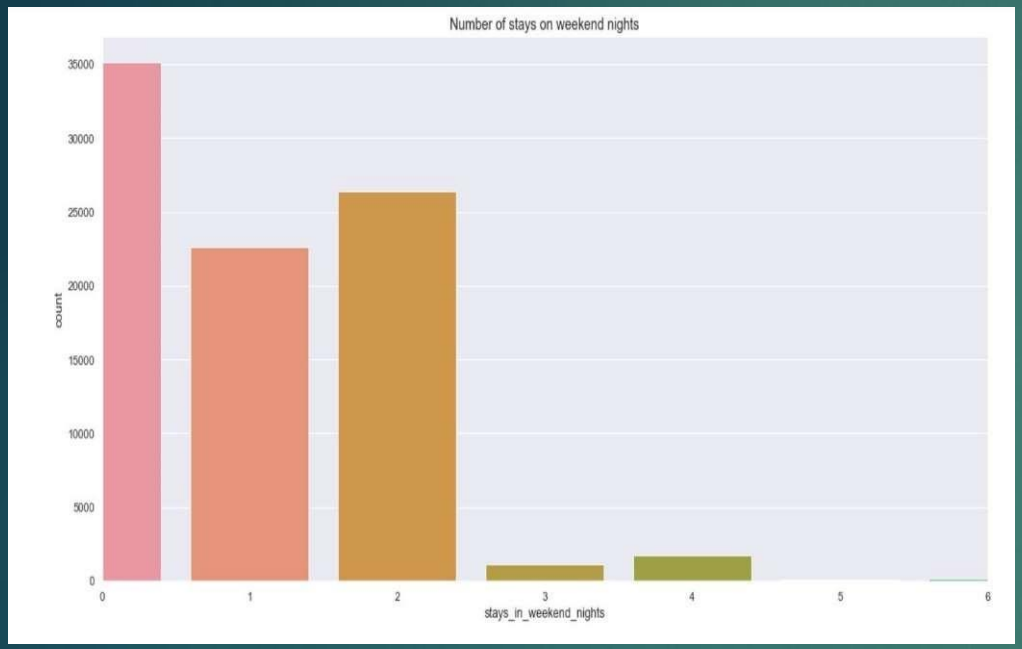


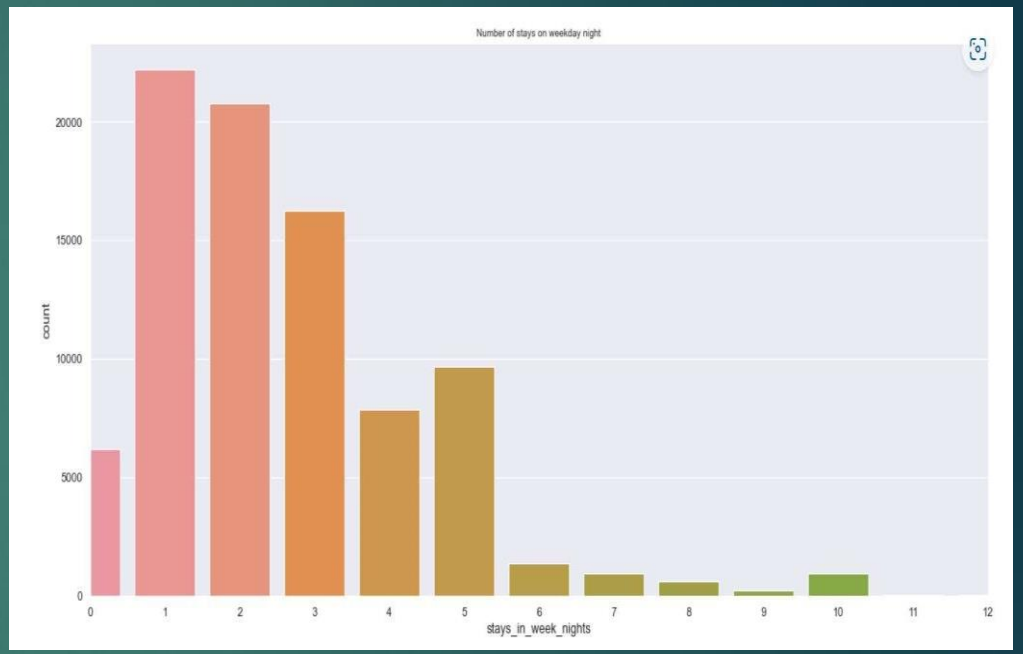


- Year-wise, month-wise and day-wise hotel booking data are represent in Figures(a, b & c), respectively. It is clearly visible in figure(a), highest booking in city hotel as well as resort hotel were in 2016. From figure(b) depicts highest booking in July and August. Summer ends in Aug followed by autumn, so it seems that summer period is a peak for hotel booking. From figure(c) shows trend for the arrival day of month has been roller coaster.

Customer centered analysis

Q4 - Whether Stay is over a weekend or weekday?

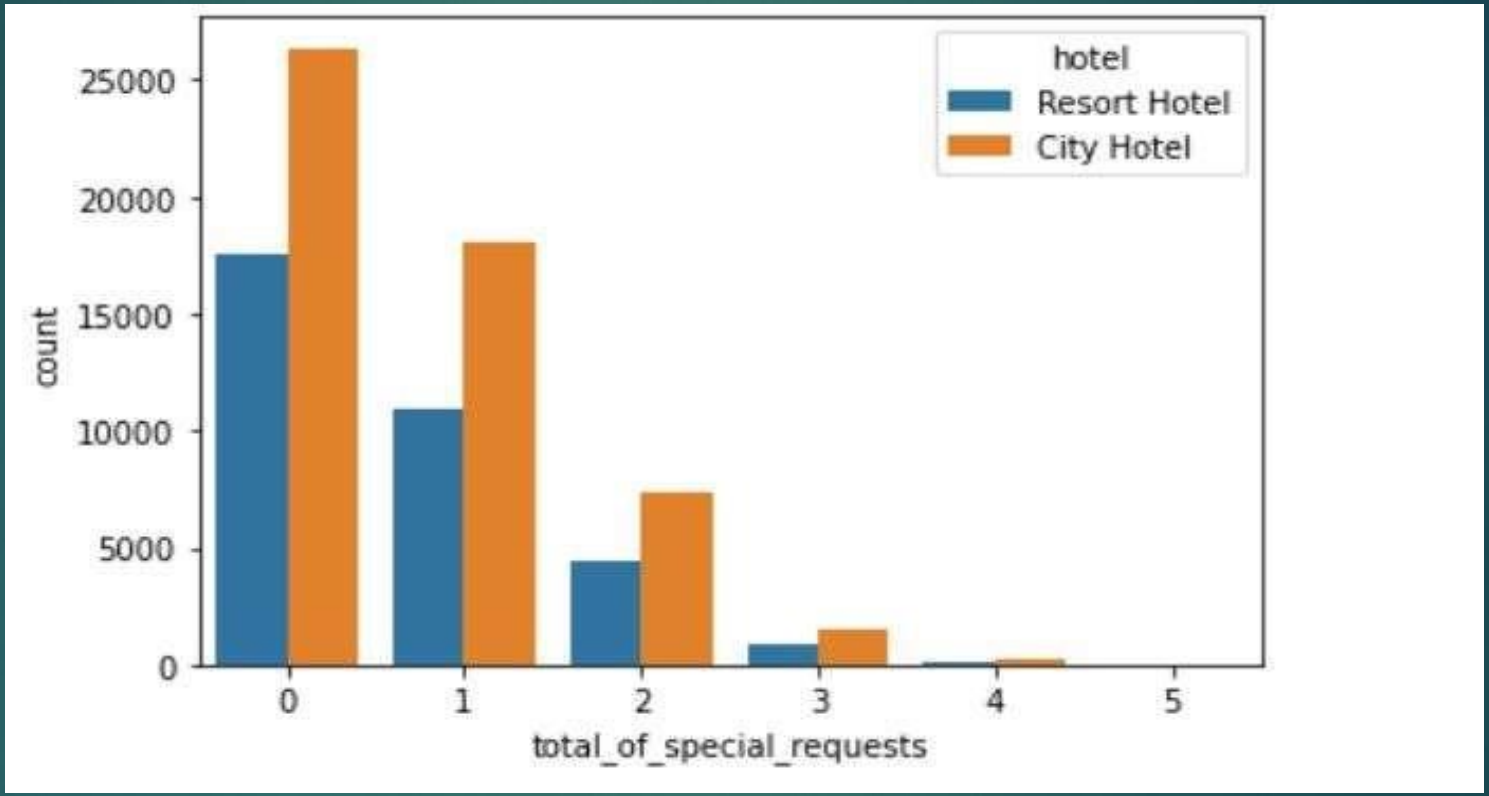




- Majority of the stays are over the weekday's night. Most of the people preferred to stay for 2-4 Nights during the weekdays however there are some bookings whose stay was as long as of 10 days. Few of the customers have checked out on the same day that's why it's showing 0 in the duration.

❖ Special Requests

Q1- Prediction of whether or not a hotel was likely to receive a disproportionately high number of special requests?



- More of special request is for city hotel.

Conclusions

- We have successfully cleaned our data in term of replace **duplicates, Null values, and convert inappropriate datatype to appropriate datatype.** We have successfully find out the relevant insights from this dataset as follow:
- **Univariate Analysis:**
  - ✓ Type 'A' room and BB(Bed & Breakfast) meal have been preferred by most of customers followed by room type 'D' and meal type HB/SC, respectively.
  - ✓ Since 98.7 % of the guests prefer No deposit type of stay. The high rate of cancellations can be due to high no deposit policies. About 91.6 % customers did not required the parking space and approximately 8.3 % customers required only 'one' parking space.
- **Hotel wise Analysis:**
  - ✓ City Hotel is most preferred by customers and significantly longer waiting time. Thus city hotels have slightly more revenue but much busier than Resort Hotel.
  - ✓ The time taken between when a customer makes a reservation and their actual arrival is called the lead time. Resort hotels have slightly high avg lead time, i.e., customers plan their trips very early to reserve resort hotels.
  - ✓ Most common stay length is less than 4 days and generally people prefer city hotel for shorter stay, but for longer stay res ort hotel is preferred.
  - ✓ Resort hotel has higher retention rate compare to city hotel that means customers are willing to stay again in resort hotel and highest number of customers are from Western Europe, namely Portugal, UK and France. So, marketing team have to be target these regions.

Conclusions

- **Booking Cancellation Analysis:**
  - ✓ Maximum cancellation has been observed by Travel Agent in 2015 and City hotels have higher booking cancellation rate of approximately 25%.
  - ✓ In month of August, longer lead time can be seen, i.e., customers had been booking their rooms so early and average daily rate is more in Resort hotel than city hotel.

- **Distribution channel wise Analysis:**
  - ✓ Highest Booking received by the hotels are through TA/OT so they are one of the most trusted booking provider. Thus most of the bookings we have received from TA/TO.
  - ✓ GDS is the most revenue generating channel but its only for city hotel. For resort hotel its contribution is negligible as compared to other channels distribution.
  - ✓ Undefined can be associated to multiple distribution channels whose data is not provided so after undefined bookings from TA/TO are generating most revenue for the Resort Hotel.
  - ✓ The majority of booking channel is from Transient and Transient Party having 82.4% and 13.4% contribution respectively. Transient parties are somewhere linked to Transient Group.

Conclusions

➢ Customer centered analysis:

- ✓ Figure (a) depicts the distribution of room assigned to the guest according to their reservation and result shows that 85% people got same room type which is assigned and 15% customers got different room. Figure (b) shows reservation status of the guest who are assigned different room and result shows that only 4.71% customers were cancelled their reservation.
- ✓ Year-wise, month-wise and day-wise hotel booking data are represent in Figures(a, b & c), respectively. It is clearly visible in figure(a), highest booking in city hotel as well as resort hotel were in 2016. From figure(b) depicts highest booking in July and August. Summer ends in Aug followed by autumn, so it seems that summer period is a peak for hotel booking. From figure(c) shows trend for the arrival day of month has been roller coaster.
- ✓ Majority of the stays are over the weekday's night. Most of the people preferred to stay for 2-4 Nights during the weekdays however there are some bookings whose stay was as long as of 10 days

➢ Special Requests:

- ✓ More of special request is for city hotel as most of families prefer city hotel.

References

YouTube
W3 schools
Pandas libraries
Numpy libraries
Stackoverflow
AlmaBetter Class material