

# CAPSTONE PROJECT-

## EDA ON HOTEL BOOKING ANALYSIS

*By*  
*Diksha Shejao*

Content:

Problem statement

Data Collection and Understanding

Data Cleaning and Manipulation

Exploratory Data Analysis(EDA)

Conclusion:

## PROBLEM STATEMENT:

- In this project I am going to analyse Hotel Booking Dataset. This dataset contains information of city hotel and resort hotel and includes information of booking time, length of stay, number of adults, children and / or babies, also have information of available parking space, among other things.
- The objective of this project is to explore and analyse the data to discover important factors that govern the booking.

## WORKFIOW:

Divide workflow into 3 steps.

Data Collection and Understanding  
Data Cleaning and Manipulation  
Exploratory Data Analysis(EDA)

## Data Collection and Understanding:

Data collection and understanding are very important .  
So I have Hotel Booking data. This data contains  
119390 rows and 32 columns. Let's understand the  
columns.

## Data Description:

- Hotel : Type of hotel(City or Resort)
- is\_cancelled: If the booking was cancelled (1) or not (0)
- Lead\_time: Number of days before the actual arrival of the guests
- 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\_month: Day of arrival date
- stays\_in\_weekend\_nights: Number of weekend nights(Saturday or Sunday) spent at the hotel by the guests.
- stays\_in\_week\_nights: Number of weeknights (Monday to Friday) spent at the hotel by the guests.
- adults: Number of adults among guests
- children: Number of children among guests
- babies: Number of among guests
- meal: Type of meal booked
- country: Country of guests
- market\_segment: Designation of the market segment

- `distribution_channel`: Name of booking distribution channel
- `is_repeated_guest` : If the bookings 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
- `reserved_room_type`: Code of room type reset
- `assigned_room_type`: Code of room type assigned
- `booking_changes`: Number of change / amendments made to the booking
- `deposit_type`: Type of the deposit made by the guest
- `agent`: Id of the travel agent who made the booking
- `company`: Id of the company that made the booking
- `Days_in_waiting_list`: Number of days the booking was on the waiting list
- `Customer_type`: Type of customer, assuming one of four categories
- `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
- `reservation_status`: Reservations status (Cancelled, Check-Out or No-Show)

- reservation\_status\_date: Date at which the last reservation status was updated

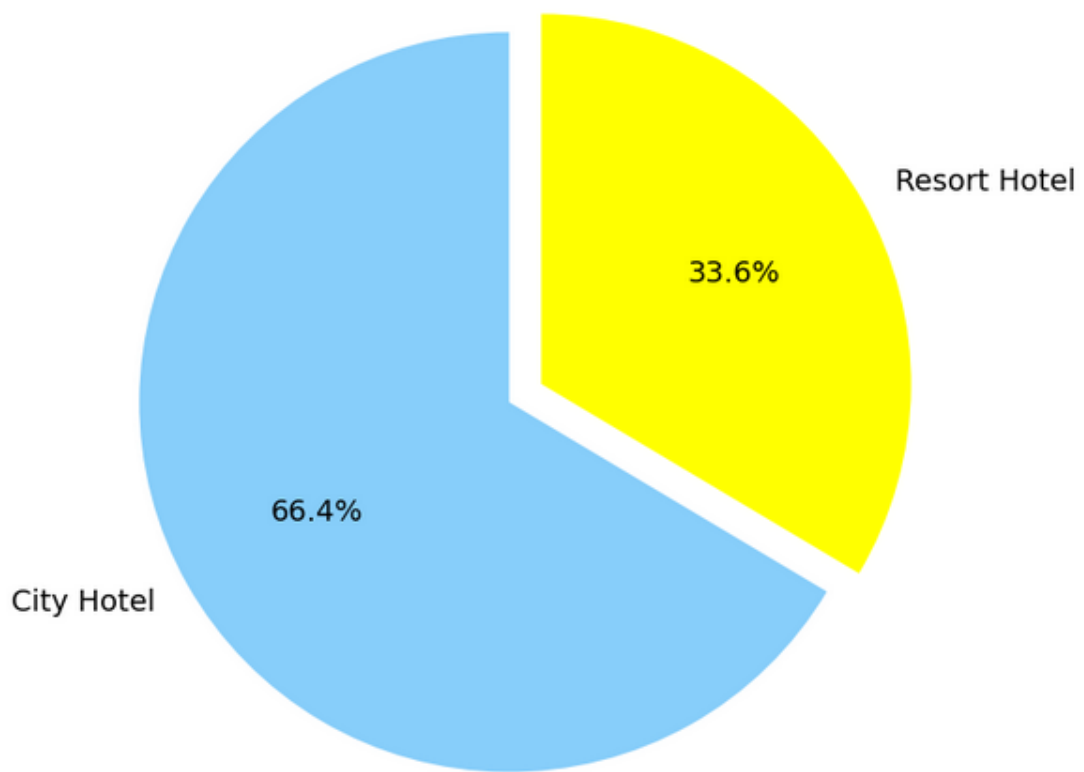
## Data Cleaning and Manipulation:

- We are replacing all the null values with 'Mean values' for 'Numerical Variable' and with 'Absent' for 'categorical Variables'.
- Data has 4 columns that have missing values and those columns are company(112593), agent(16340) , country(488) and children(4). So the missing values of columns are company,agent and children are replaced by 0

and the missing values of column country are replaced by others by using `.fillna()`.

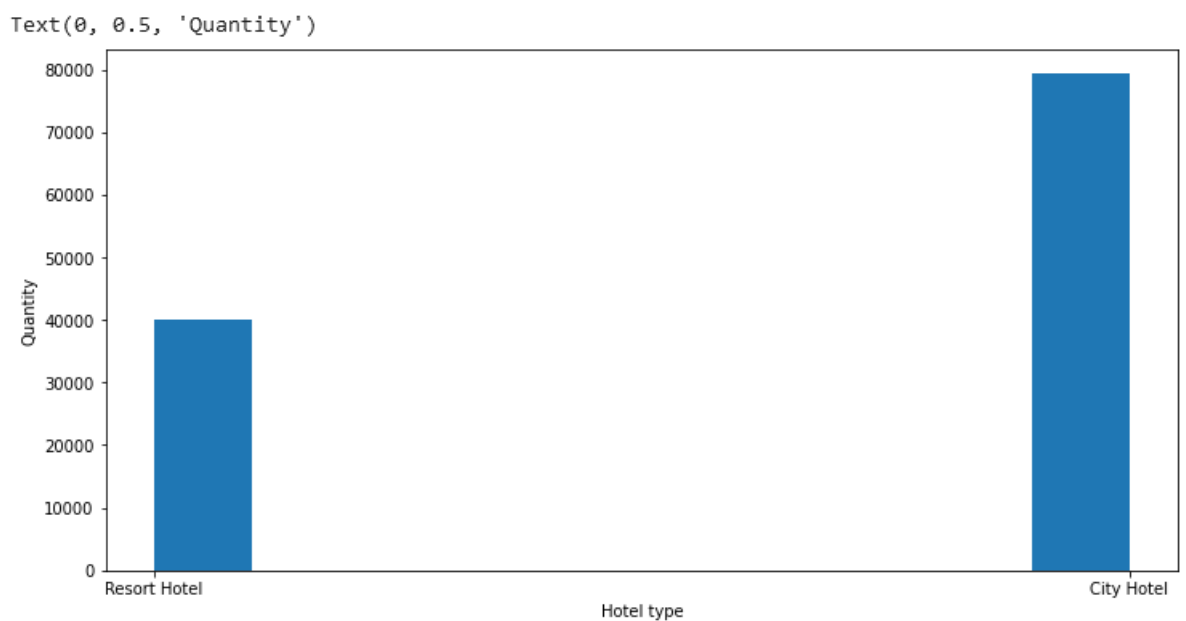
Exploratory Data Analysis(EDA):  
Pie chart of most preferred hotel

City hotel:





Hotel type:

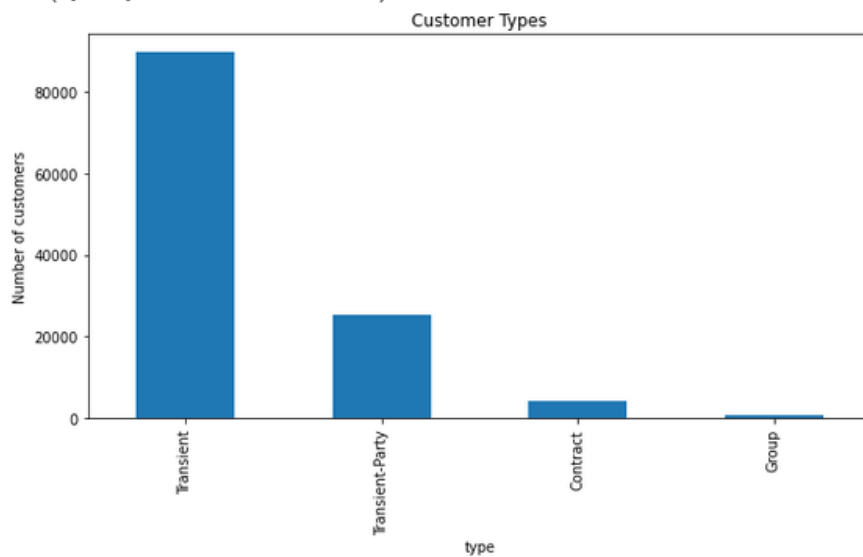


Conclusion:

- (1) City hotel is the most preferred hotel and the percentage is 66.4% means city hotel is the busier hotel type
- (2) Hotel type are resort hotel and city hotel

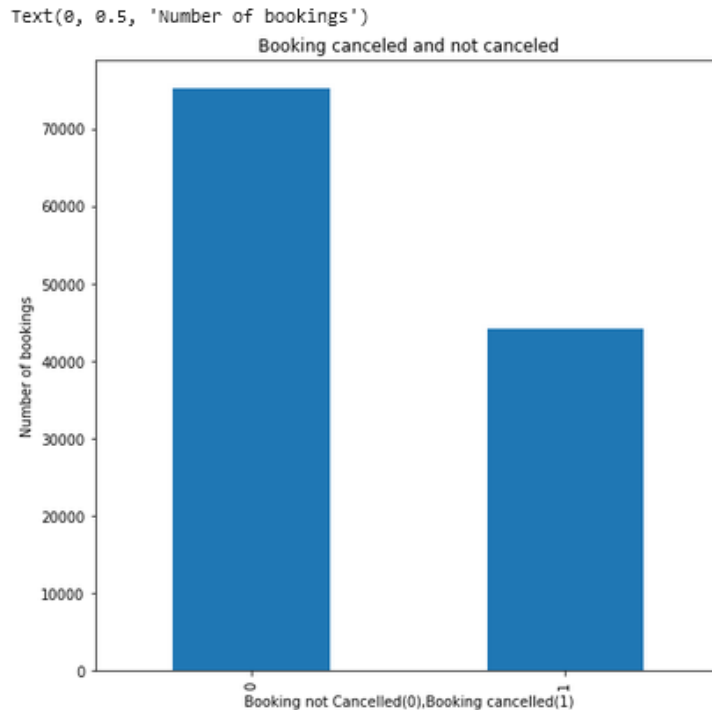
## Custom Type:

Text(0, 0.5, 'Number of customers')



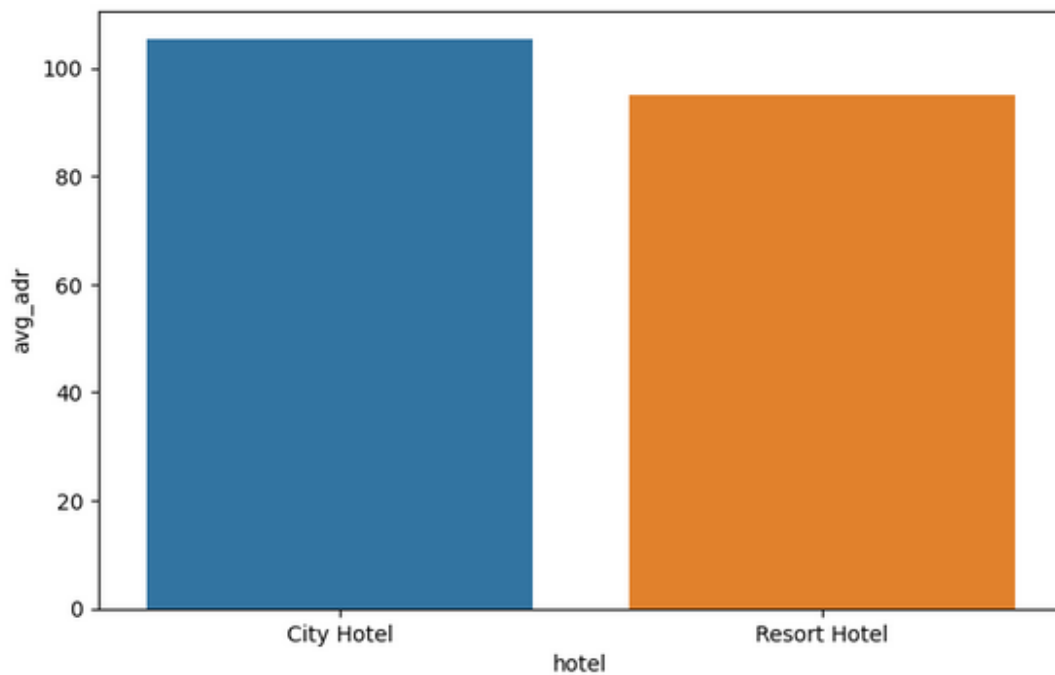
Custom types are transient,transient\_party,contract,group.  
Most of the custom type are transient

## Booking cancelled and not cancelled:



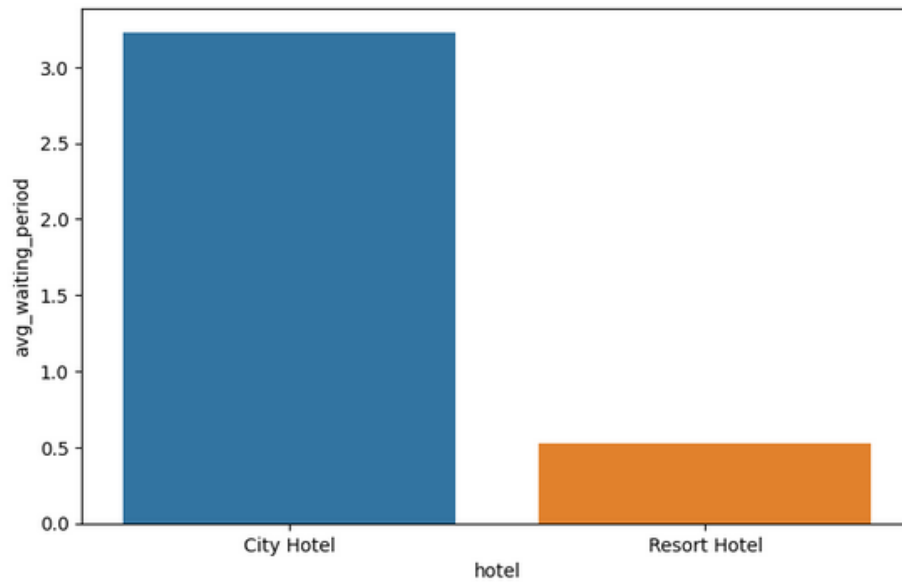
zero(0) shows the most of the booking not cancelled and one (shows) booking cancelled  
zero(0) shows 62% is booking not cancelled and  
one (1) shows 37% booking cancelled.

Hotel seems to make more revenue:



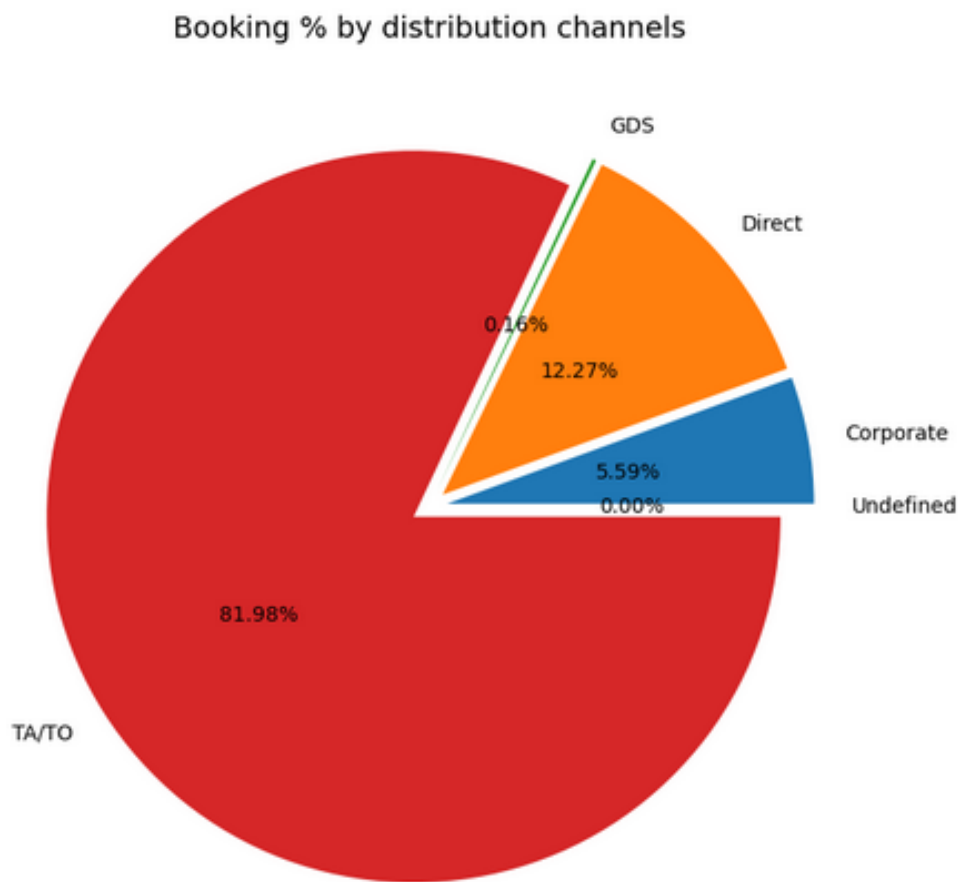
The Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.

Hotel has longer waiting time:

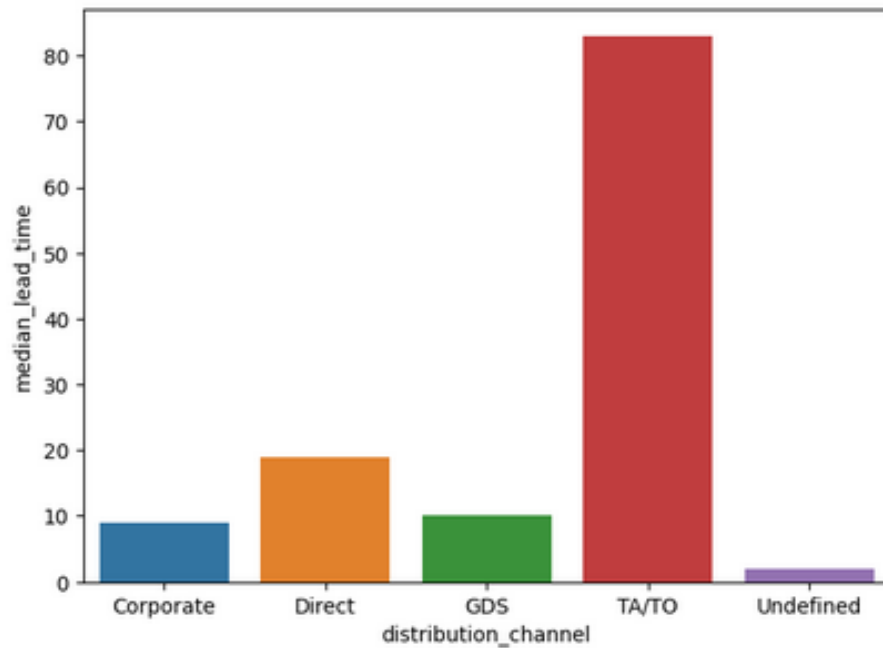


City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.

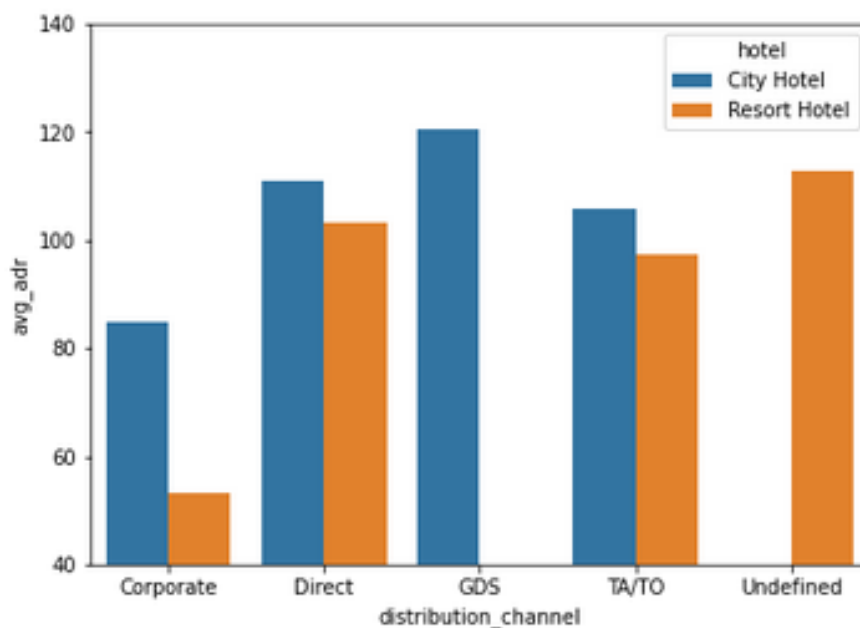
## Most common channel for booking hotels:



- Mostly used for early booking of hotels are
- TA/TO is mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred.



- Distribution channel brings better revenue generating deals for hotels

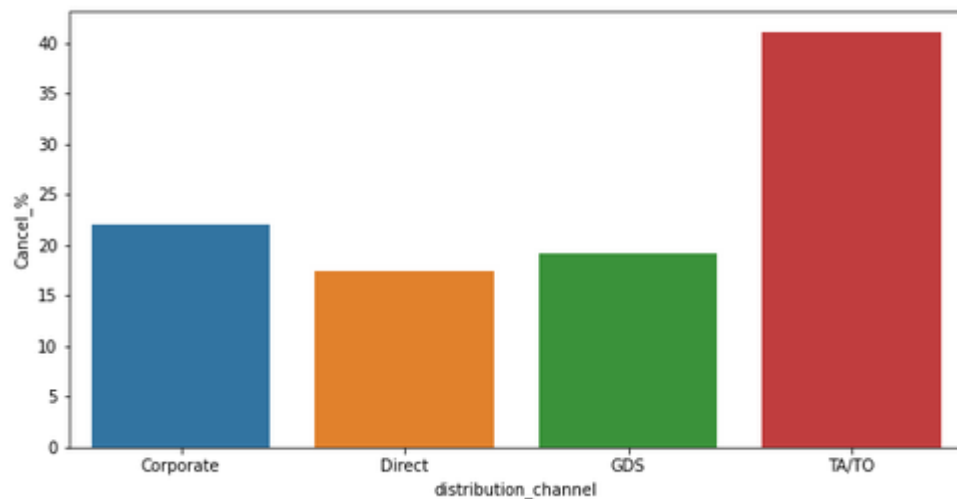


- GDS channel brings higher revenue generating deals for City hotel, in contrast to that most

bookings come via TA/TO. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals.

- Resort hotel has more revenue generating deals by direct and TA/TO channel. Resort Hotel need to increase outreach on GDS channel to increase revenue.

Significant distribution channel has highest cancellation percentage



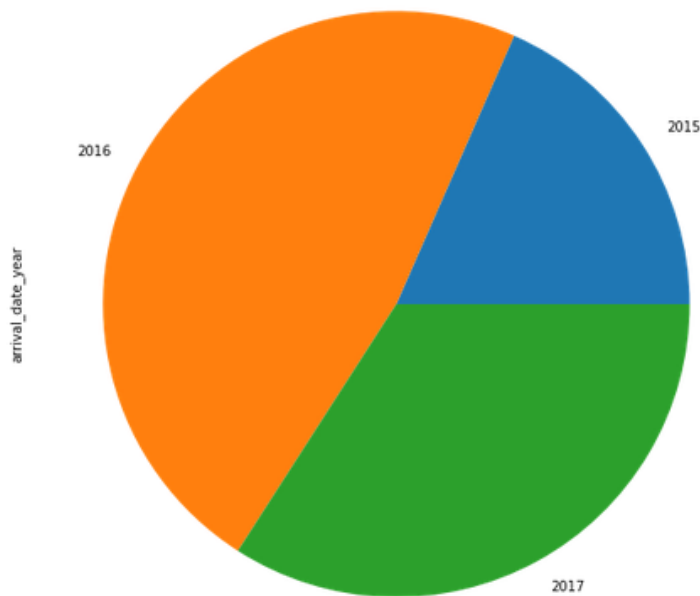
- TA/TO has highest booking cancellation %.  
Therefore, a booking via TA/TO is 30% likely to get cancelled.



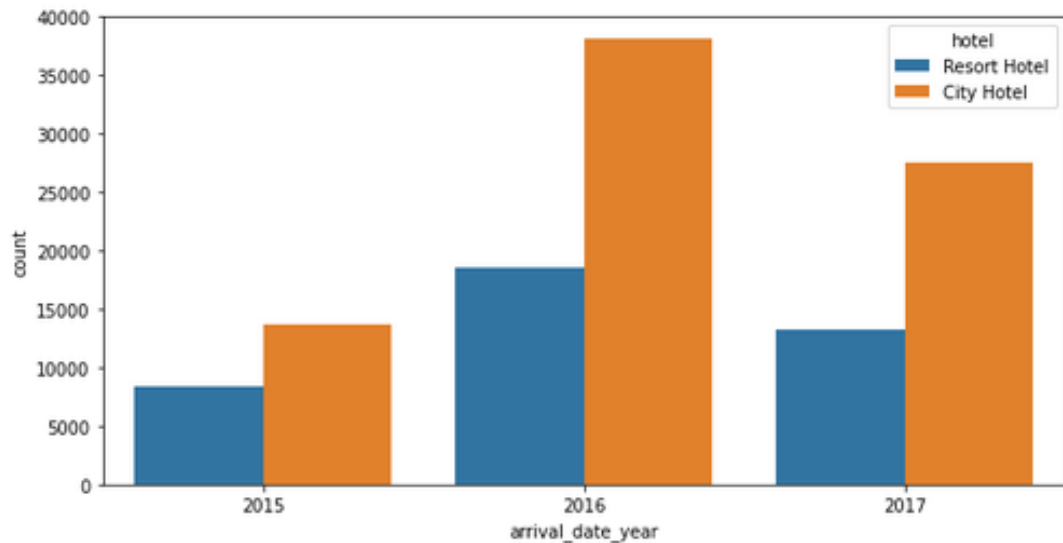
## Total Number of Bookings across different years:

Text(0.5, 1.0, 'Number of bookings on different years')

Number of bookings on different years

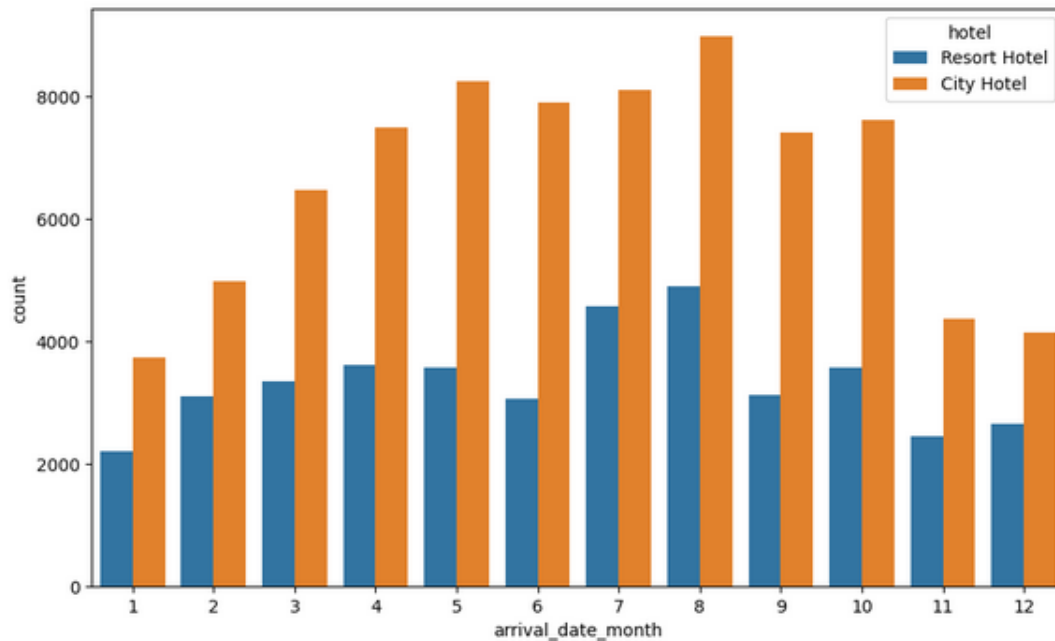


- We can tell from the above graph that more than 50 percent of the bookings are done on the year 2016, following 2017 and 2015.
- Booking across different year for Resort Hotel and City Hotel are.



- We can validate from the above graph for the previous pie chart that 2016 was the year most bookings were done.
- And from the above plot, we can infer that city hotel has the demand when compared to resort hotels. The number of bookings for city hotel has gone as high as more than 35000 for the year 2016, more than 25000 for the year 2017 and more than 10000 for the year 2015.
- Whereas, the resort hotel bookings were around 17000 for the year 2016, 10000 for 2017 and 8000 for the year 2015

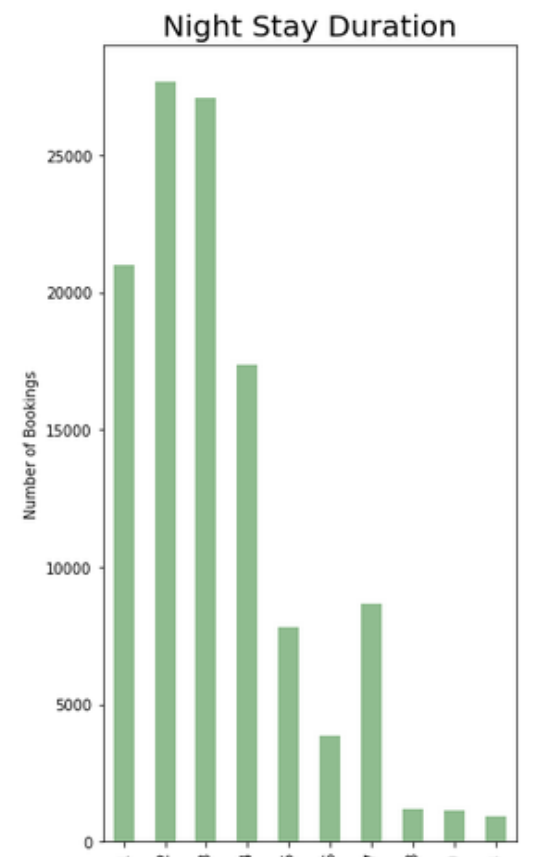
Number of bookings done across different months:



- The above graph brings the insights about bookings made across different months, So we see that number of bookings are high on the months of July and august following, may, june, september and october.
- The most probable reason for this would be that these months are most ideal seasons for vacation as much weather 1000 fluctuation will not be present.

## Night Stays:

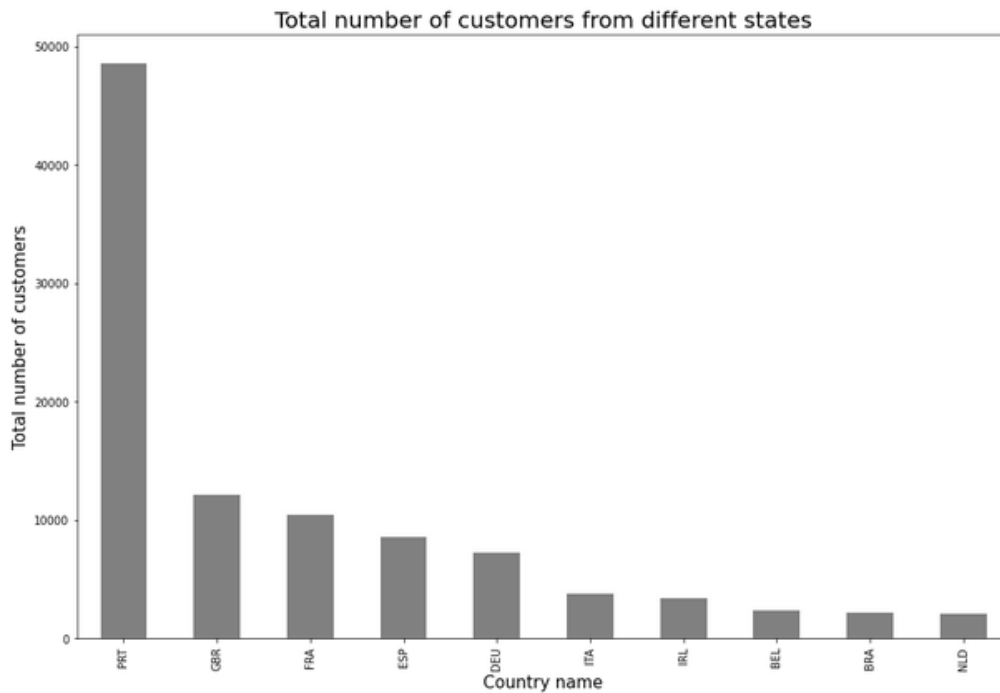
- Total Number of night stays for all bookings



## Number of nights

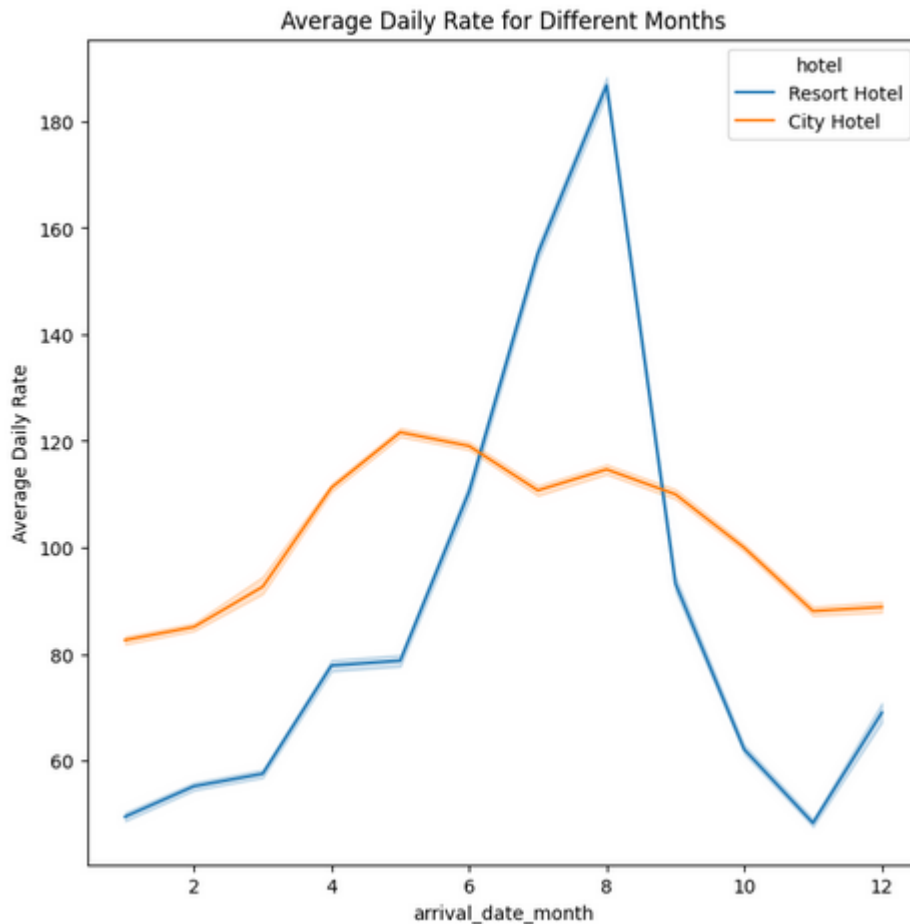
- Most number of people stays for 2-3 days in the hotel and even for one day.
- The demand for long stay in hotels are very less

## Top Countries from different states:



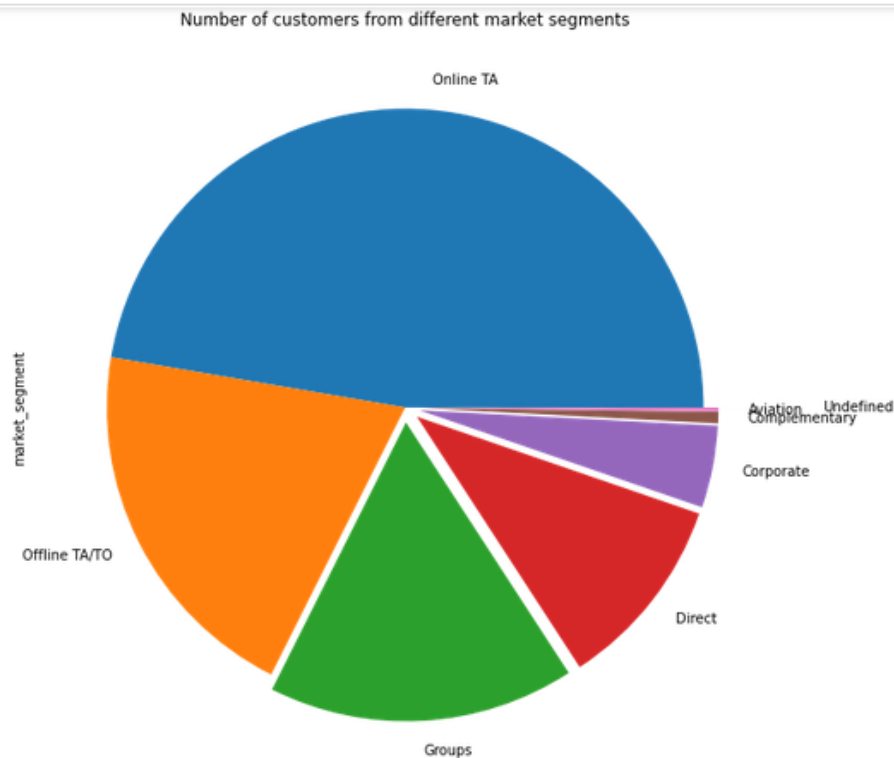
- Portugal leads to a outstanding number of bookings of which not even double of the second position country can beat.
- United Kingdom, France, Spain and Germany are at the position of second, third, fourth and fifth respectively.

## Average Daily Rates for Different months:



- The average daily rates are high during the months of may, june, july and august for City hotels
- The average daily rates are high during the months of july, august and september for Resort Hotels

Number of customers from different market segments:



- 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.

## Conclusion:

1. Almost 60% Bookings are for City hotel and 40% Bookings are for Resort hotel, therefore City hotel is busier than resort hotel
2. 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.
3. Mostly guests stay for less than 5 days in Hotel and for longer stays Resort hotel is preferred
4. Almost 30% of bookings via TA/TO are cancelled
5. Majority of the deposit type is of **'No deposit'** type, which itself concludes the high rate of cancellation rate.
6. From the given dataset, we can see 2016 is the year in which hotel bookings are the highest.
7. 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.
8. City hotels are of in high demand compared to resort in all aspects due to its reliability of majority of the population.
9. Most of the customers prefer to stay for one, two and three nights and majoritily its two and three night which tops.
10. Among the top 10 countries with respect to number of bookings, most of them are European countries which



belongs to the list of which Portugal stands first with 48.5k bookings.

11. For resort hotels, ADR is more expensive during June, July, August and September whereas for city hotels, ADR is more expensive during May, June July and August.

So, Overall ADR of both cities are more expensive between May and September.

12. Online Travel agencies provides high amount of leads and customers following Offline TA, Groups, Direct, etc.,

13. For Customers, generally the longer stays (more than 15 days) can result in better deals in term of low adr.

So, we can conclude that we need to focus our marketing area on online TA since majority of the visitors tend to reach out to them.

