EDA Capstone Project

Hotel Booking Analysis

Prajwal D U

Hotel Business Industry

The hotel industry is one of the most important components within the service industry, catering for customers,
who require overnight accommodation,
It is closely associated with the travel industry and the hospitality industry, although there are notable differences in scope.

INTRODUCTION

- This data set comprises booking information for a city hotel and a resort hotel, including dates of booking, duration of stay, number of adults, children, and/or newborns, and available parking spots, among other things. This dataset has removed all personally identifiable information.
- To gain understanding from the data, we will use Python to perform exploratory data analysis.

POINTS FOR DISCUSSION

- Data Summary
- Data Cleaning and Preparation
- Which year hotels have highest footfall ?
- Top 10 countries from where guests visits the hotels
- Booking according to market segment
- Price per night guest pays to stay in hotels
- Booking relation between Resort hotel and City hotel
- Top 20 agents with highest booking
- Booked accommodation as per Single,Couple and Family or Friends
- Busiest month for hotels
- Booking cancellation as per month
- ADR(Average Daily Rate) over months of year
- Accommodation tenure
- Hotel wise accommodation tenure
- Correlation matrix
- Conclusion
- References

DATA SUMMARY (Column Information)

- **Hotel**
  - Resort hotel
  - City hotel
- **is_canceled**
  - 1: Canceled
  - 0: Not canceled
- **lead_time**
  - No of days that elapsed between entering date of booking into property management system and arrival date
- **arrival_date_year**
  - Year of arrival date (2015-2017)
- **arrival_date_month**
  - Month of arrival date (Jan - Dec)
- **arrival_date_week_numberr**
  - Week number of year for arrival date (1-53)
- **arrival_date_day_of_month**
  - Day of arrival date

- **stays_in_weekend_nights**
  - No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel
- **stays_in_week_nights**
  - No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel
- **Adults**
- **Children**
- **Babies**
- **meal**
  - Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
- **country**
- **market_segment** (a group of people who share one or more common characteristics, lumped together for marketing purposes)
  - TA: Travel agents
  - TO: Tour operators
- **distribution_channel** (A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)
  - TA: Travel agents
  - TO: Tour operators
- **is_repeated_guest** (value indicating if the booking name was from repeated guest)
  - 1: Yes
  - 0: No

- **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. Code is presented instead of designation for anonymity reasons.
- **assigned_room_type**
  - Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- **booking_changes**
  - Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
- **deposit_type**
  - Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
- **agent**
  - ID of the travel agency that made the booking
- **company**
  - ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons

- **day_in_waiting_list**
  - Number of days the booking was in the waiting list before it was confirmed to the customer
- **customer_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
- **adr (average daily rate)**
  - average daily rate = Sum Of All Lodging Transaction/Total Number Of Staying Night
- **reservation_status**
  - Canceled – booking was canceled by the customer;
  - Check-Out – customer has checked in but already departed;
  - No-Show – customer did not check-in and did inform the hotel of the reason why
- **reservation_status_date**
  - Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel
- **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)

Data Cleaning and Preparation

	index	Null Values	Percentage Null Values
0	hotel	0	0.000000
1	is_canceled	0	0.000000
2	lead_time	0	0.000000
3	arrival_date_year	0	0.000000
4	arrival_date_month	0	0.000000
5	arrival_date_week_number	0	0.000000
6	arrival_date_day_of_month	0	0.000000
7	stays_in_weekend_nights	0	0.000000
8	stays_in_week_nights	0	0.000000
9	adults	0	0.000000
10	children	4	0.003350
11	babies	0	0.000000
12	meal	0	0.000000
13	country	488	0.408744
14	market_segment	0	0.000000
15	distribution_channel	0	0.000000
15	distribution_channel	0	0.000000
16	is_repeated_guest	0	0.000000
17	previous_cancellations	0	0.000000
18	previous_bookings_not_canceled	0	0.000000
19	reserved_room_type	0	0.000000
20	assigned_room_type	0	0.000000
21	booking_changes	0	0.000000
22	deposit_type	0	0.000000
23	agent	16340	13.686238
24	company	112593	94.306893
25	days_in_waiting_list	0	0.000000
26	customer_type	0	0.000000
27	adr	0	0.000000
28	required_car_parking_spaces	0	0.000000
29	total_of_special_requests	0	0.000000
30	reservation_status	0	0.000000
31	reservation_status_date	0	0.000000

Data Cleaning and Preparation

Data Cleaning:

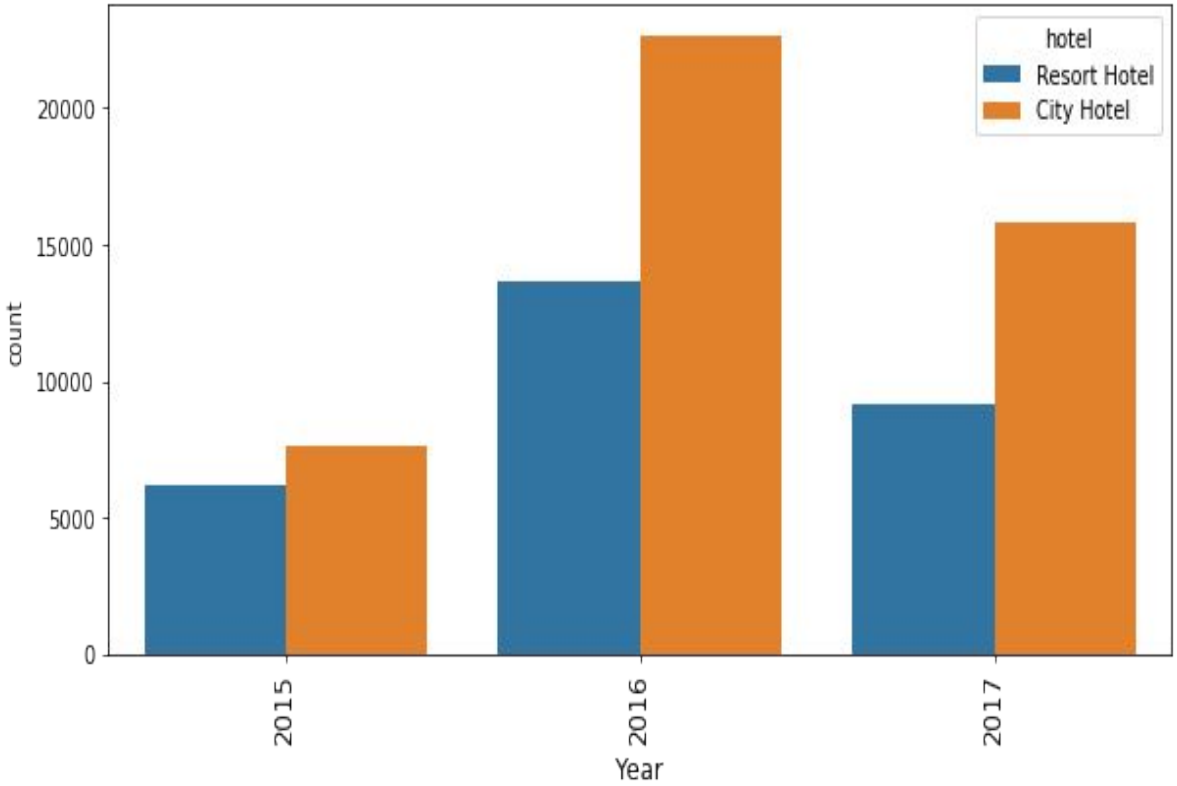
- As we check the dataset we have 4 features that contains missing values, those are company, agent, country and children column.
- More than 94% null values in 'company' column we need to drop that column to avoid further errors in analysis.
- Similarly column 'Agent' have 13 percent null values, whenever using 'Agent' column in future we will drop that null rows or we can replace null values by zero '0'.
- We can also see some values that are null in 'country' column we can replace them by Other Countries
- In our dataset there are some rows in which adult,children and babies have value 0 at same time which is not possible , so we will drop that rows.

EDA

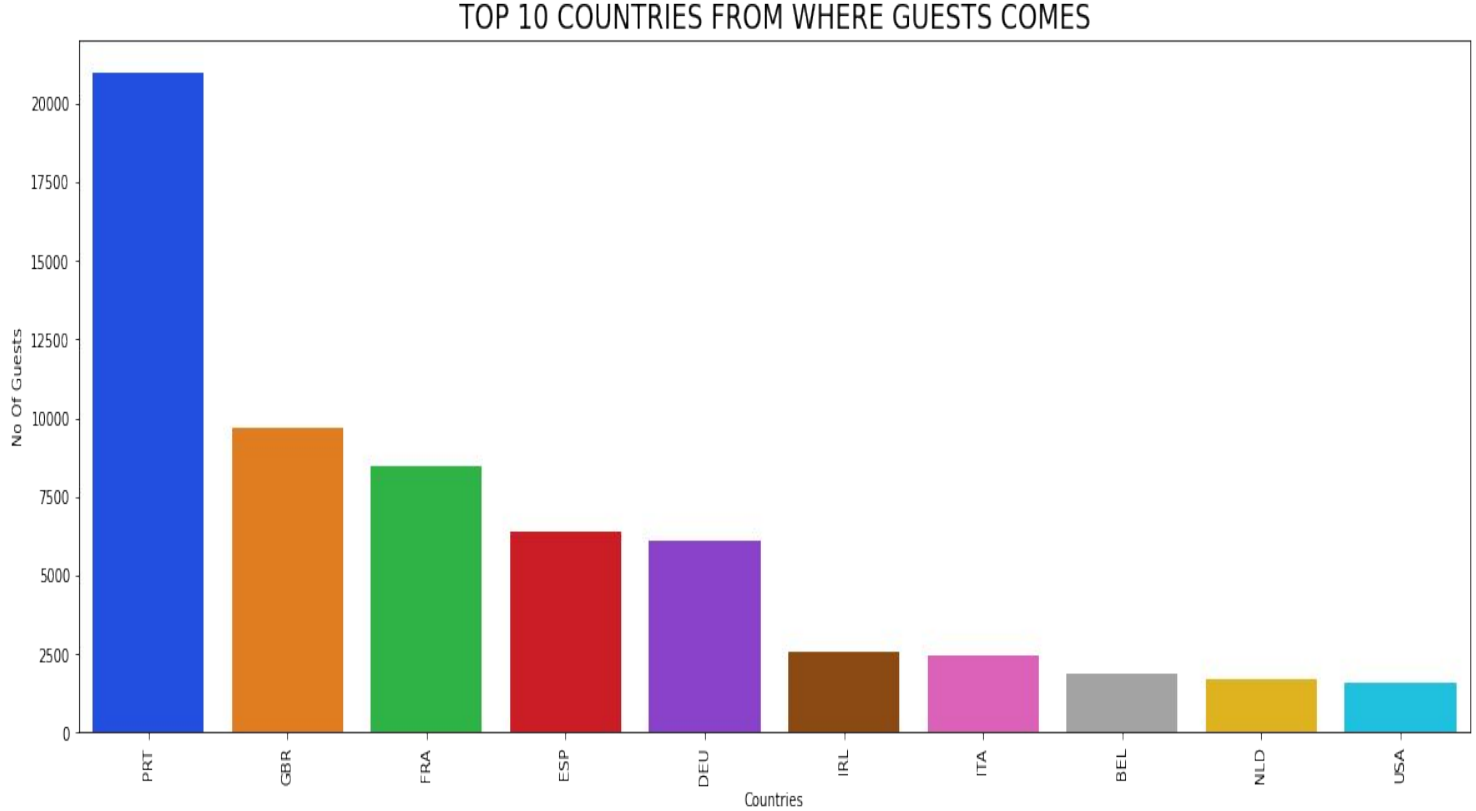
- After complete the data cleaning part and data Preparation part we jump into EDA part
- Now let's do the fun part, extract the information from our data and try to answer some questions.

Number of Bookings Across Different Years

- Most Number of the bookings are done in the year 2016 following 2017 and 2015 for the dataset given comprising the data for these 3 years.
- When, sub plotted, we can see that City hotel has high demand rather than Resort hotels due to its reliability and price difference compared to resort hotels.



Top 10 countries from travelers and tourists visit the hotels

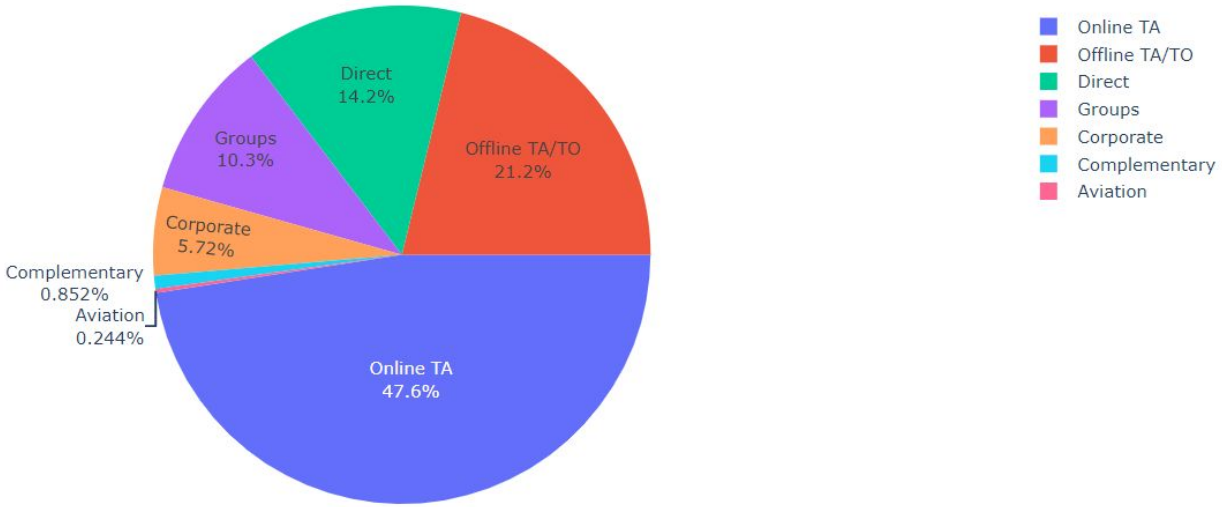


Travelers and Tourists from Top 10 countries visited the hotels

- After analyzing the dataset, we found that Portugal top the position with 48590 customers followed by UK with 12129, France with 10415, Spain with 8568 and Germany with 7287 customers.
- USA sits back with least number of customers among the top 10 Countries.

What type of Bookings is majorly seen in the hotels

BOOKING ACCORDING TO MARKET SEGMENT



- Online Travel agency segment gives the high amount of leads for the hotel booking than that of any other sources of Market segments.
- We can report that we need to target our marketing area on online TA websites or apps and focus majorly on online TA.
- The following majority market segments are offline travel agencies, groups and direct customers.

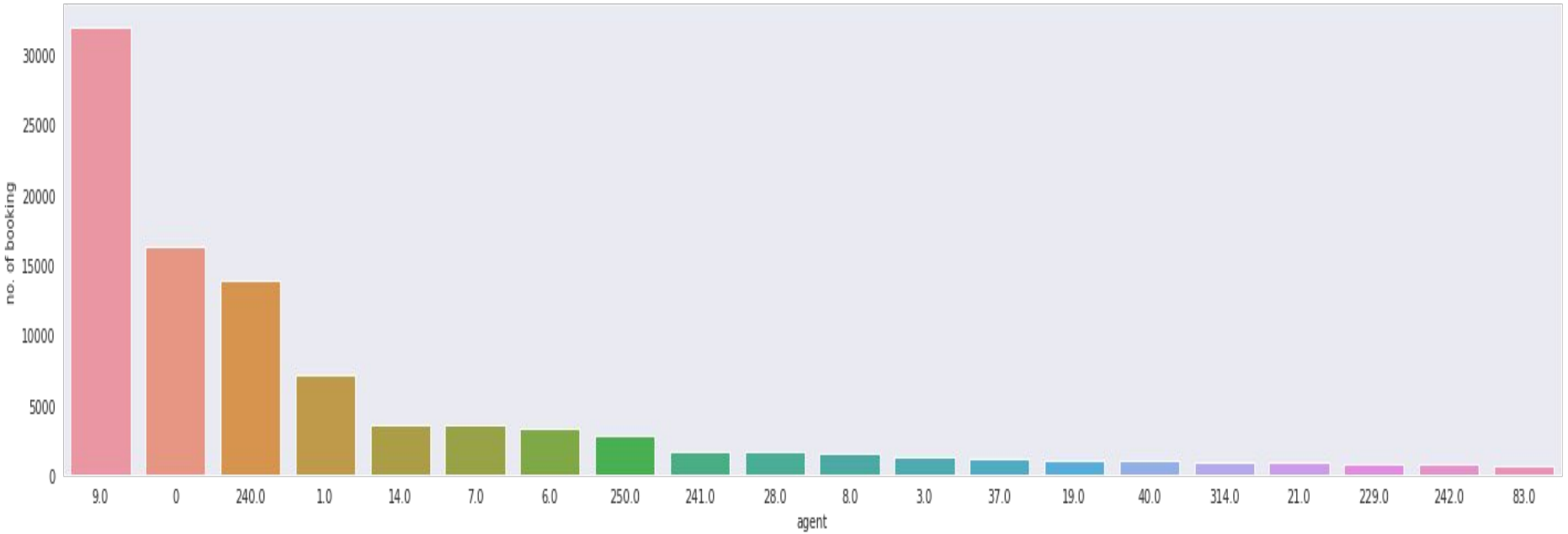
Customer pay for accomodation per night

Above Box plot shows reserved rooms type with respect to ADR(Average Daily Rate), here we see in A,D,G,E,F,H and B has outliers, Yellow shows Resort hotel and blue shows City hotel.

Relation between Resort Hotel and City Hotel?

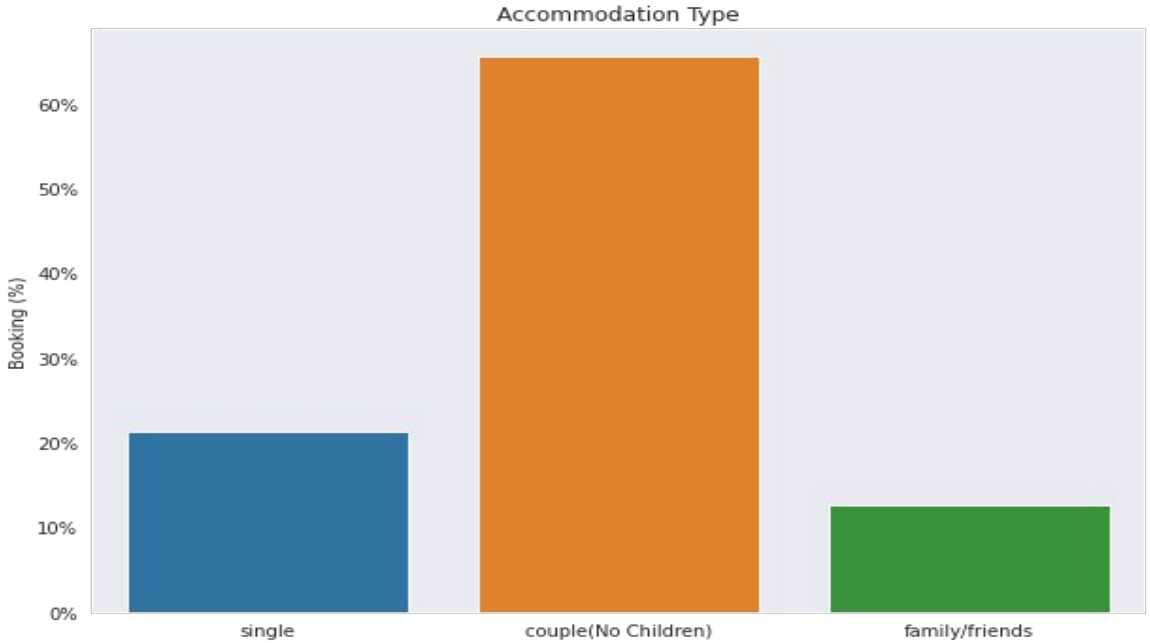
As we see graph blue shows city hotel and orange shows resort hotel, More than 60% of the population booked the City Hotel.

Top 20 agents with highest bookings



This graph shows the top 20 agents with their booking contribution and be able to use this data to gain more understanding and to make business decision.

The most booked accommodation type(single,couple,family)



As we seen graph more than 60% of booking Couple(or adults) is the most popular accommodation type. So we can make plans accordingly

In which month Hotels busiest

- Resort has more bookings at July and August month, and least booking in June, September, and start and end of the year.
- Similarly Hotel has more booking in July and August, and the least bookings were made at the start and end of the year.

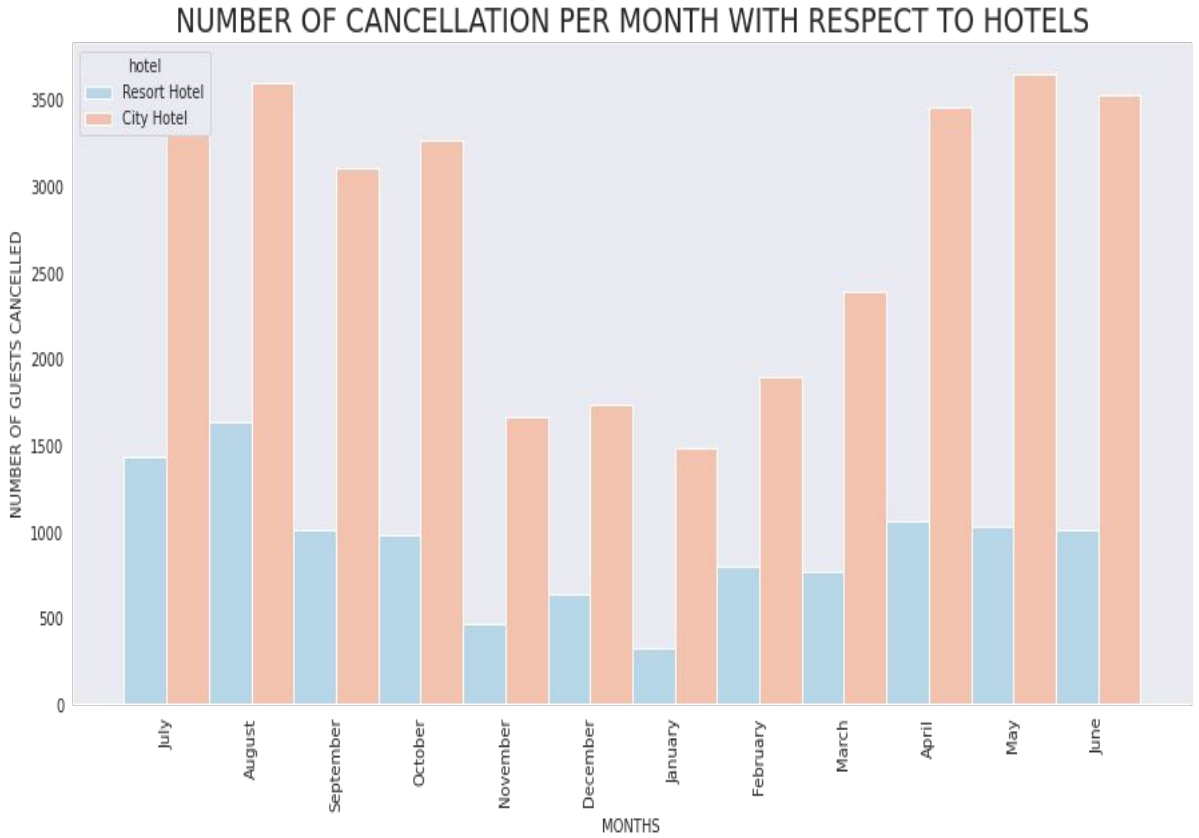
Booking Cancellation Per Month

CITY HOTEL :

- For city hotel most of the cancellation was done between April to August.
- And least cancellations are between November to January.

RESORT HOTEL :

- For resort hotel most of the cancellation was done in July and August.
- Least cancellation was between November to March.



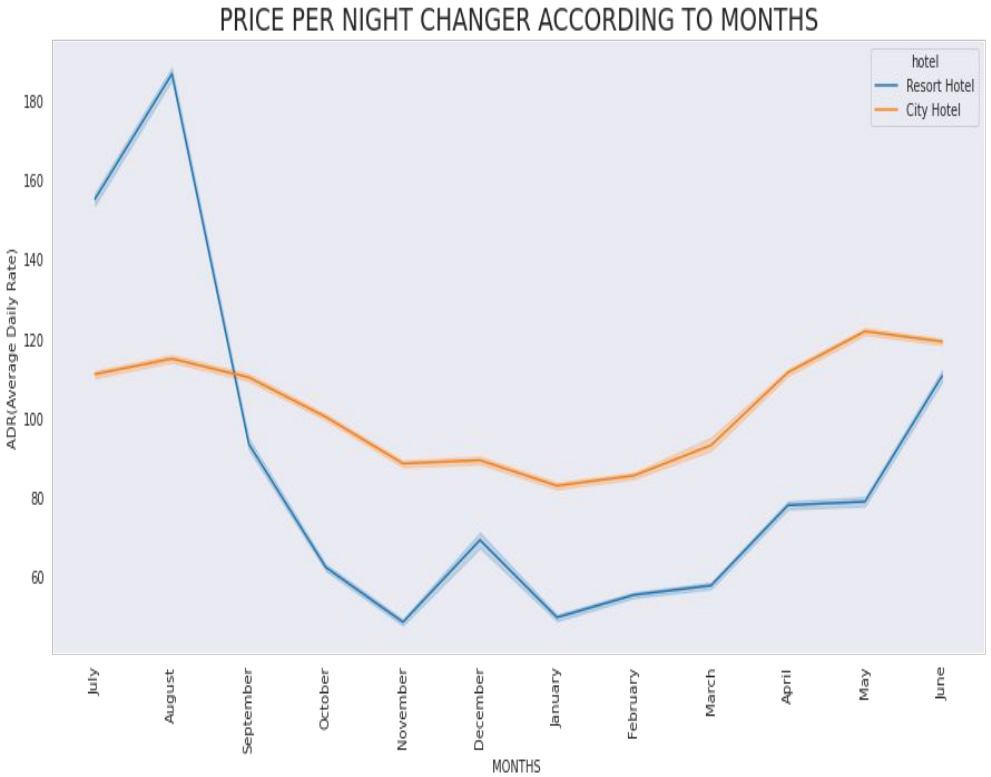
ADR (Average Daily Rate) Over Months Of The Year

RESORT HOTEL :

- The adr price for Resort hotel is quite high in the month of the August.
- It show very steep fall after the month of August and it is lowest in the November and January.

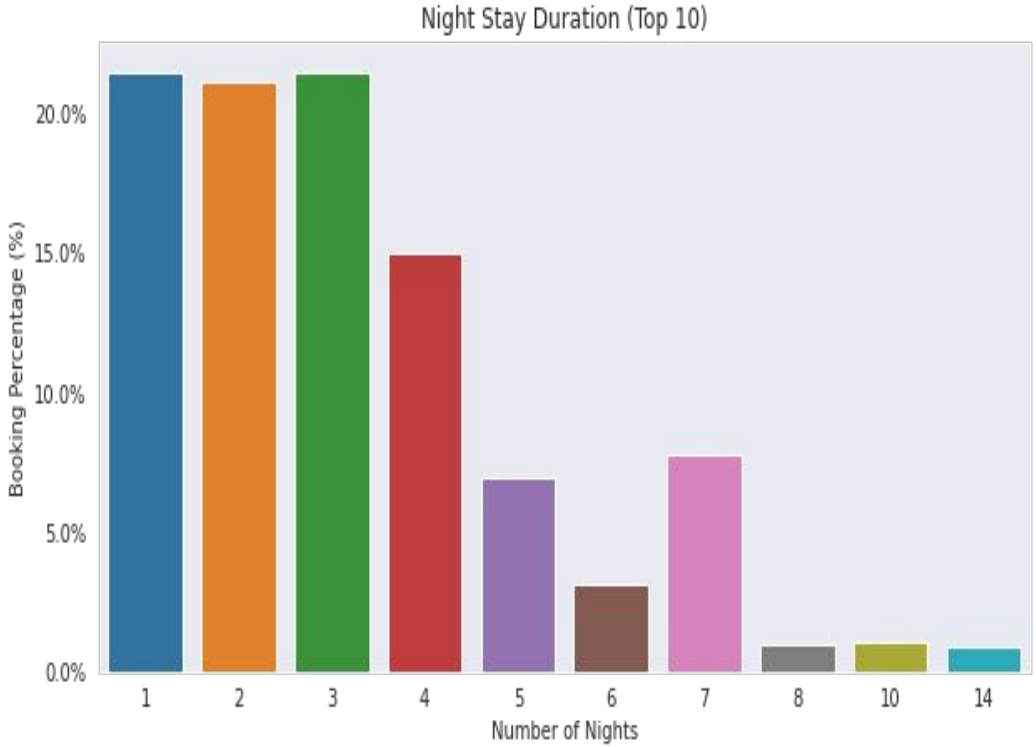
CITY HOTEL :

- The City hotel shows smooth curve with respect to Adr , Its range lies between 80$ to 130$.
- For the City hotel price per night is at peak in the month of May and least in the month of January,however the price range does not changes much.



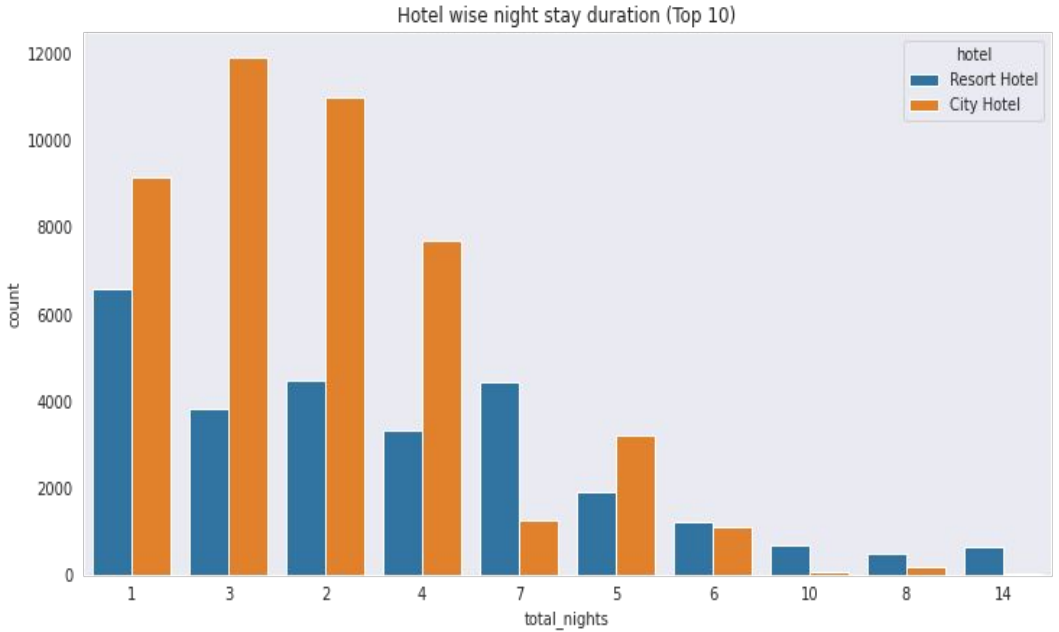
Accommodation Tenure

- Nearly 60% of the guest that visits stays in the hotel for the tenure of 3 days or less than 3 days.
- There are such visitors also who stays in hotel for more the 3 days and less than week ,but they are less in percentage.
- The guests that stays in hotels for more than a week are in very few with respect to percentage.

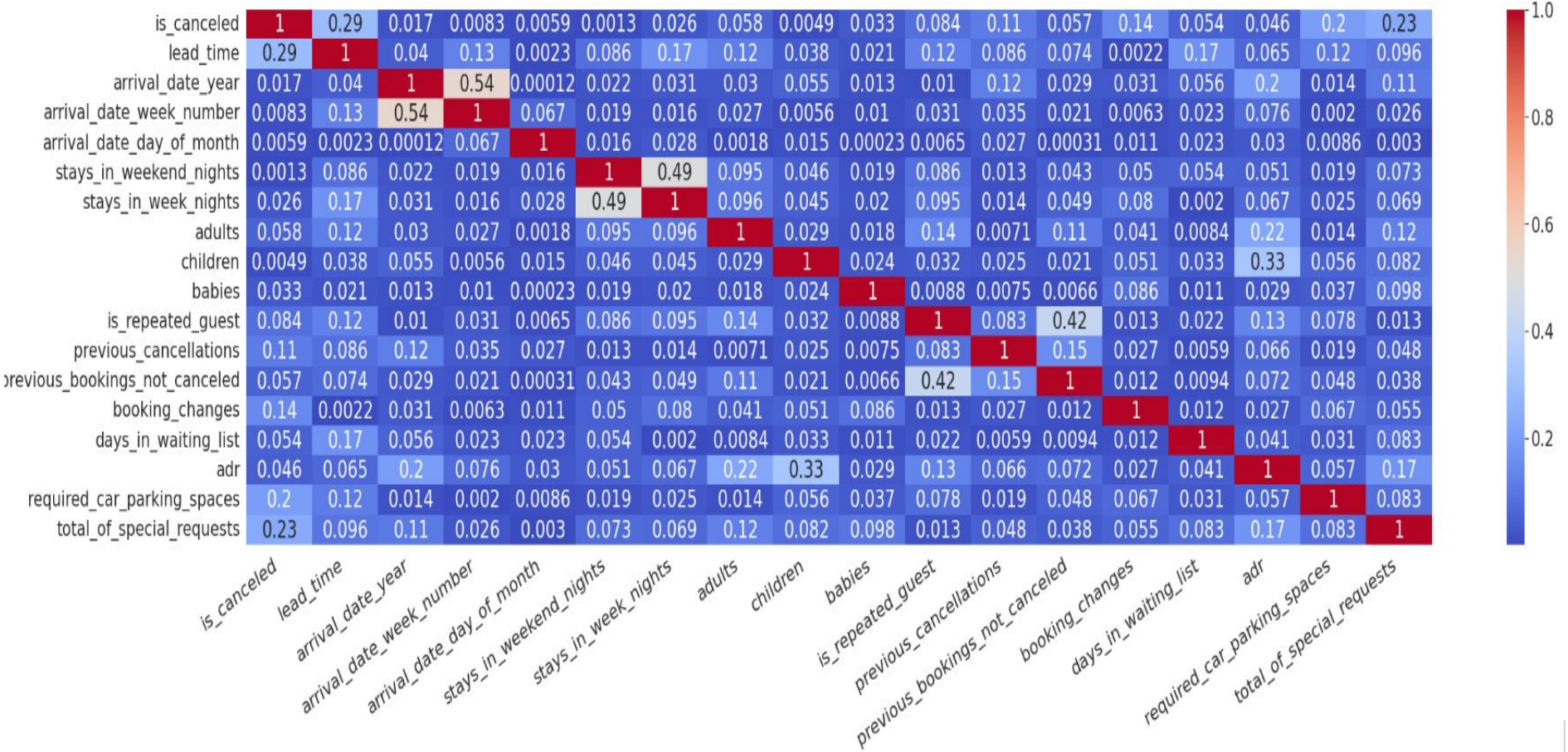


Hotel Wise Accommodation Tenure

- The highest number of guests stays in Resort hotel is for 1 day are more than 6000 guests.
- Looking at the bars of the City hotel we get information that most of the guests stays for the duration of 3 days range is 10000 to less than 12000.
- Above 5 days are least.



Correlation Matrix



Correlation Matrix

- In the heatmap, It shows some high correlations between few variables, that's because we have created some new columns from existing columns and have not dropped it later.
- It tells us how one variable depends on other variables that mean if we change one variable value how it’s affect our dataset.
- arrive_date_week_number and arrive_date_year are high correlated.
- Above 50% we say data is correlated.

Conclusion

- 'City hotels' and 'Resort hotels' are two types of hotels present in the dataset, out of which 'City hotels' are more preferred by the customers than the latter(66.4% customer prefers 'City hotels' where as 33.6% customer prefers 'Resort hotels').
- Out of 119000 customer dataset, 75166 customers checked in hotel while 44224 customers cancelled their bookings, that is about 37% of the booking got cancelled whereas 63% customers checked in the hotel.
- Majority of the deposit type is of 'No deposit' type, which itself concludes the high rate of cancellation rate.
- From the given dataset, we can see 2016 is the year in which hotel bookings are the highest.
- We can also see the trend in the middle of the year as those season has less weather condition and holidays are more during that season. We can also infer that winter season has the lowest number of bookings around the globe and we can assume it because of the weather condition
- Out of all months 'August' witnessed highest number of hotel bookings whereas 'January' witnessed the least.
- City hotels are high demand compared to resort hotels in all aspects due to its reliability of majority of the population.

References

- https://pandas.pydata.org/
- https://seaborn.pydata.org/
- https://www.geeksforgeeks.org/
- https://stackoverflow.com/
- https://matplotlib.org/

AI

Thank You

PRAJWAL D U