HOTEL BOOKING ANALYSIS

Vaitul Sidhdhapara & Drashti Shah

Data Science Trainees,
AlmaBetter, Bangalore

Abstract

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. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series or decision trees.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

Problem Statements:

1. We will be analyzing some key metrics for hotel bookings like:
   - The number of cancellations
   - Number of bookings on weekday vs weekends
   - Most preferred meal types
   - Country wise bookings
   - New customers acquired
   - Customer lifetime value of the existing customers
   - Type of rooms preferred by customers
   - Booking types,
   - Hotels available for booking
   - The revenue of the hotels

2. We will be using various lenses to look through the data to analyze patterns associated with each segment such as:
   - The type of hotel
   - Day of week
   - Type of customers
   - Type of rooms

3. Finally, we will also try to predict the future bookings either based on time series analysis or decision tree.

Data Descriptions:

hotel	Type of Hotel whether 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 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 number of year for arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number of adults
children	Number of children
babies	Number of 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	Country of origin.
market_segment	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
distribution_channel	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
is_repeated_guest	Value indicating if the booking name was from 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. 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.
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	Type of deposit made for booking: 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
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 booking: 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 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: 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

Introduction

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, among other things

We are provided with a hotel bookings dataset. Our main objective is performed 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. We perform the codes by using theses libraries. 1) NumPy, 2) Pandas, 3) Matplotlib, 4) Seaborn.

Type of libraries:

1) NumPy
2) Pandas
3) Matplotlib
4) Seaborn

Reason for using libraries:

NumPy: Stands for numerical python. In data we come across lot of numerical calculation. To make those easier instead of coding it from the scratch we can use NumPy library functions. There are also more function in numpy. It also has functions for working in domain of linear algebra, Fourier transform, and matrices

Pandas: Pandas is mainly used for data analysis and associated manipulation of tabular data in DataFrames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel.

Matplotlib: Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB.

Seaborn: It is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataFrames and the Pandas library. The graphs created can also be customized easily.

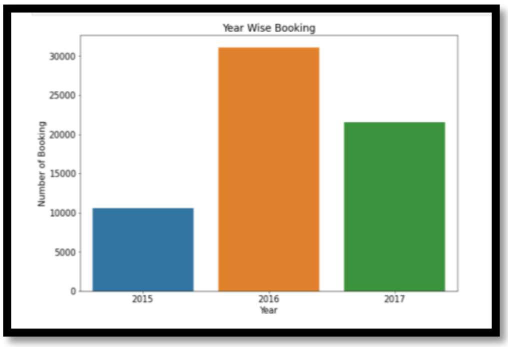
Challenges

(1) There was a lot of duplicate data
(2) Data was present in wrong data type format
(3) Choosing appropriate visualization techniques to use was difficult.
(4) A lot of null values were there in the dataset.

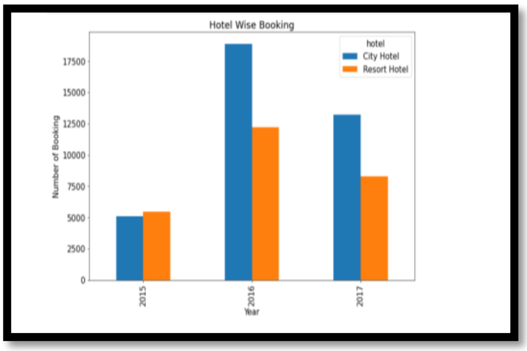
First of all, we clean it and then we start our EDA.

- **We have tried to answer these following questions**

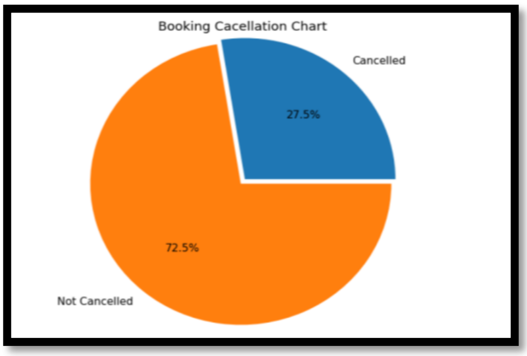
1. **Which year have a more number of bookings?**



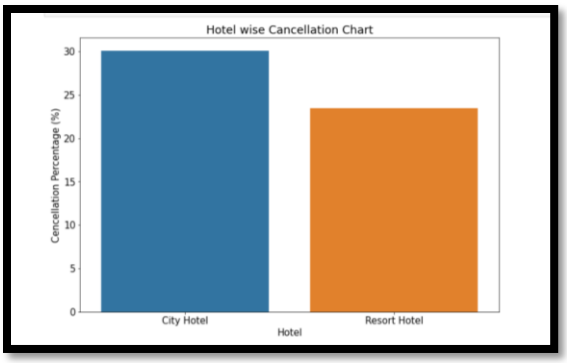
2. **How many bookings are done in every year according to the hotel?**



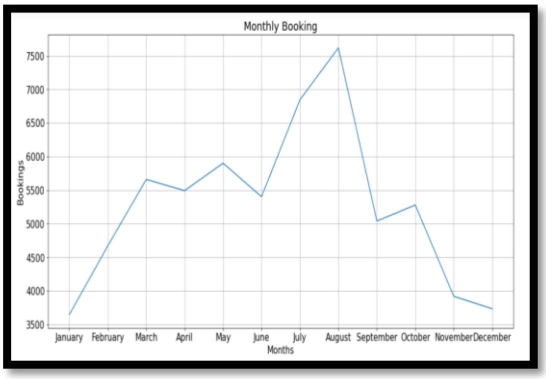
3. **What is the booking cancellation ratio?**



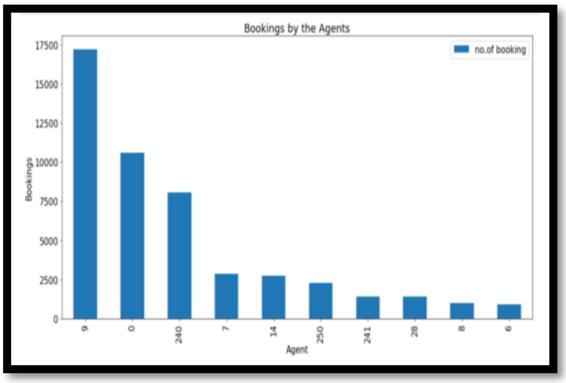
4. **Which hotel has higher bookings cancellation rate?**



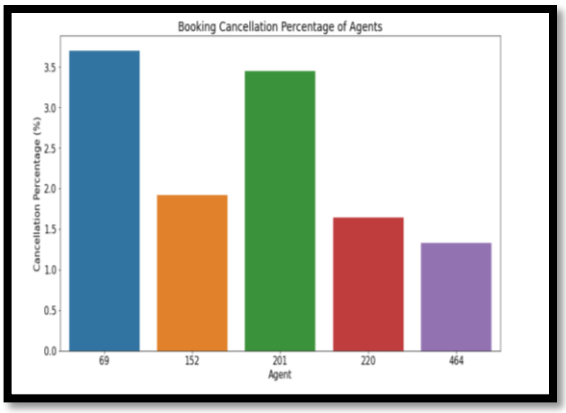
5. **Which month has maximum bookings?**



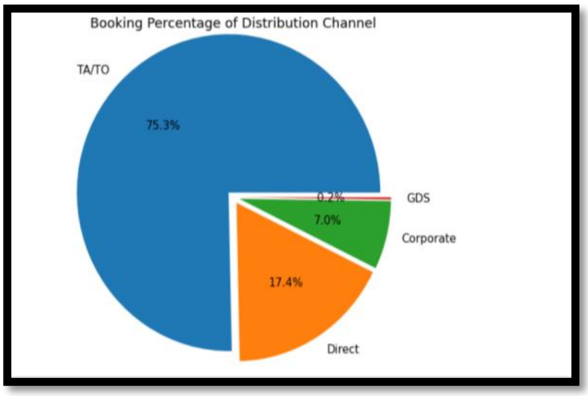
6. **Which agent does the most bookings?**



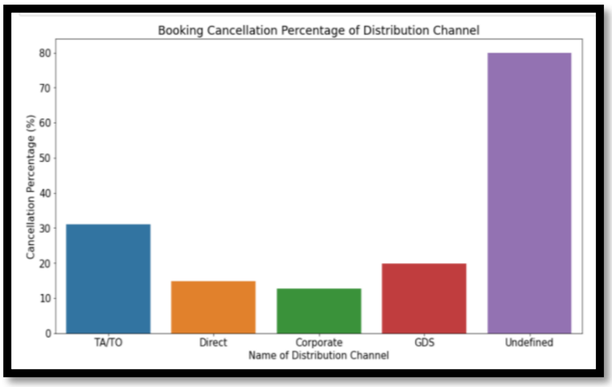
7. **Which agent has the lowest booking cancellation ratio?**



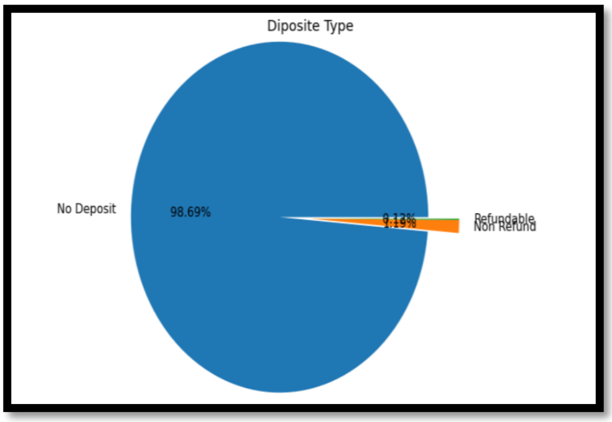
8. **Which distribution channel has highest booking rate?**



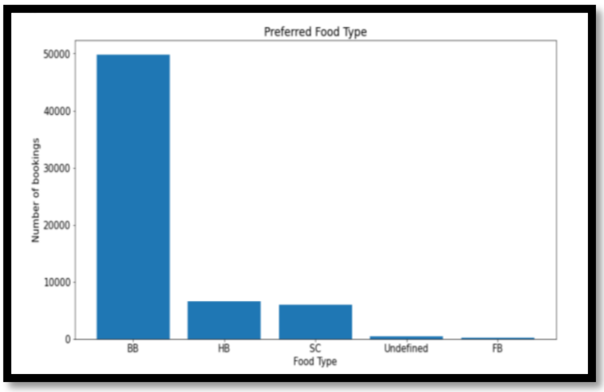
9. **Which distribution channel has highest cancellation percentage?**



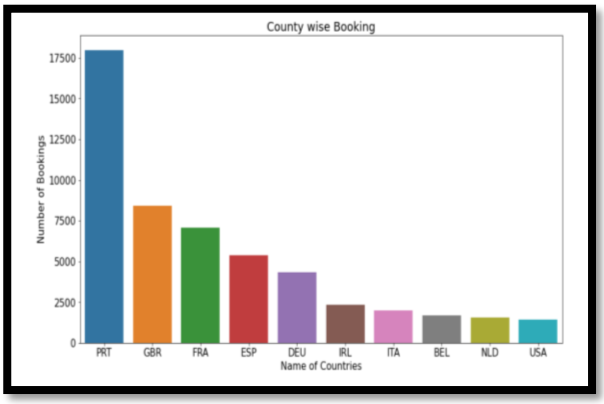
10. **Which type of deposit is preferred by customers?**



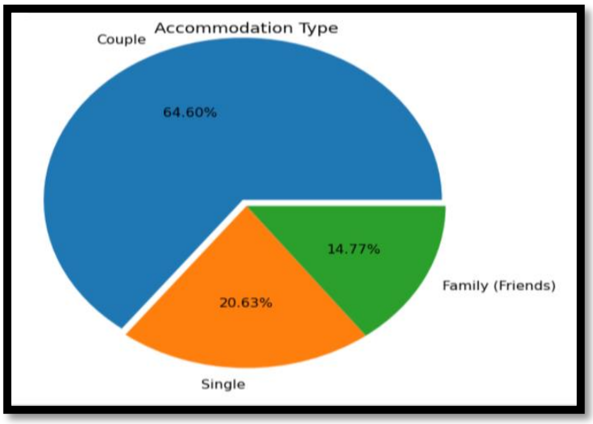
11. **Which type of food(meal) is preferred by customers?**



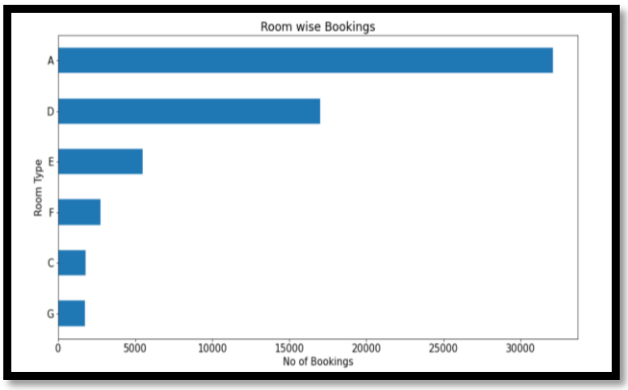
12. **Which country has the highest number of bookings by customers?**



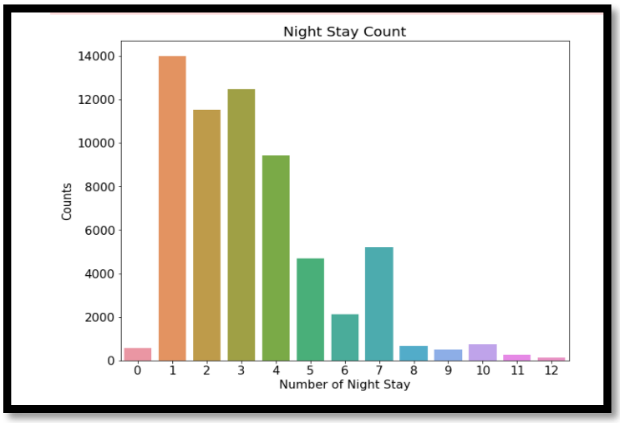
13. **Which is the most booked accommodation type (Single, Couple, Family)**

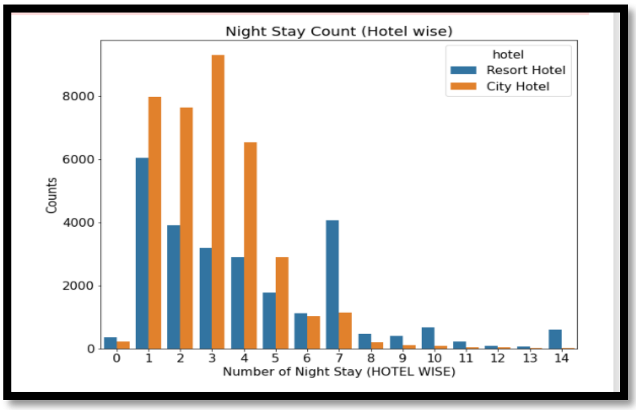


14. **Which type of room is most in demand?**

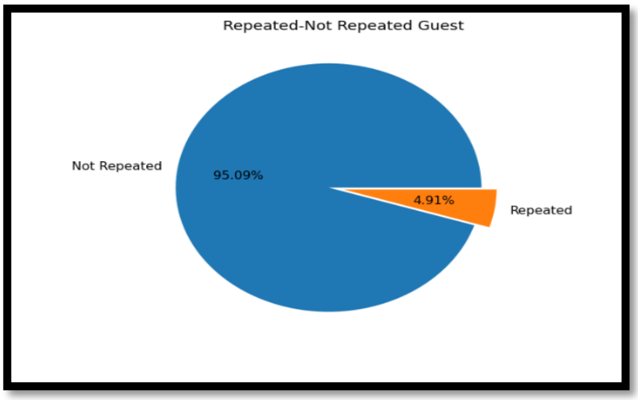


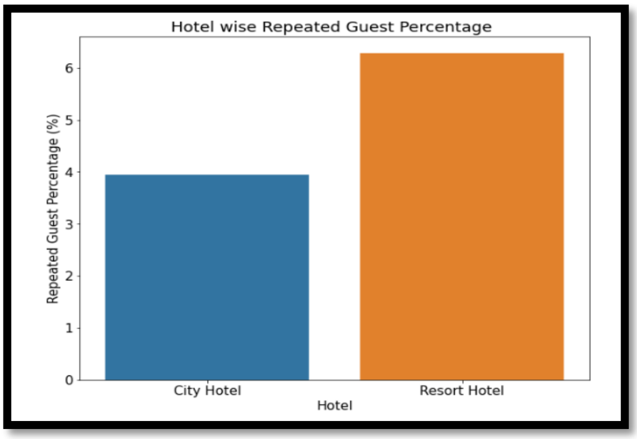
15. **How many nights guests choose to stay the hotel?**



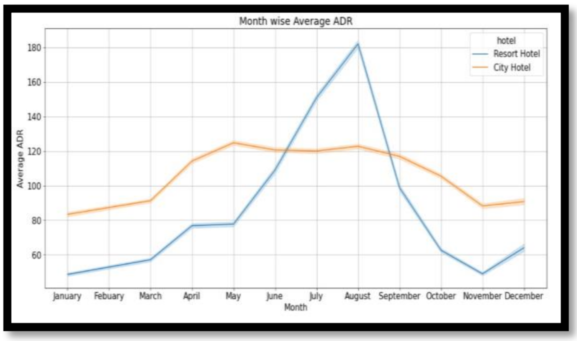


16. **What is the ratio of repeated guest?**

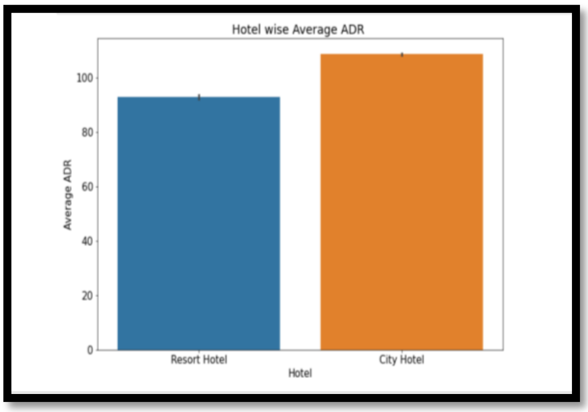




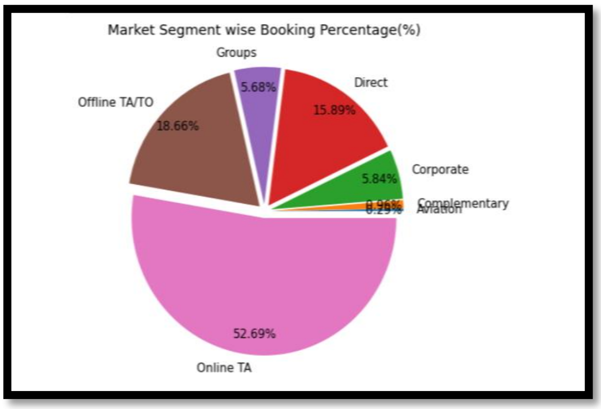
17. **Which month has the highest ADR (Hotel wise)?**



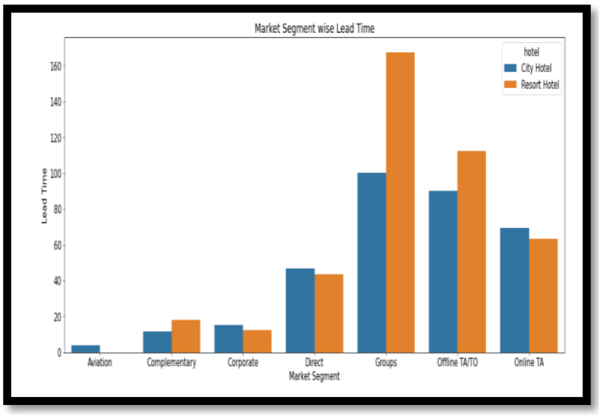
18. **What is the average ADR of hotel?**



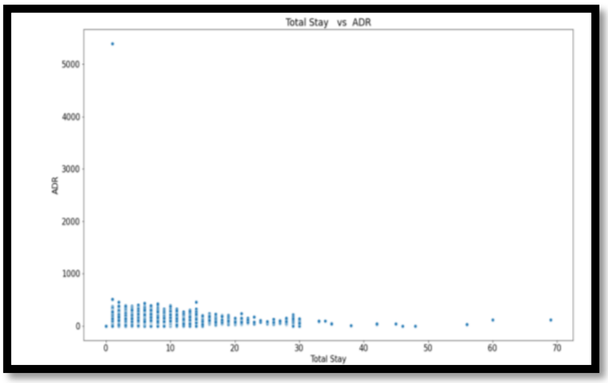
19. **By which market segment are the most bookings done?**



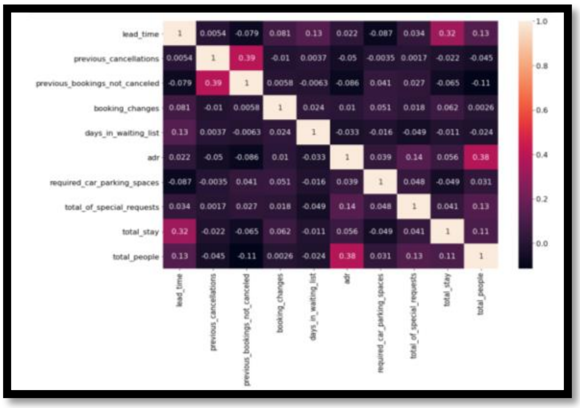
20. **Which market segment has the longest lead time?**



21. **Relationship between number of Total Stay and ADR.**



22. **Correlation between two parameters.**



Conclusion

City Hotel is busier than Resort Hotel. Also, the overall adr of City hotel is slightly higher than Resort hotel. Mostly guests stay for less than 4 days in hotel and for longer stays Resort hotel is preferred. Both hotels have significantly higher booking cancellation rates and very few guests less than 4 % return for another booking in City hotel. 6% guests return for stay in Resort hotel. Most of the guests came from European countries, with most of guests coming from Portugal. Guests use different channels for making bookings out of which most preferred way is TA/TO. Almost 30% of bookings via TA/TO are cancelled. July-August are the most busier and profitable months for both of hotels. Couples are the most common guests for hotels; hence hotels can plan services according to couples needs to increase revenue. For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr. Overall booking cancellation ratio is 27.5%. Agent 9 is done more number of bookings and Agent 464 has lowest cancellation ratio. For longer hotel stays people generally plan little before the actual arrival.