Capstone Project 1

EDA Hotel Booking Analysis

Submitted by:
Shubham.A.Digrase

Points Of Discussion :

- Problem Statement.
- Data Summary.
- Types of customers w.r.t the hotels.
- Repeated and New Customers.
- Visitors from countries.
- Type of customers make change in bookings.
- Percentage of bookings in City and Resort hotel.
- Which hotel has higher lead time?.
- Total number of canceled Bookings by hotel type.
- Total number of Bookings & Cancellations through market segments?
- Which year had the highest bookings?
- Which hotel has longer waiting time?
- Cancellation percentage of bookings.
- Percentage of repeated guests.
- The Percentage Distribution of Deposit type.

Points Of Discussion continue.....

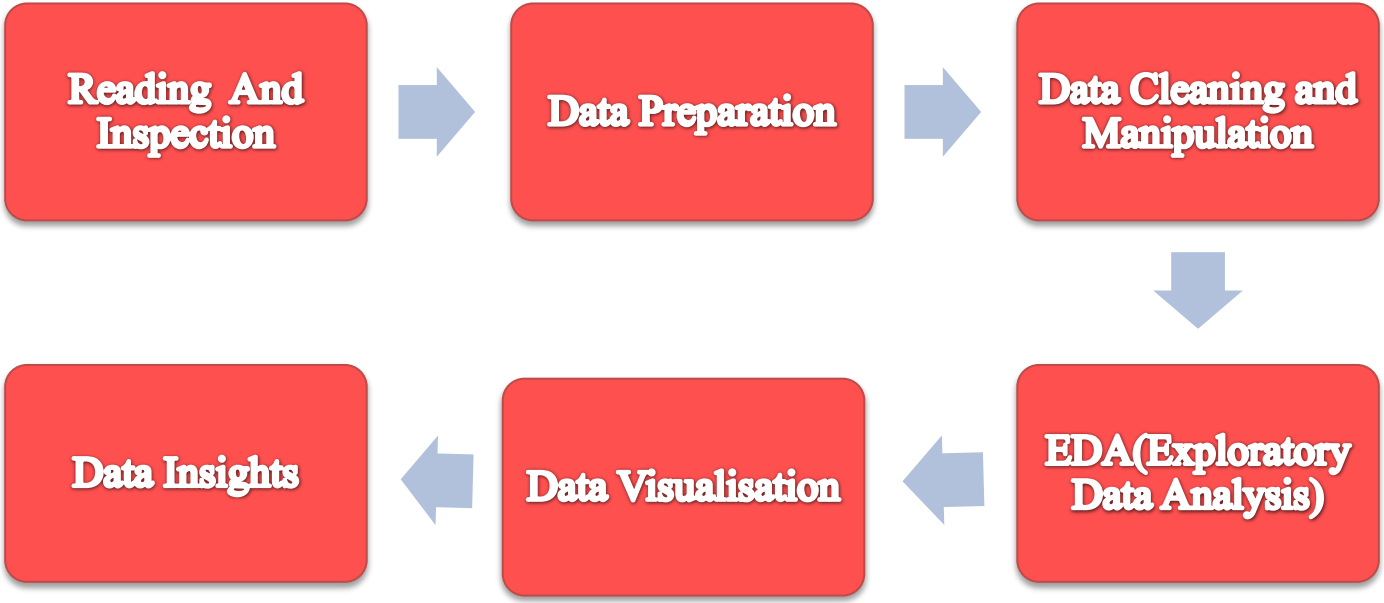
- Type of food is mostly preferred by the guest.
- Most preferred room type by the customers.
- Which month most of the bookings happened?
- Distribution channel is mostly used for hotel bookings.
- Which distribution channel has the highest cancellation rate?
- Which Hotel type has the highest ADR(Average Daily Rate)?
- ADR across different market segment.
- Total of Special requests with respect to customer.
- How long do people stay at the hotels?
- Requirement of parking spaces.
- Correlation Of Heat map.
- Conclusion.
- Solution to business objective.

Problem Statement

➢The hotel industry is one of the most important components of the wider service industry, catering for customers who require overnight accommodation. It is also closely associated with the travel industry and the hospitality industry. For this project we are analyzing hotel booking data of a city hotel and a resort hotel of few years. The information includes the booking time, canceled bookings, room and meal type, customers stay time, available parking spaces, visitors, cancellation cases.

➢The major goal of this project is to investigate and analyze data in order to identify significant elements that influence reservations and provide knowledge to hotel management so that it may implement various campaigns to increase sales and performance.

Work Flow


Reading And Inspection
Data Preparation
Data Cleaning and Manipulation
EDA(Exploratory Data Analysis)
Data Visualisation
Data Insights

Data Collection and understanding dataset input

Field	Description
hotel	City and Resort hotel
is_canceled	indicating booking cancelled (1) or not canelled (0)
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week no of year for arrival date
arrival_date_day_of_month	day of arrival date
stays_in_weekend_nights	no of weekends night
stays_in_week_nights	no of week nights
adults	no of adults
children	no of children
babies	no of babies
meal	Kind of meal opted for
country	country of origin
market_segment	Which segment the customer belongs to
distribution_channel	How the customer accessed the stay- corporate booking/Direct/TA.TO
is_repeated_guest	Guest coming for first time or not

Data Collection and understanding dataset input

Field	Description
Previous cancellations	Was there a cancellation before
Previous bookings	Count of previous bookings
reserved_room_type	Type of reserved room
assigned_room_type	Type of assigned room
booking_changes	Count of changes made to booking
deposit_type	Deposit type
agent	Booked through agent
days_in_waiting_list	Number of days in waiting list
customer_type	Type of customer
adr	Average daily rate
required_car_parking_spaces	If car parking is required
total_of_special_requests	Number of additional special requirements.
reservation_status	Reservation of status
reservation_status_date	Date of the specific status

Dataset Input data summary

Numeric

- 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, adr, required_car_parking_spaces, total_of_special_requests

,

Binary

- is_canceled, is_repeated_guest

Categorical

- Hotel, arrival_date_month, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type

Data Collection and understanding dataset input

Prerequisites

- ✓ Import Python libraries.
- ✓ Mount google drive to google colab
- ✓ Authorize notebook to access google drive files

Understanding dataset input

- ✓ Find out the total columns and rows of dataset
- ✓ Find the data type of each column.
- ✓ Find individual distribution for some of the columns
- ✓ Also check the correlation between dependent columns

Data cleaning and manipulation

- ✓ Extract the unique values of each column content from the hotel booking dataset.

  Dataset size : 119390 rows × 32 columns

- ✓ Identify duplicated rows and remove the same.

  Dataset size : 87396 rows × 32 columns.

- ✓ Replace NaN values with 0 for heading Agent & company
- ✓ Replace NaN values with their mean values for heading children
- ✓ Replace NaN values with 'others' for heading Country
- ✓ Modify datatype from float to int64 for heading Agent, Company, Children

Data cleaning and manipulation

- The columns of company, agent, country, children has missing values.

```
[22]  1 # Handling missing values
      2 df_hotel_bookings.isnull().sum().sort_values(ascending = False)[:6]
```

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

```
[23]  1 #We will replace null values by 0 in these columns:
      2 df_hotel_bookings[['company','agent']] = df_hotel_bookings[['company','agent']].fillna(0)
```

```
2 df_hotel_bookings['children'].fillna(0 , inplace = True)
```

```
2 df_hotel_bookings['country'].fillna('others', inplace = True)
```

```
1 # Check all null values are removed.
2 df_hotel_bookings.isnull().sum().sort_values(ascending = False)[:6]
```

```
hotel                        0
previous_cancellations       0
reservation_status_date      0
reservation_status           0
total_of_special_requests    0
required_car_parking_spaces  0
dtype: int64
```

Data Cleaning and Manipulation

•Duplicate values are 31994 .we dropped it from the data

```
[12]  1 # Dataset Duplicate Value Count
      2 len(hotel_df[hotel_df.duplicated()])
```

```
31994
```

```
[13]  1 # Dropping duplicate values
      2 hotel_df.drop_duplicates(inplace = True)
```

```
[14]  1 hotel_df.shape
```

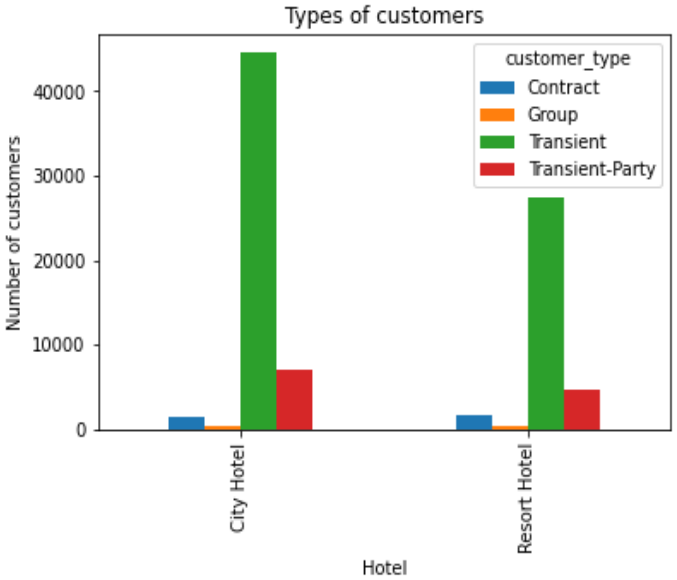
```
(87396, 32)
```

•We create new column : total_nights = stay in weekend nights and week nights.

```
1 df_hotel_bookings['total_nights'] = df_hotel_bookings['stays_in_weekend_nights'] + df_hotel_bookings['stays_in_week_nights']
```

EDA (Exploratory Data Analysis)

•**Types of customers w.r.t the hotels**.

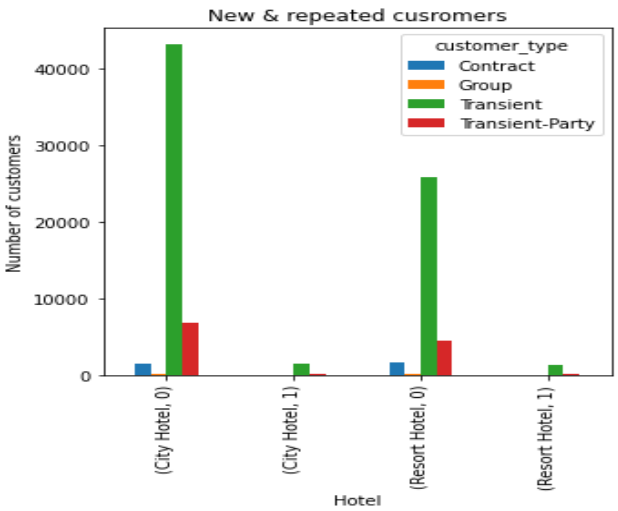


Findings:

•From the above bar plot we find that both the City Hotel & Resort Hotel are attracting 'Transient' type of customers the most followed by 'Transient-Party' type. Whereas, 'Group' type of customers are the least attended by both the hotels.

EDA (Exploratory Data Analysis)

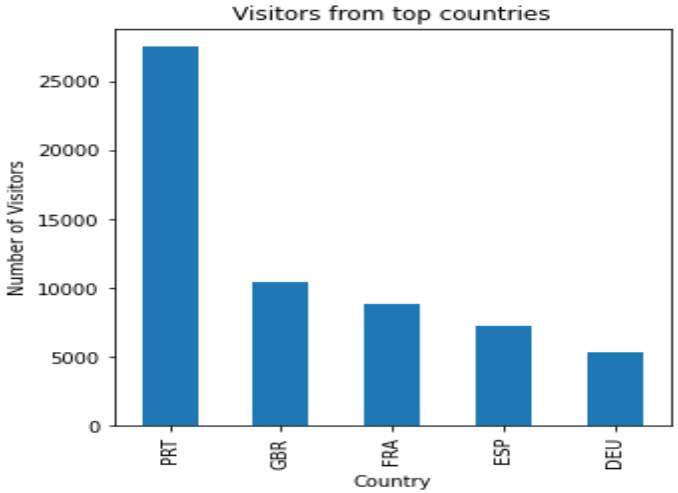
•Repeated and New Customers



Findings:

•The 'Transient Type' of customers are more repeated and the 'Contract Type' customers has less repeated.

•Top Countries customers



Findings:

•From above bar plot chart the most of th e visitors are from Portugal(PRT).

EDA (Exploratory Data Analysis)

•Type of customers make change in bookings

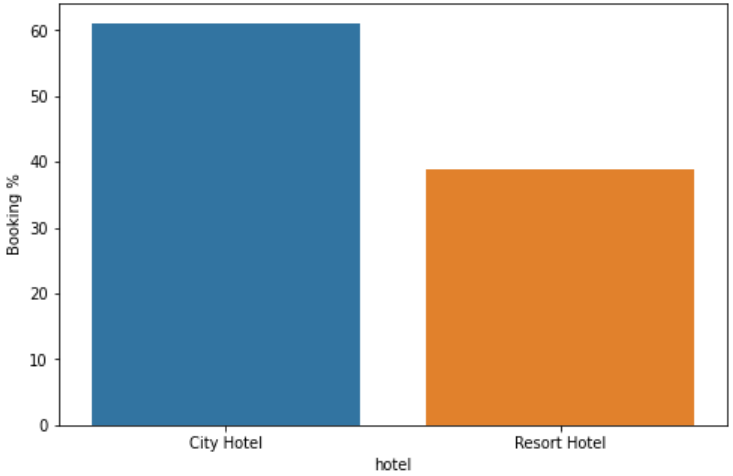
•

Findings:

•The 'Transient Type' customer are making more changes in bookings and the 'Group Type' customer are making less changes in bookings.

EDA (Exploratory Data Analysis)

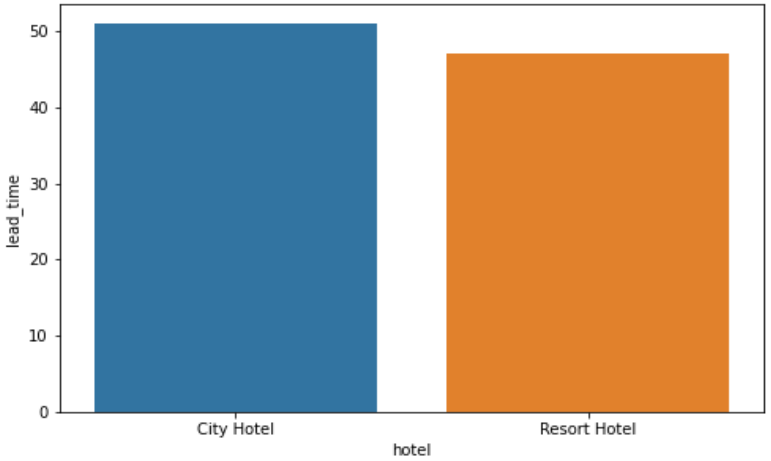
•What is percentage of bookings in City and Resort hotel?



Findings :

•Around 60% bookings are for City hotel and 40% bookings are for Resort hotel.

•Which hotel has higher lead time?

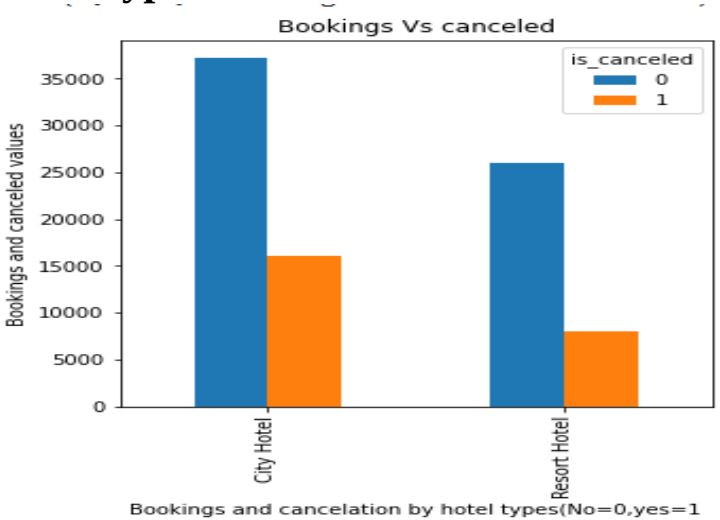


Findings :

•City hotel has more lead time as compared to resort hotel.

EDA (Exploratory Data Analysis)

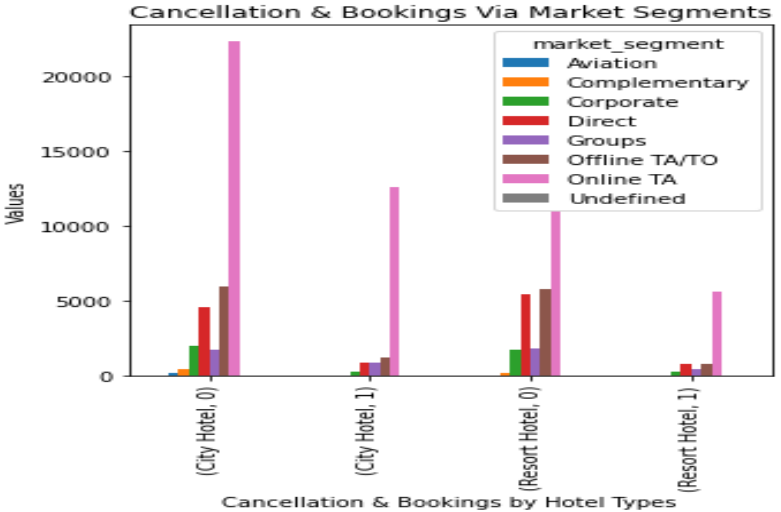
•What is total number of canceled Bookings by hotel type?



Findings:

•'City Hotel' has highest bookings and higher cancellation also as compared to resort hotel.

•What is total number of Bookings & Cancellations through market segments?

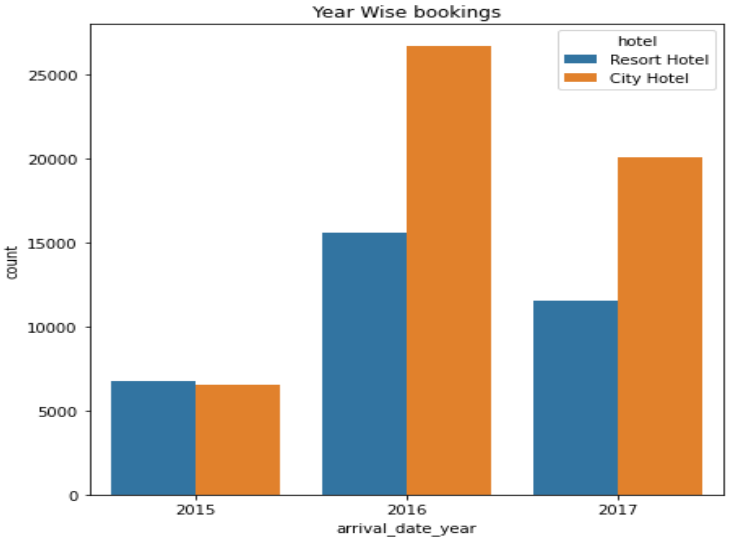


Findings:

•The highest segment from where the booking and cancellation done are Online TA .

EDA (Exploratory Data Analysis)

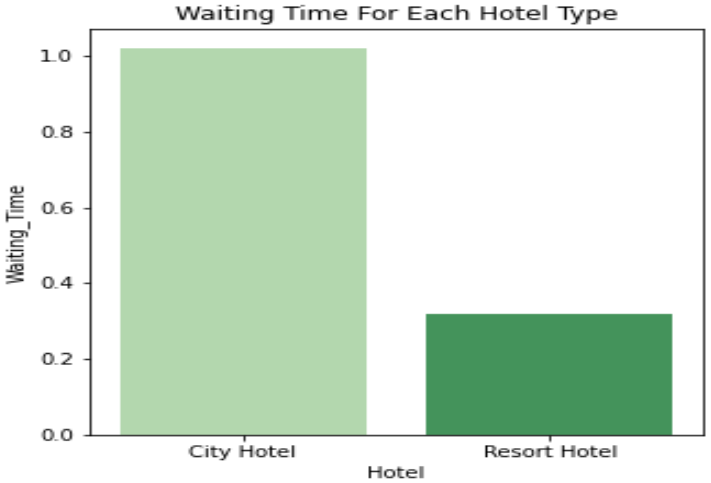
•Which year had the highest bookings?



Findings:

•2016 had the highest booking.
•2015 had the lowest booking.
•Overall City hotels had the most of the bookings.

•Which hotel has longer waiting time?

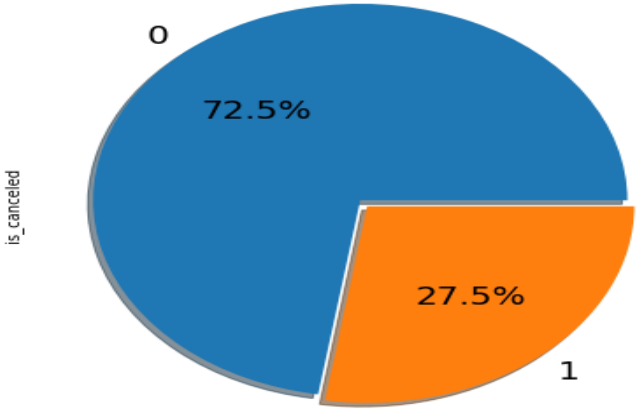


Findings:

• The City Hotels has longer waiting period than the Resort Hotels. Therefore that City Hotels are much busier than the Resort Hotels

EDA (Exploratory Data Analysis)

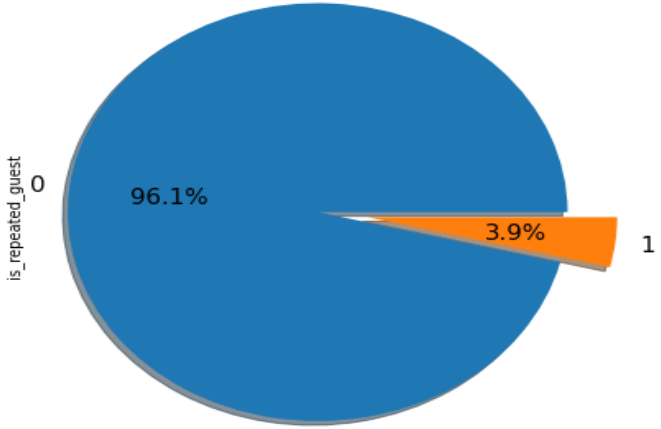
•What is the cancellation percentage of bookings?



Findings:

•27.5% of the bookings were cancelled.

•What is the Percentage of repeated guests?

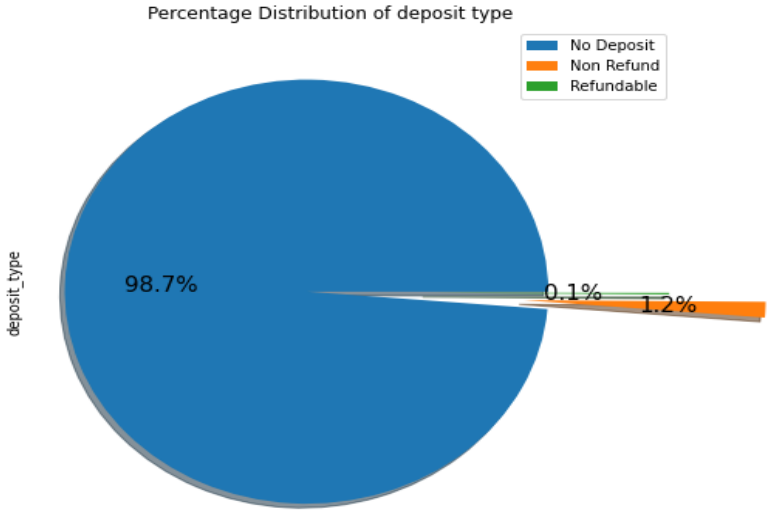


Findings:

•Repeated guests are only 3.9%, so to retain the guests, it is important to take feedback from them.

EDA (Exploratory Data Analysis)

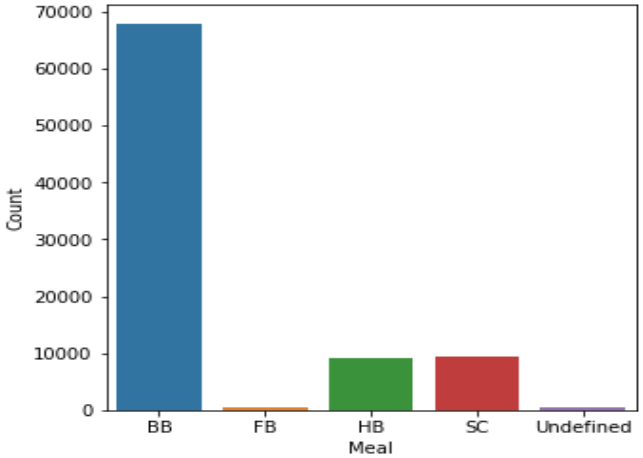
•What is The Percentage Distribution of Deposit type ?



Findings:

•98.7 % of guests prefer "No Deposit" type of deposit.

•Which type of food is mostly preferred by the guests?



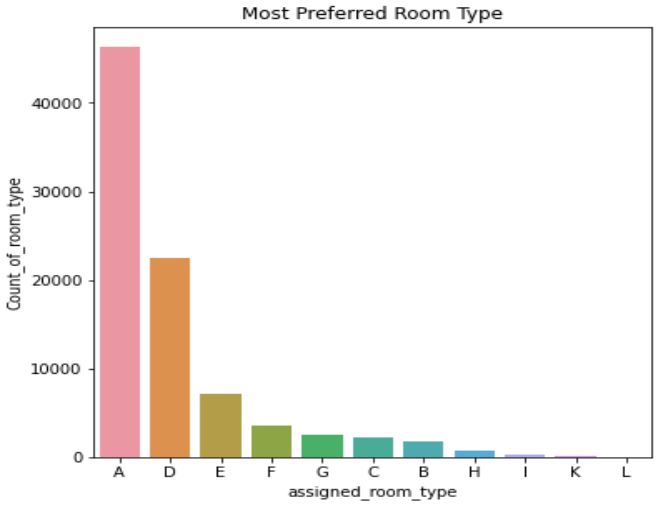
Findings:

• The most preferred meal type by the guests is BB(Bed and Breakfast).

• HB- (Half Board) and SC- (Self Catering) are equally preferred.

EDA (Exploratory Data Analysis)

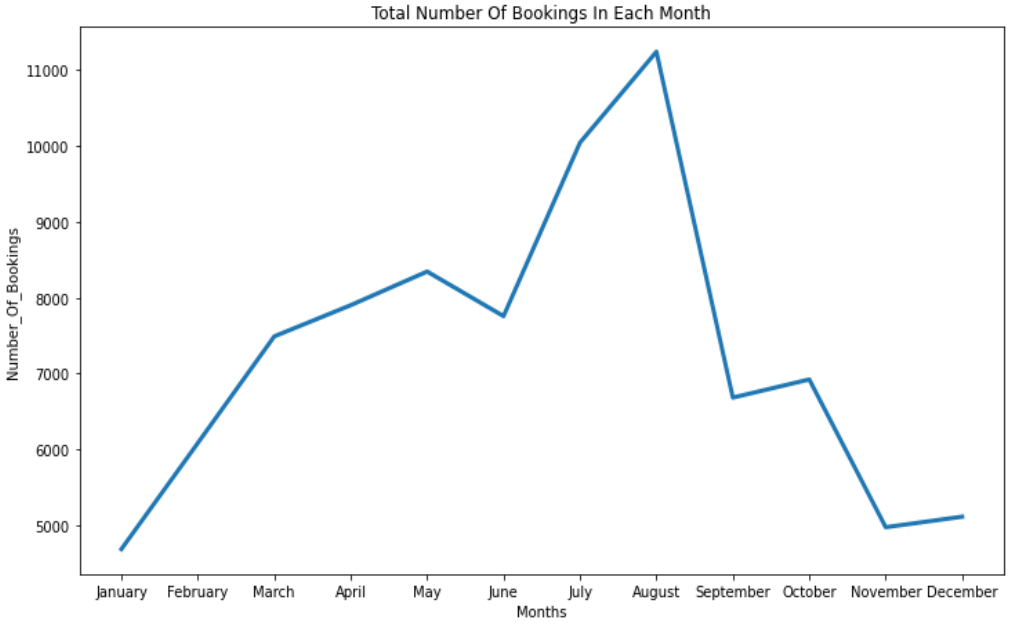
•Which is the most preferred room type by the customers?



Findings:

•Mostly guests prefer to stay in the room type"A".

•In which month most of the bookings happened?

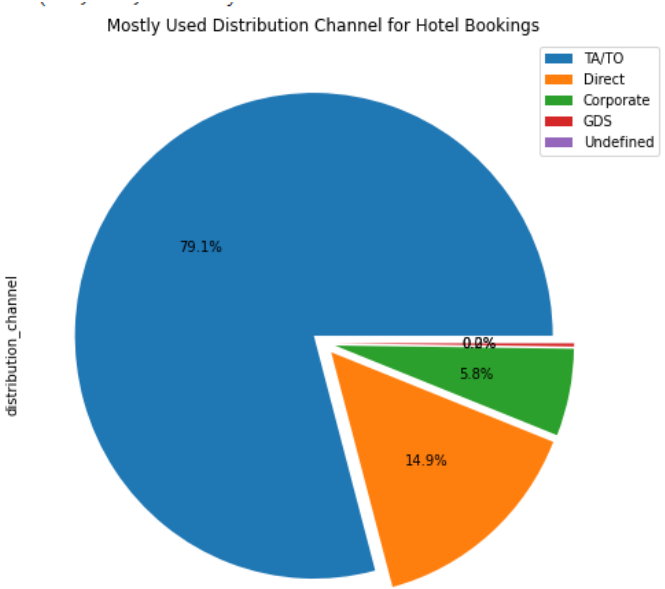


Findings:

•July and August months had the most Bookings.

EDA (Exploratory Data Analysis)

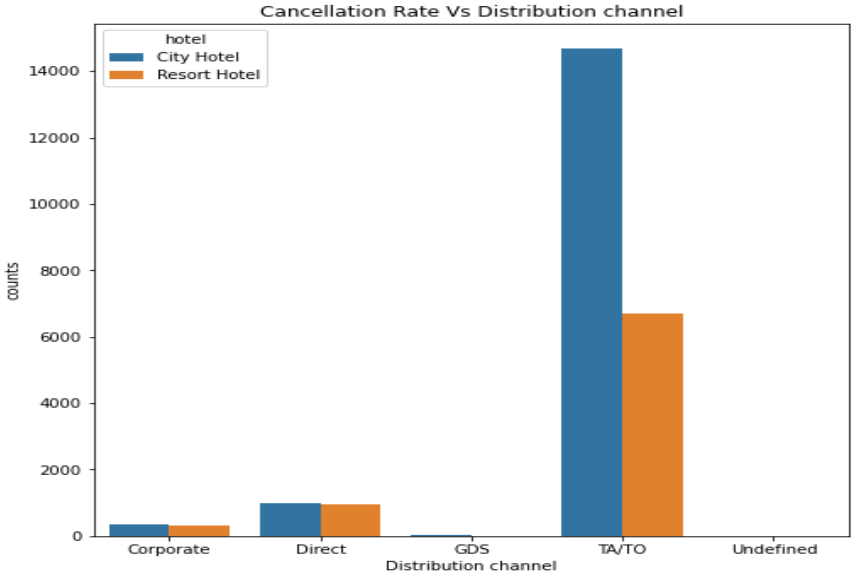
•Which Distribution channel is mostly used for hotel bookings?



Findings:

•The 'TA/TO' is mostly(79.1%) used for booking hotels.

• Which distribution channel has the highest cancellation rate?

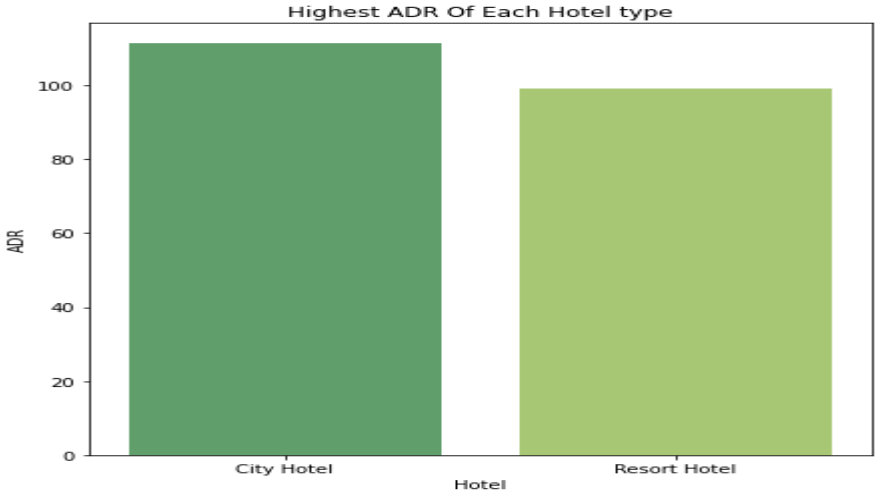


Findings:

•"TA/TO" City hotels has the high cancellation rate compared to resort hotels.

EDA (Exploratory Data Analysis)

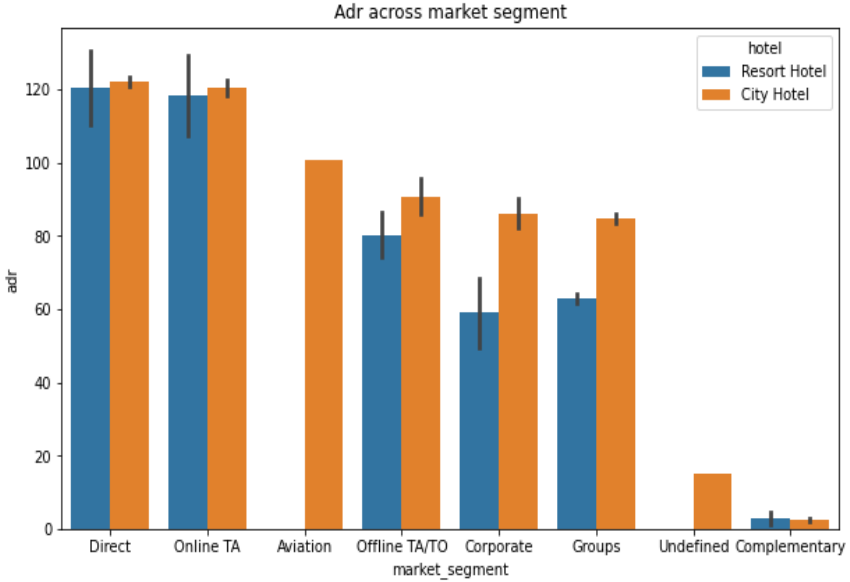
•Which Hotel type has the highest ADR(Average Daily Rate)?



Findings:

•City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels.

•ADR across different market segment.

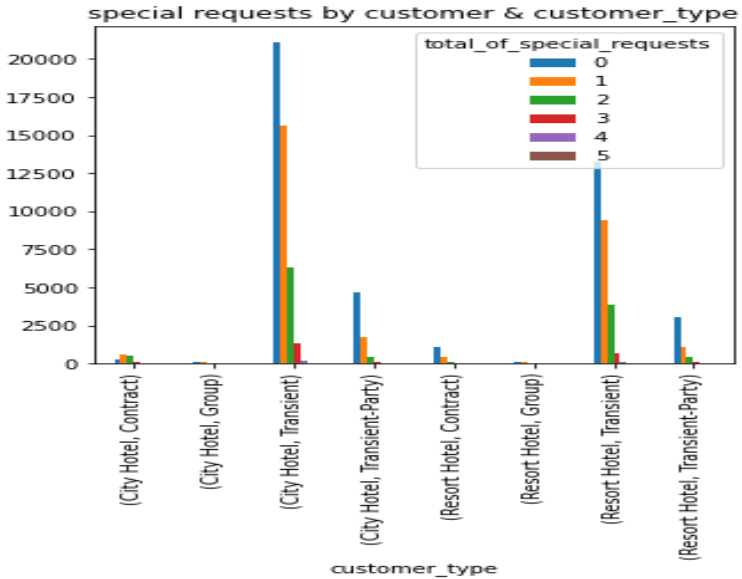


Findings:

• 'Direct' and 'Online TA' contribute the most in both types of hotels.

EDA (Exploratory Data Analysis)

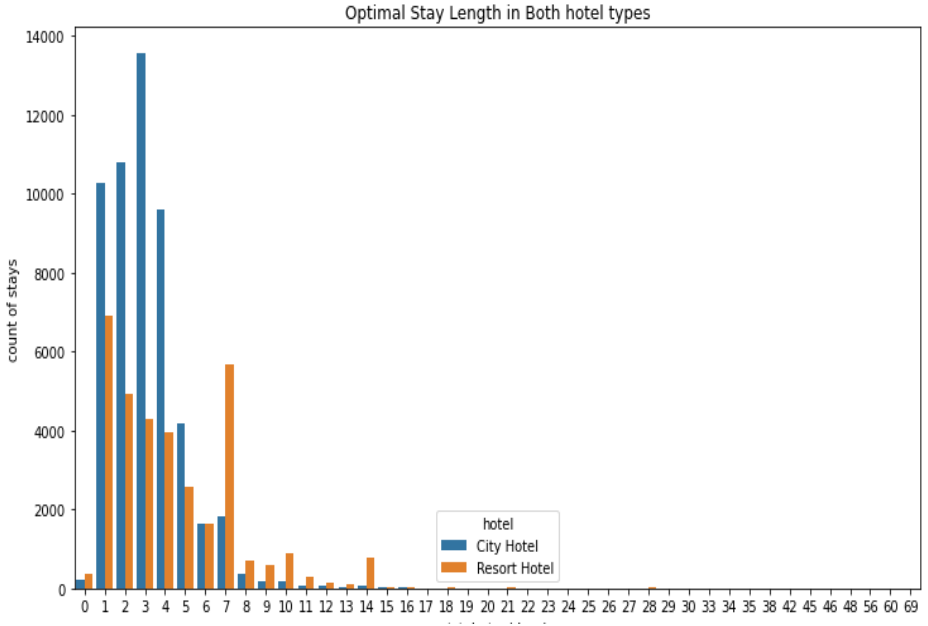
•Finding total of Special requests with respect to customer.



Findings:

•'City Hotel' with 'Transient type' of customer has highest no of special requests.

•How long do people stay at the hotels?

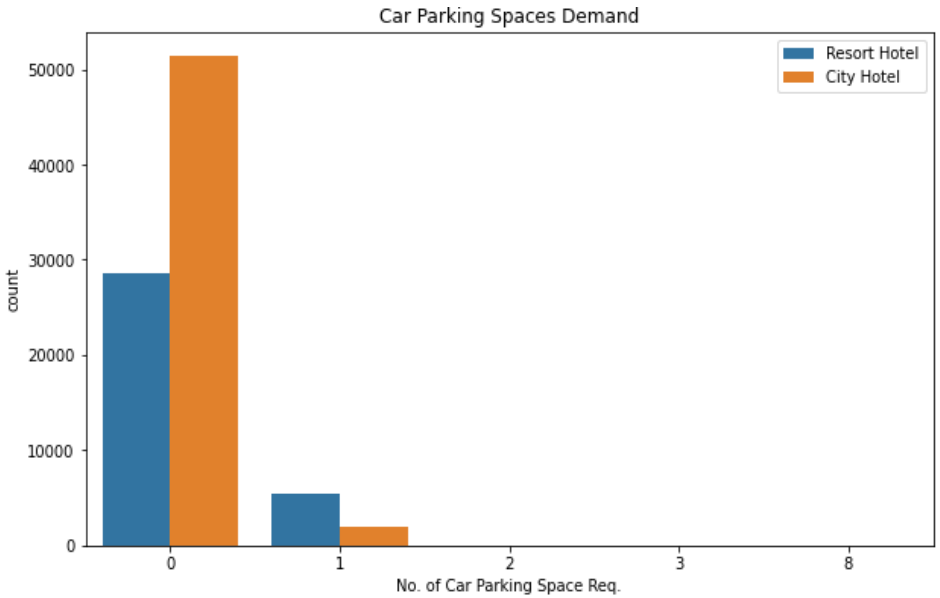


Findings:

•The optimal stay in both types of hotels is less than 7 days.

EDA (Exploratory Data Analysis)

- **Requirement of parking spaces.**

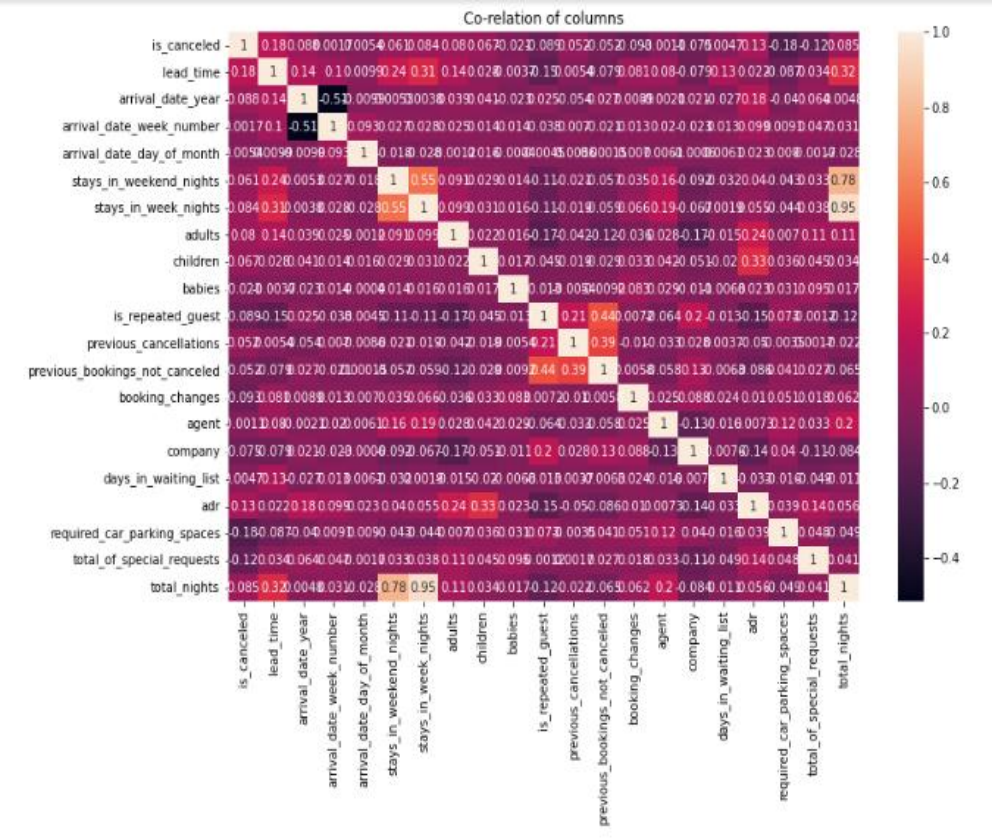


Findings:

- 'City Hotel' and 'Resort Hotel' maximum guests required zero space. Very less guests in both type of hotel required one space for car parking.

EDA (Exploratory Data Analysis)

•Correlation Of Heat map



Findings:

•'is_canceled' and 'total_nights' are correlated to each other.If the customer does stay on nights in hotels, there is no cancelation of bookings.

•'lead_time' and 'total_of_special_requests' are also correlated to each other. As we provide more special requests then the period of time taken by customer is more that is lead time.

•'adr' and 'is repeated guests' are also correlated to each other. More number of repeated guests the more are average daily rate.

Conclusion

1. City hotels are 60% bookings that is most preferred hotel type by the guests
2. City hotel has more lead time as compared to resort hotel.
3. 'Online TA' has most no bookings and cancellations in market segment.
4. 'City Hotel' has highest no of cancellations as compared to resort hotel.
5. Most no of bookings are in year 2016 and lowest bookings are in year 2015.
6. City Hotels has longer waiting period than the Resort Hotels. Therefore that City Hotels are much busier than the Resort Hotels.
7. 27.5% of the bookings were cancelled.
8. Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
9. 98.7 % of guests prefer "No Deposit" type of deposit.
10. BB(Bed & Breakfast) is the most preferred type of meal by the guests.
11. Most guests prefer to stay in the room type "A".
12. July and August months had the most no Bookings in hotel.
13. 79.1 % bookings were made through TA/TO (Travel agents/Tour operators).
14. "TA/TO" City hotels has the high cancellation rate compared to resort hotels and in "Direct" both the hotels has almost same cancell ation rate.
15. ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
16. Maximum number of guests were from Portugal country.
17. 'Direct' and 'Online TA' contribute the most in both types of hotels.
18. 'City Hotel' with 'Transient type' of customer has highest no of special requests.
19. The optimal stay in both types of hotels is less than 7 days.
20. The 'City Hotel' and 'Resort Hotel' maximum guests required zero space.Very less guests in both type of hotel required one space for car parking.

Solution For Business:

- Create a Loyalty Program
- Provide Excellent Customer Service
- Promote the unique benefits of resort hotels
- Ask for Feedback
- Utilize Social Media

Challenges:

- Dataset contains a lot of duplications.
- Against few columns having a lot of Null values.
- Few dataset columns with wrong data type format.

Thank You