Capstone Project

EDA on Hotel Booking Analysis

Nehal S Jambhulkar

Problem Statement

- ✓ For this project we will be analyzing Hotel Booking data. 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.
- ✓ Hotel industry is a very volatile industry and the bookings depends on above factors and many more.
- ✓ The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management ,which can perform various campaigns to boost the business and performance.

- So we will divide our work flow into following 3 steps

Data Collection and Understanding

Data Cleaning and Manipulation

Exploratory Data Analysis(EDA)

EDA will be divided into following 3 analysis.

1. **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data you are analysing is only one variable.
2. **Bivariate analysis:** Bivariate analysis is where you are comparing two variables to study their relationships.
3. **Multivariate anlysis:** Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

Data Collection and Understanding:

- ❖ After collecting data it’s very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let’s understand this 32 columns.

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.

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.

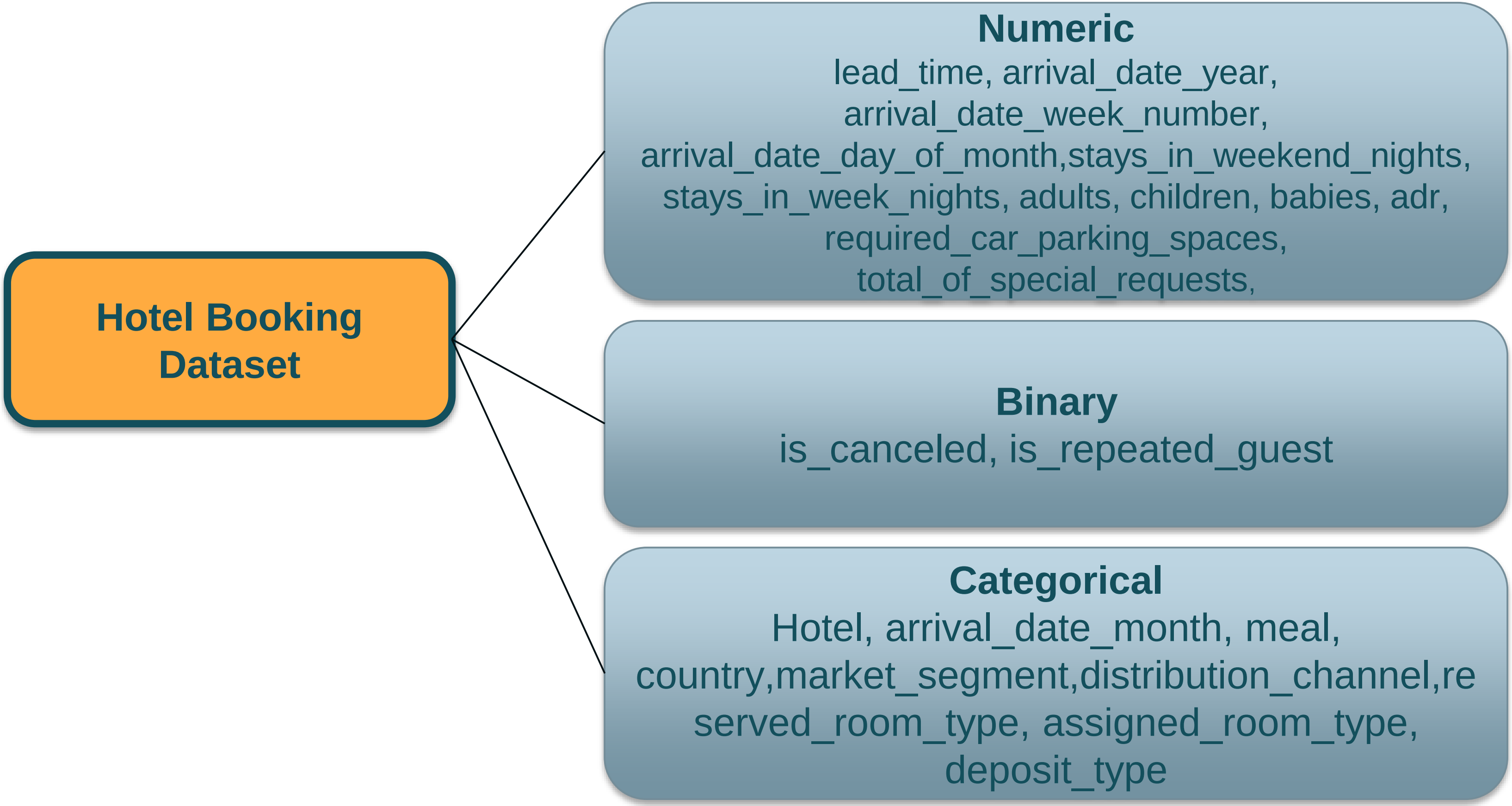
adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

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.

Dataset Input data summary



Hotel Booking Dataset

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 Cleaning and Manipulation:

1. Company, agent, country and children columns with missing values.I replaced missing values as per requirement.
2. Dropping company column because more then 90% data is missing
3. Data had 31994 duplicates values. So I dropped it from the data.
4. I created 2 new columns

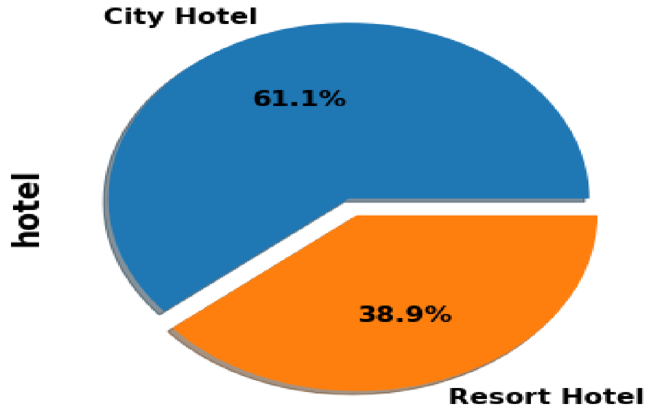
   A)‘total_People’ = from the Children, adults, babies.

   B) ‘total_stay ’ = From weekend nights and weekdays night.

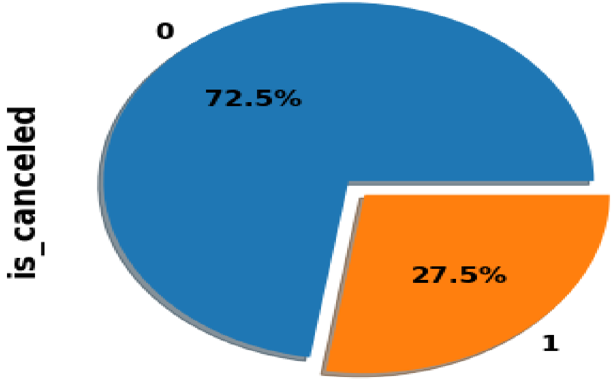
Exploratory Data Analysis (EDA) :

Univariate Analysis

Pie Chart for Most Preffered Hotel



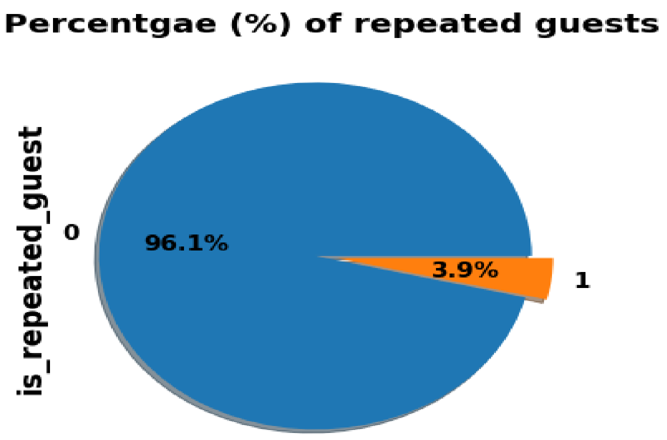
Cancellation and non Cancellation

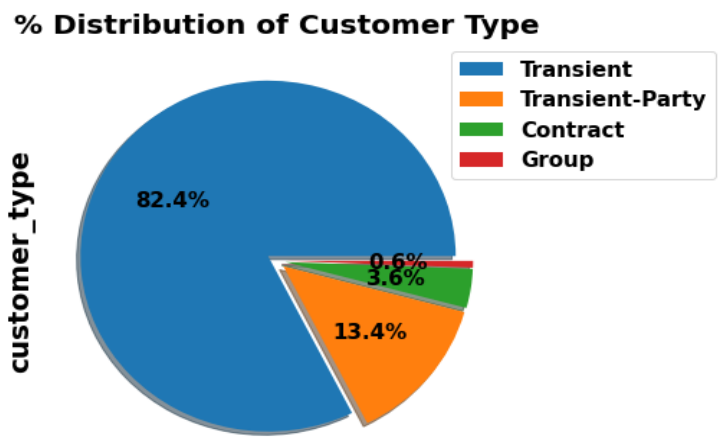


Conclusions:

1) City hotels is the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.

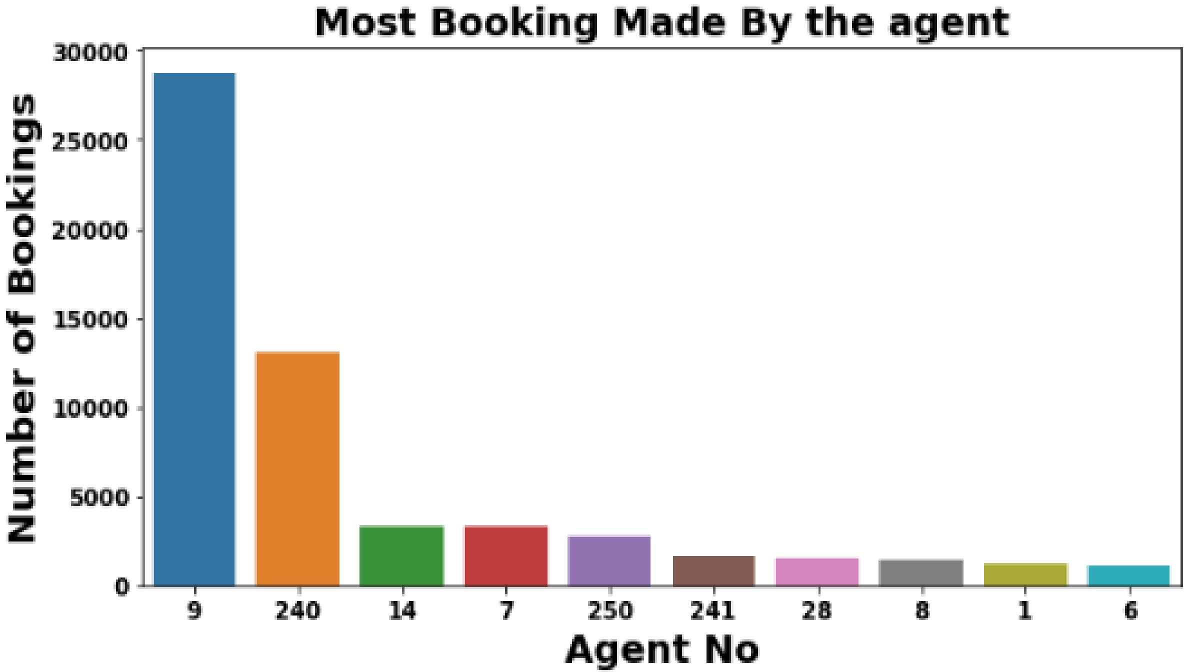
2) 27.49 % bookings were got cancelled out of all the bookings





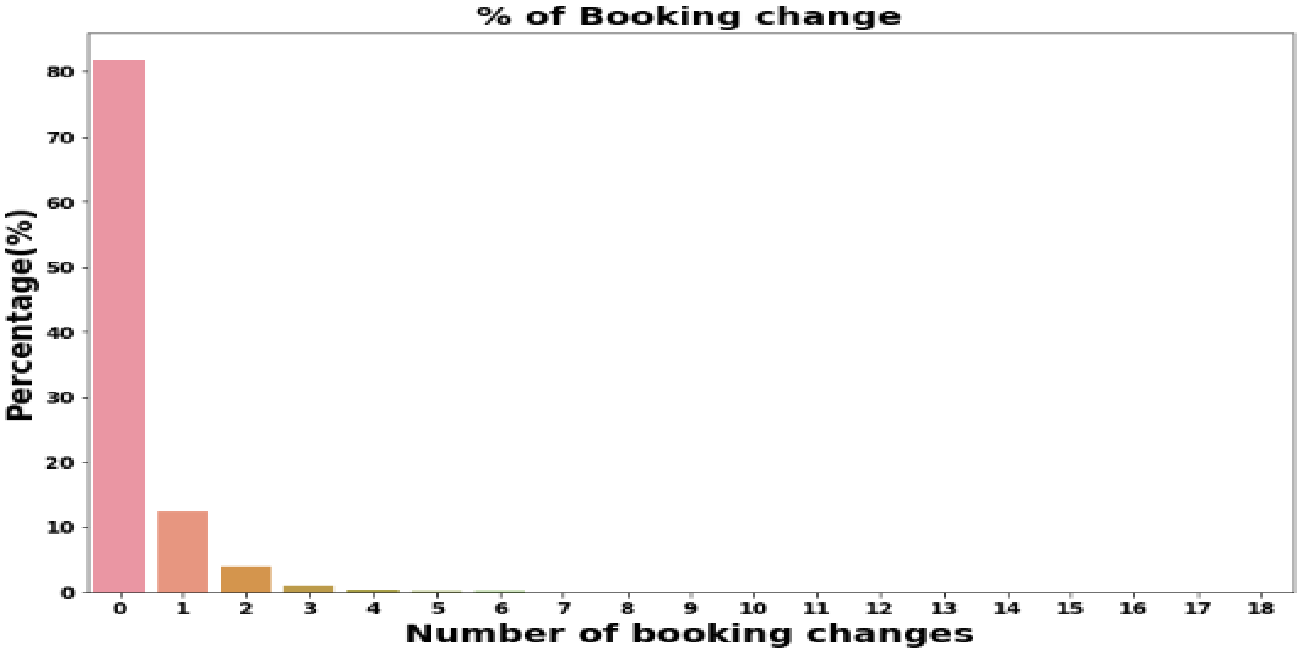
Conclusion:

- Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- Most of the customers/guests were Transient type(82.4%). And transient party were 13.4% and 0.6 belongs to group. Remaining guests belongs to Contract type.



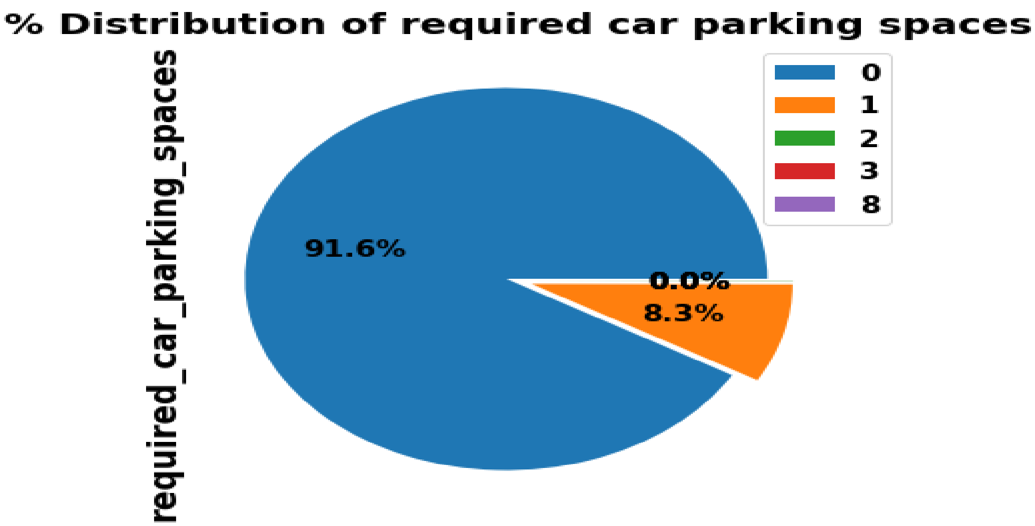
Conclusions:

➢ Agent Id no -9 made the highest bookings which is more than 28721.



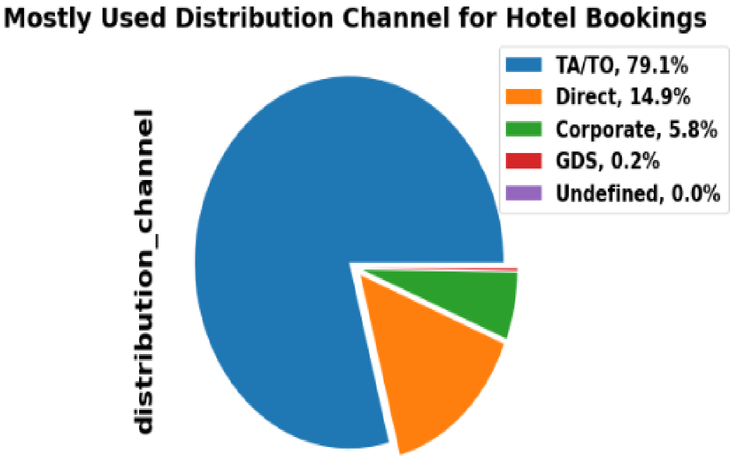
Conclusions:

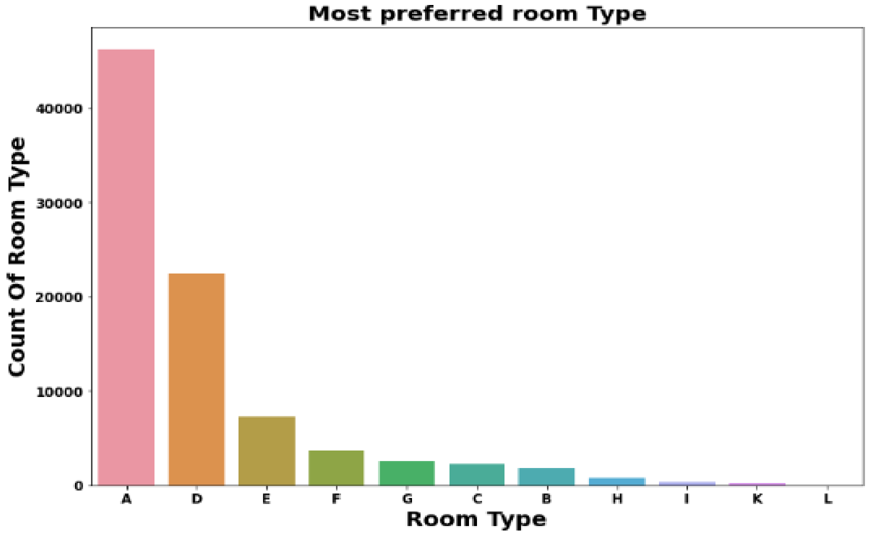
- The percentage of 0 changes made in the booking was more than 82 %.
- Percentage of Single changes made was about 10%.



Conclusions:

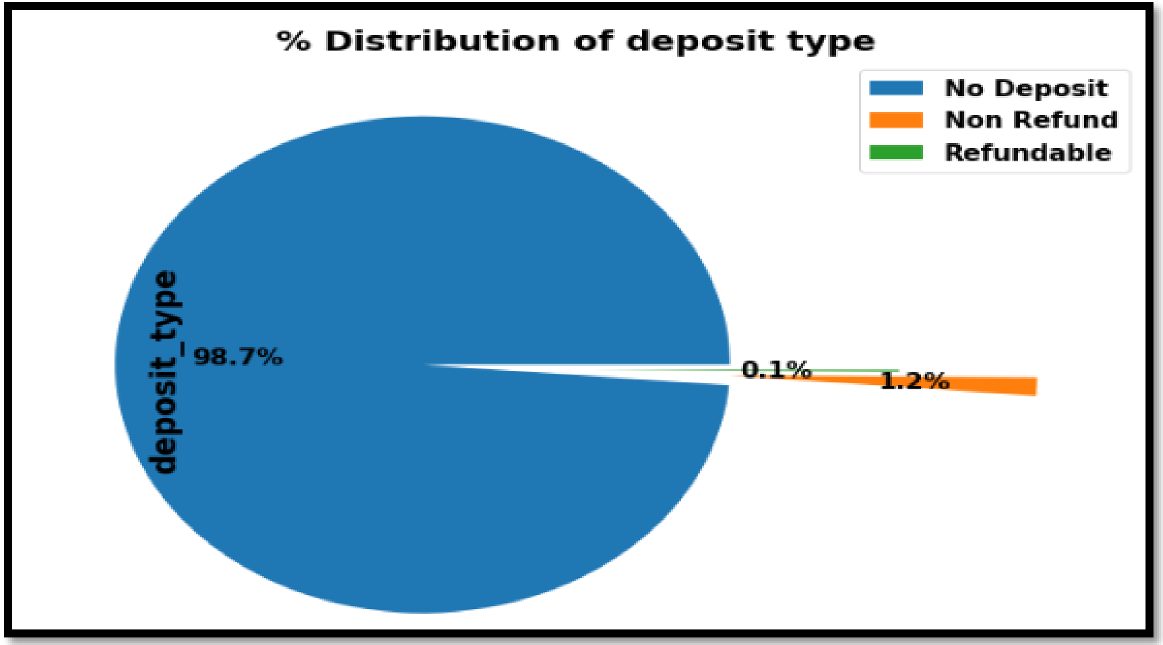
- Most of the customers(91.6%) do not require car parking spaces. Only 8.3 % people required only 1 car parking space.





Conclusions:

- 79.1 % bookings were made through TA/TO (travel agents/Tour operators). Second most channel is direct.
- Room type 'A' is most preferred by the guests second most preferred is 'D'.



Conclusions:

- ➢ Almost 98.7% of the guests prefer ‘No deposit’ type of criterion while booking hotels.

PRT- Portugal

GBR- United Kingdom

FRA- France

ESP- Spain

DEU - Germany

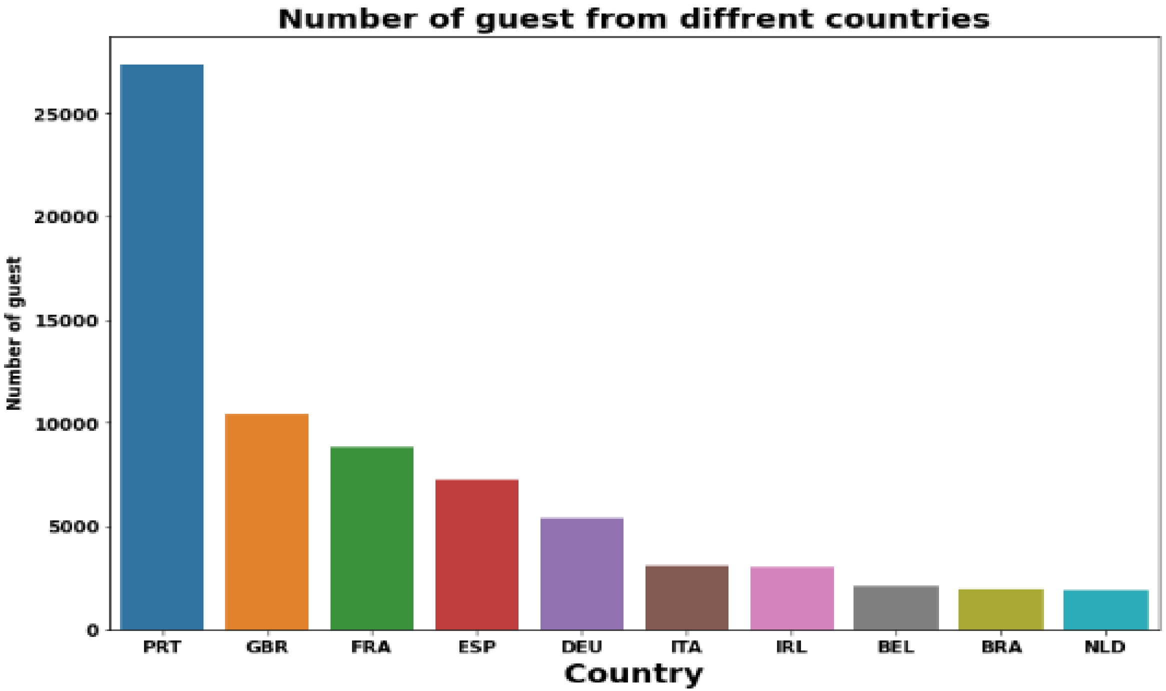
ITA –Italy

IRL - Ireland

BEL -Belgium

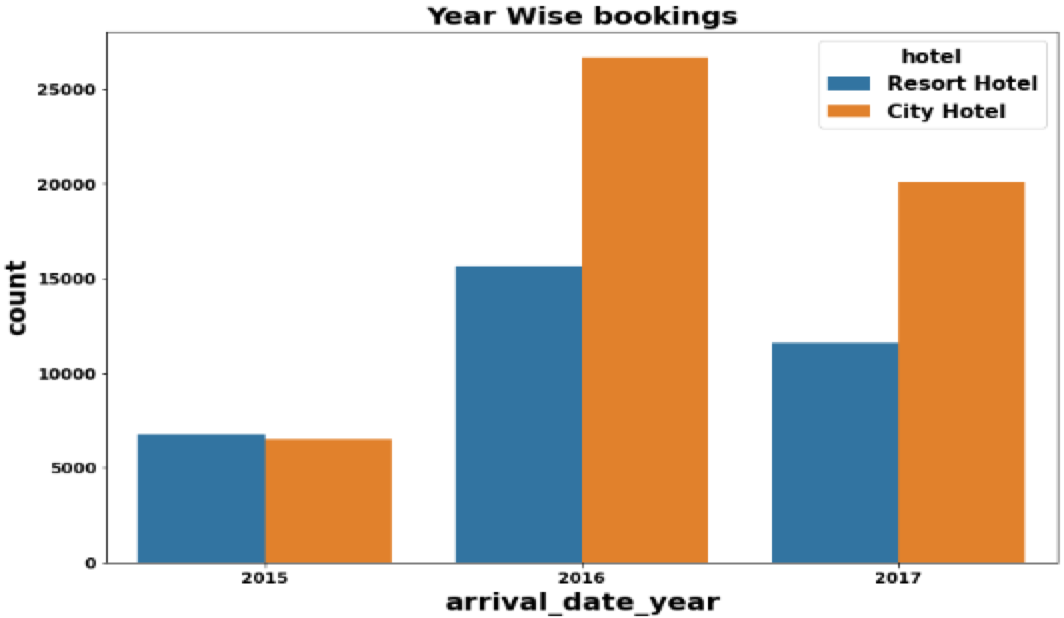
BRA -Brazil

NLD-Netherlands



Conclusions:

- Most of the guests are coming from Portugal i.e. more 25000 guests are from Portugal

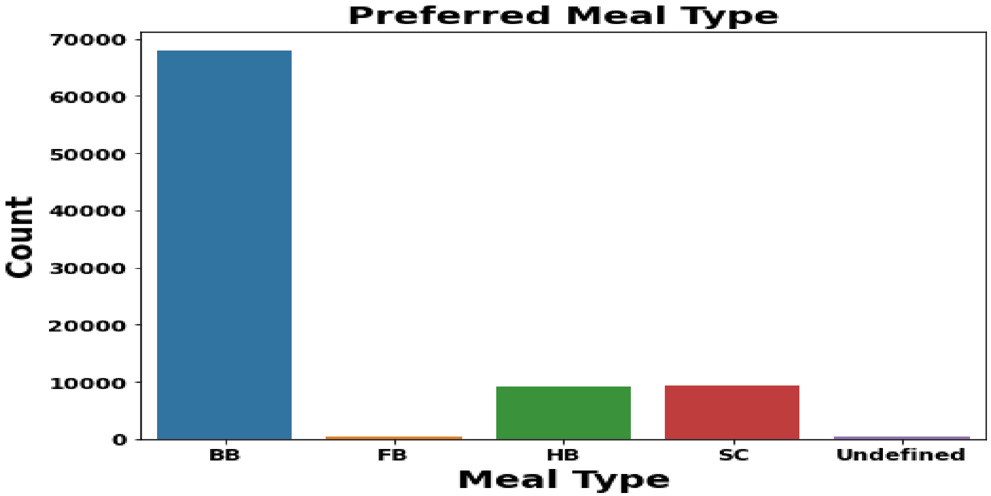


Conclusion:

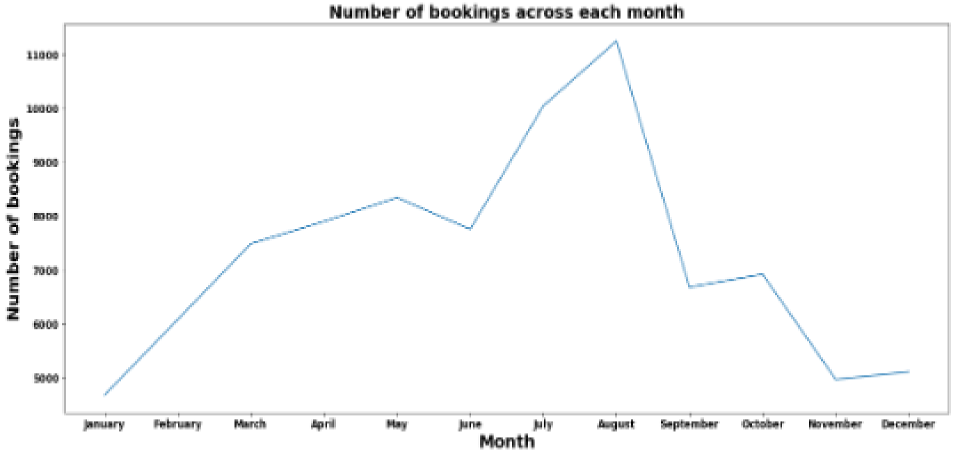
- Most of the bookings for City hotels and Resort hotel were happened in 2016. As we can see Most of the bookings were for City hotels

Conclusions:

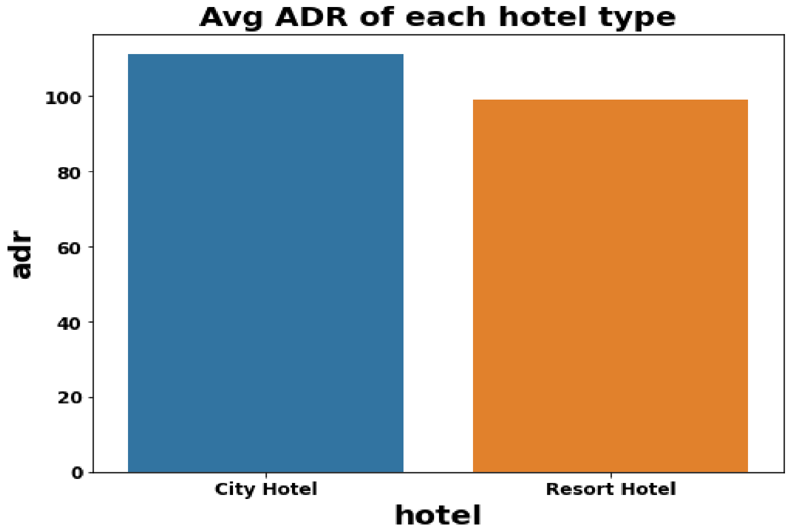
- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Full Board i.e. FB is least preferred.
- HB (Half Board) and SC(Self Catering) are equally preferred.

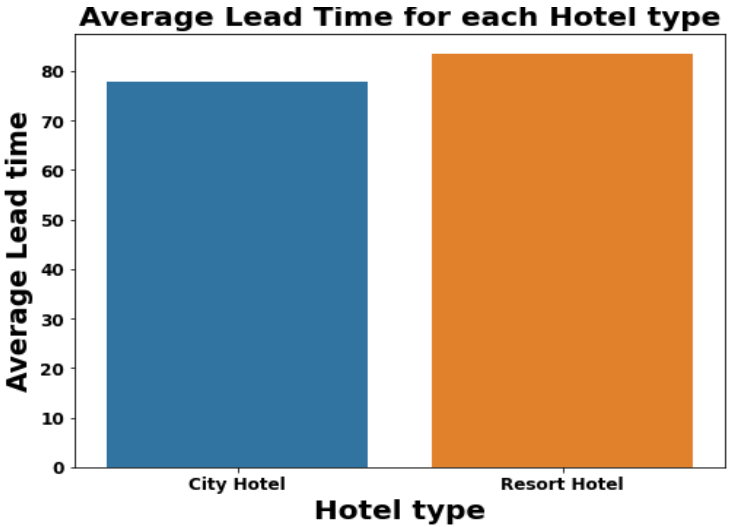


- As we can see in the line chart,from June to September most of the bookings happened. It’s Summer time. After September bookings Starts declining.



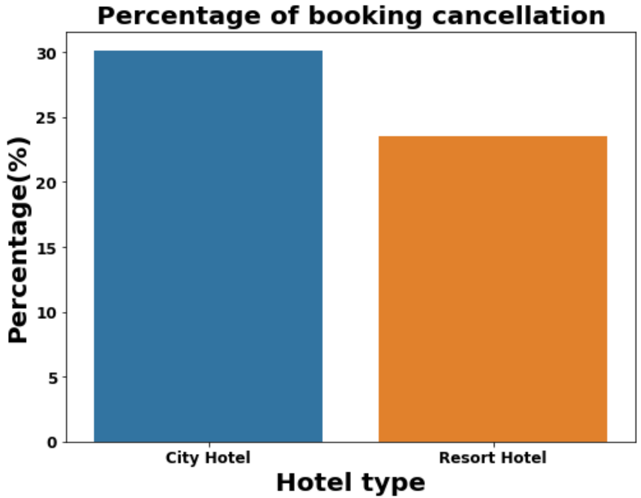
Bivariate and Multivariate Analysis



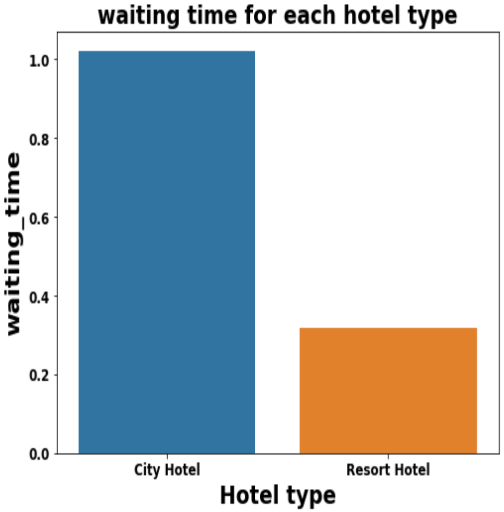


Conclusions:

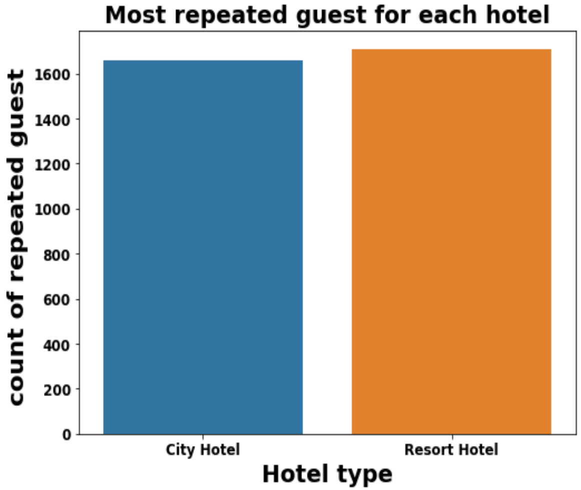
- ➢ Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- ➢ Average lead time for resort hotel is high. It means people plan their trip too early. Usually people prefer resort hotels for longer stays. That’s why people plan early



Percentage of booking cancellation



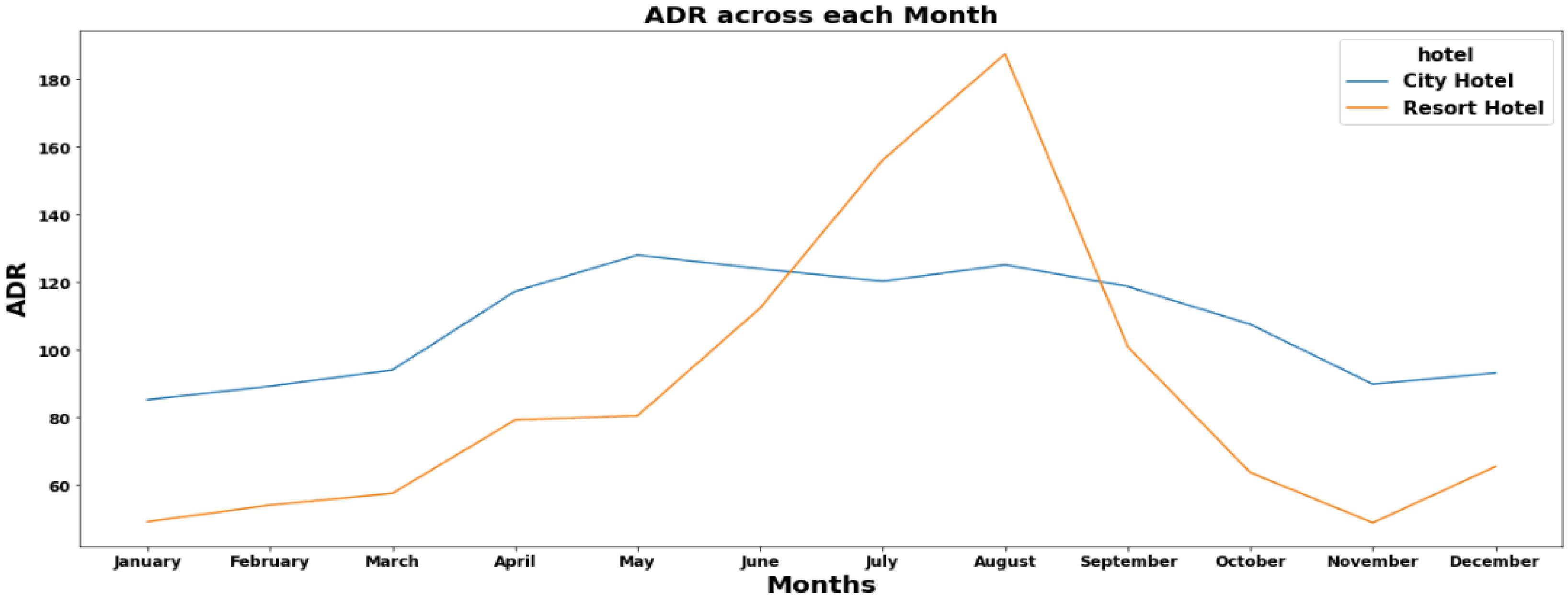
waiting time for each hotel type



Most repeated guest for each hotel

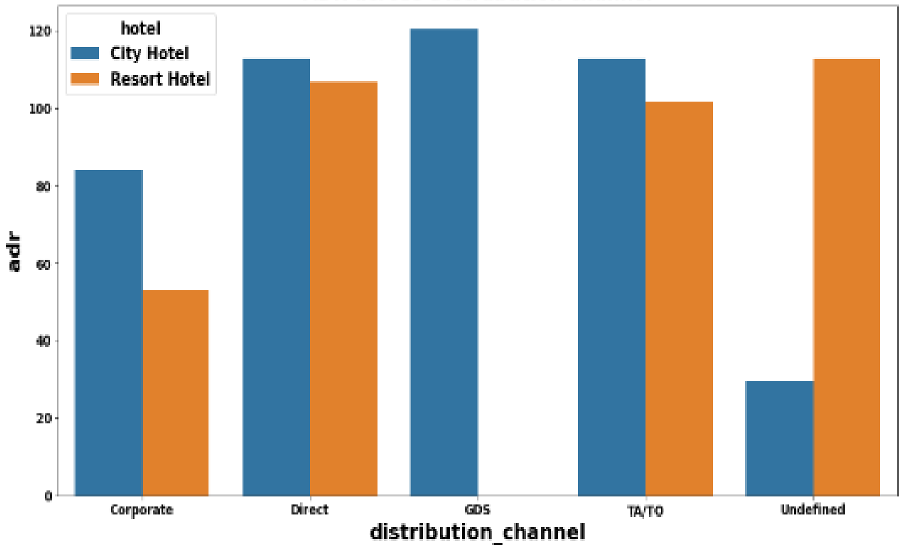
Conclusions:

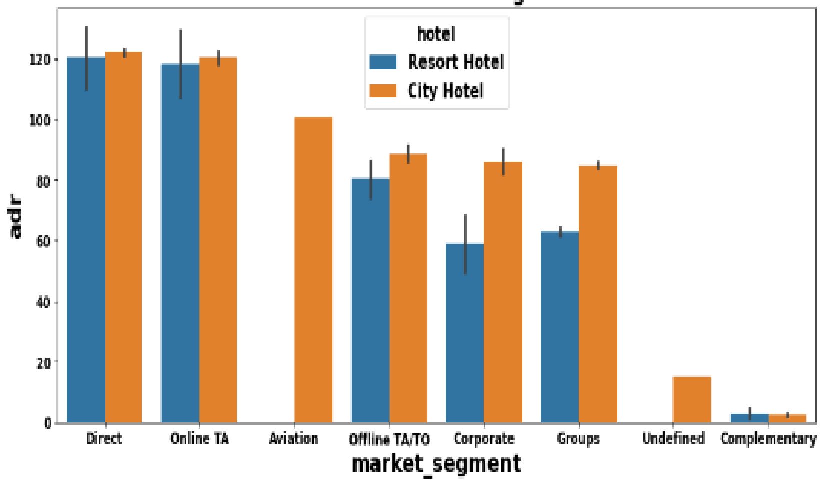
- Booking cancellation rate is high for City hotels which almost 30 %.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Resort hotels has the most repeated guests. In order to get increase the count of repeated guests hotel management need to take the valuable feedbacks from the guests and try to give good service.



Conclusions:

- Resort hotels had the highest adr in June ,July and August than the City hotels. But in other months adr of Resort hotel was less than the City hotels.
- Thus we can say that, the January, February, March, April ,November and December are the good months for customers to get good adr


ADR across Distrubution channel


Adr across market segment

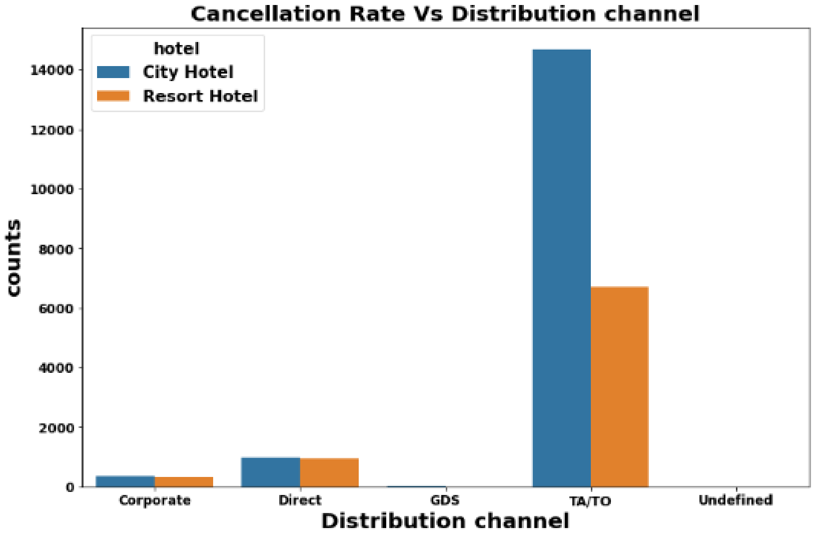
Conclusions:

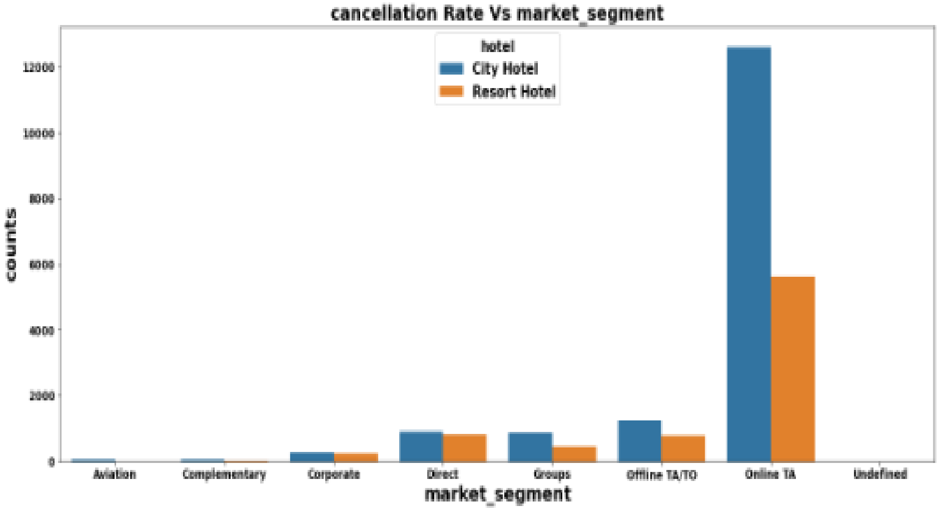
Distribution channel:

- 'Direct' and 'TA/TO' has almost equal adr in both type of hotels which is high among other channels.
- GDS has high adr in 'City Hotel' type. GDS needs to increase Resort Hotel bookings. From this we can say that “Direct” and ‘TA/TO’ are generating more revenue than the other channels.

Market Segment:

- Here “Direct” and ‘Online Travel Agency’ has high adr for both hotel types. Aviation segment needs to increase Resort hotel bookings.





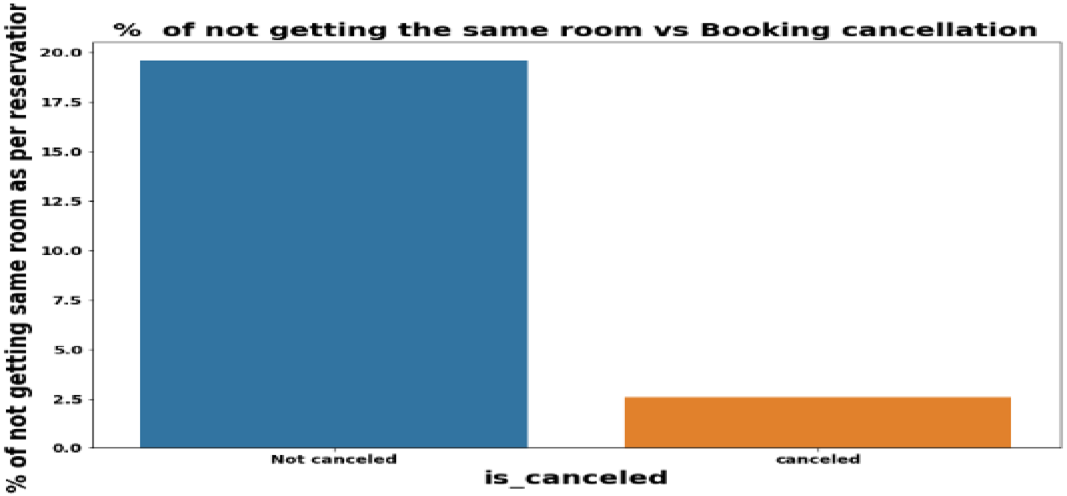
Conclusions:

Distribution channel:

➢ ‘TA/TO’ distribution channel has highest cancellations for city hotels and more than 6000 cancellations for resort hotels. In order to reduce the cancellations they should improve their cancellation policies and deposit policies.

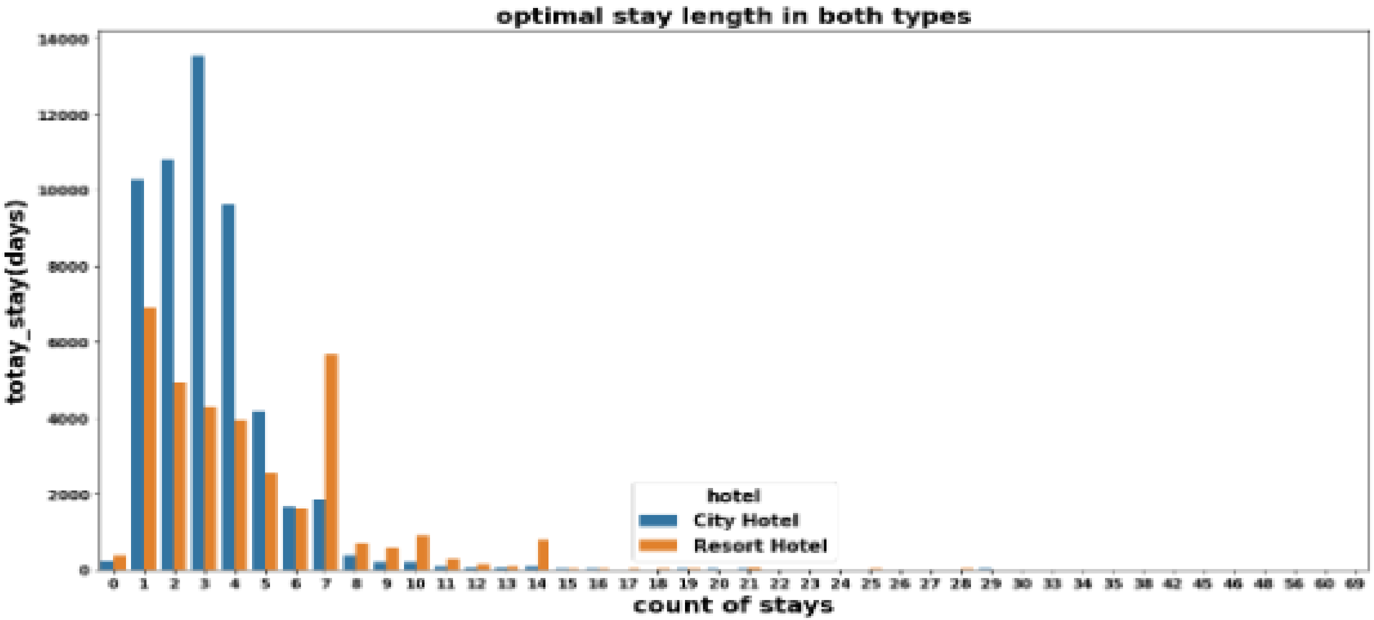
Market Segment:

➢ ‘Online TA/TO’ market segment has highest cancellations for city hotels.



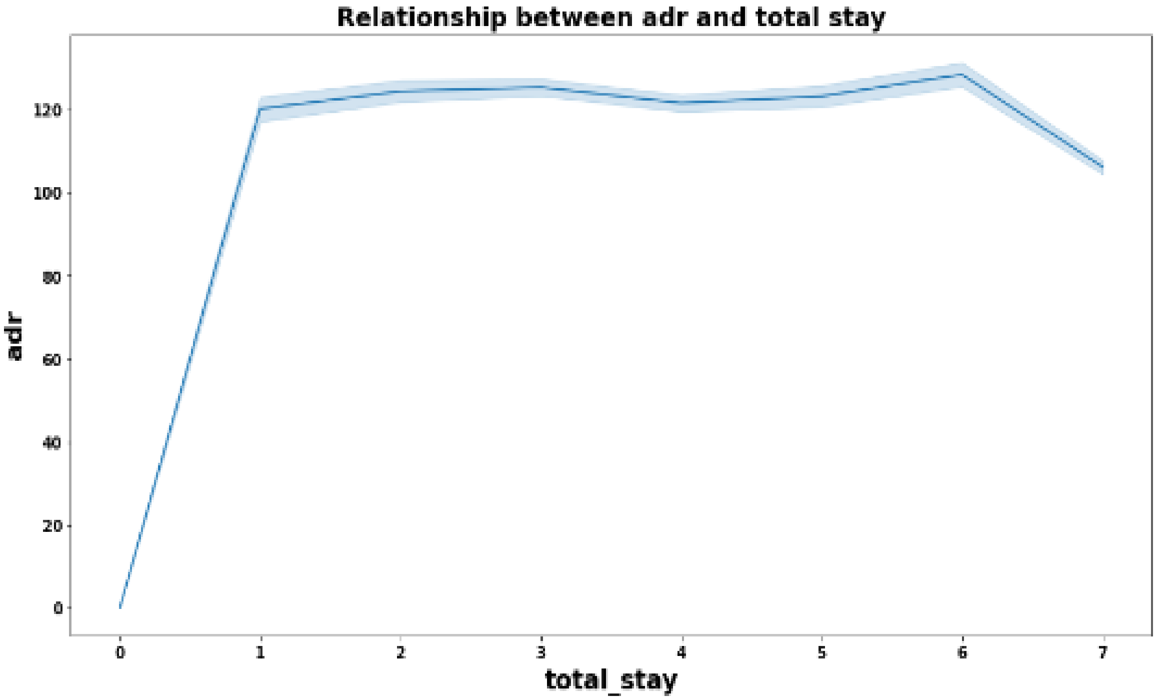
Conclusions:

- ➢ Almost 19 % people did not canceled their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.
- ➢ Thus not getting the same room as per reserved room is not the reason for booking cancellations.



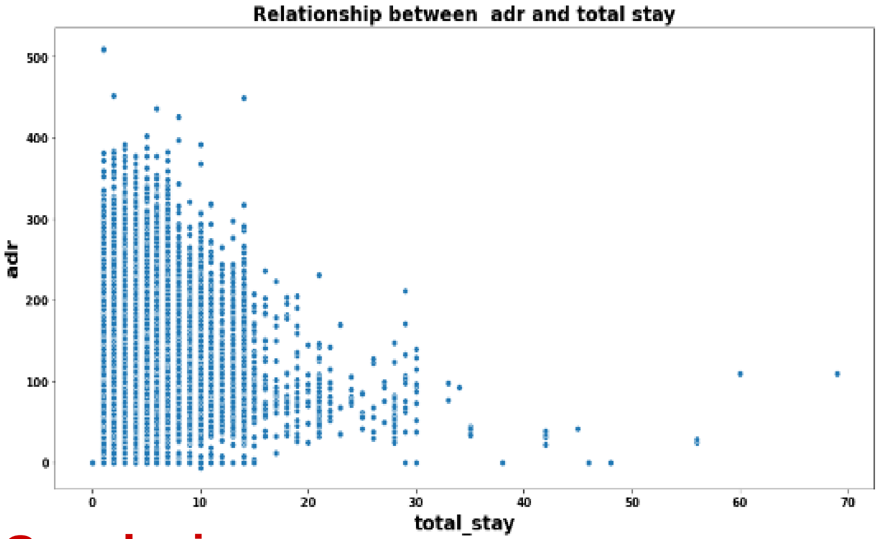
Conclusions:

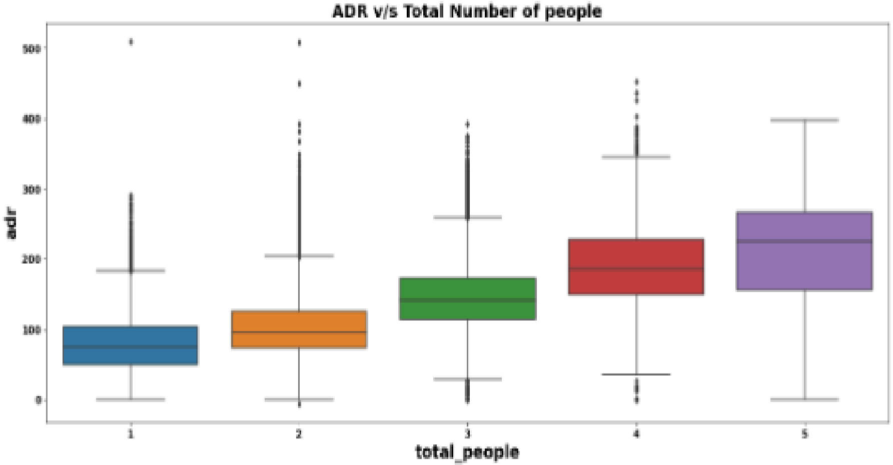
- Optimal stay in both the type hotel is less than 7 days.



Conclusions:

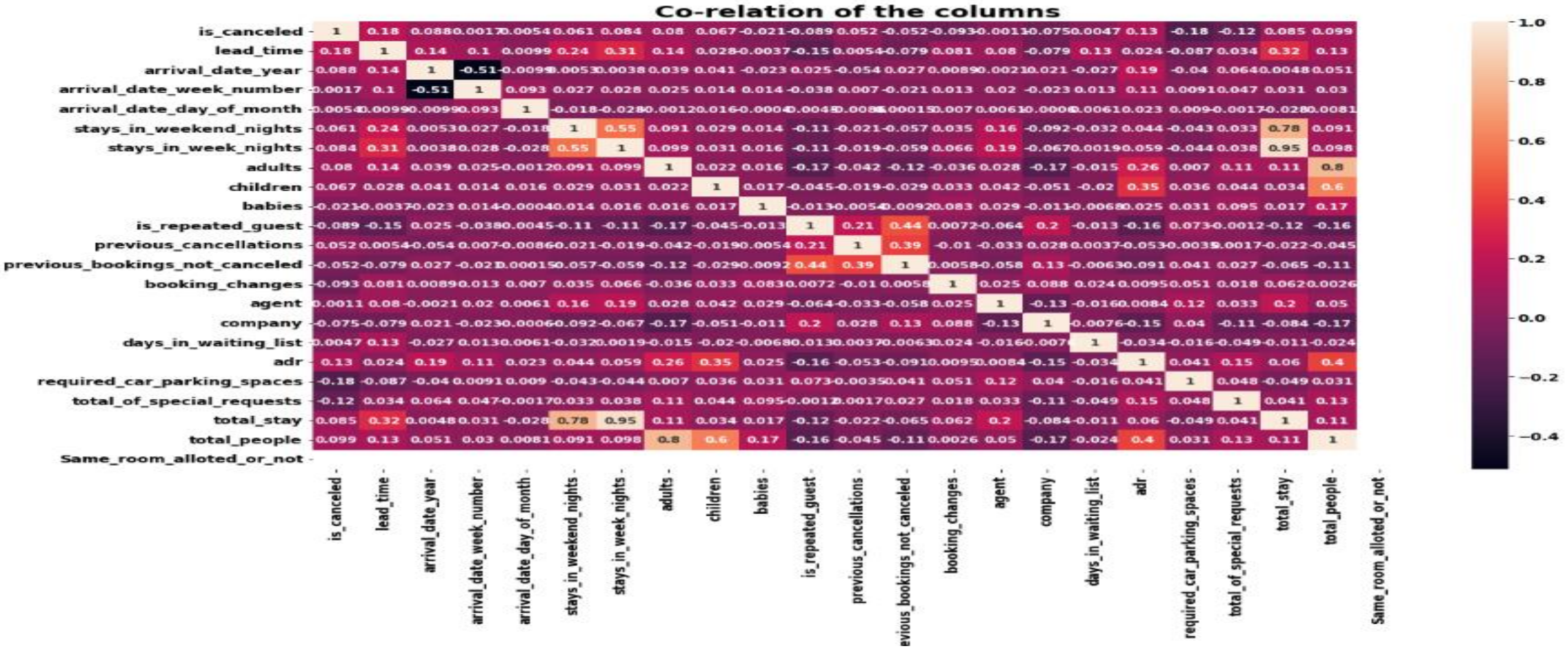
- As the total stay increases the adr also increases.





Conclusions:

- As the total number of people increases adr also increases.
- Thus adr and total people are directly proportional to each other.
- From above scatter we can say that as the stay increases adr is decreasing. Thus for longer stays customer can get good adr.



	is_canceled	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	is_repeated_guest	previous_cancellations	evious_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_stay	total_people	Same_room_alloted_or_not
is_canceled	1	0.18	0.088	0.0017	0.0054	0.061	0.084	0.08	0.067	-0.021	-0.089	0.052	-0.052	-0.093	0.0011	-0.075	0.0047	0.13	-0.18	-0.12	0.085	0.099	
lead_time	0.18	1	0.14	0.1	0.0099	0.24	0.31	0.14	0.028	-0.0037	-0.15	0.0054	-0.079	0.081	0.08	-0.079	0.13	0.024	-0.087	0.034	0.32	0.13	
arrival_date_year	0.088	0.14	1	-0.51	-0.0099	0.0053	0.0038	0.039	0.041	-0.023	0.025	-0.054	0.027	0.0089	0.0021	0.021	-0.027	0.19	-0.04	0.064	0.0048	0.051	
arrival_date_week_number	0.0017	0.1	-0.51	1	0.093	0.027	0.028	0.025	0.014	0.014	-0.038	0.007	-0.021	0.013	0.02	-0.023	0.013	0.11	0.0091	0.047	0.031	0.03	
arrival_date_day_of_month	0.0054	0.0099	0.0099	0.093	1	-0.018	-0.028	0.0012	0.016	-0.0004	0.0045	0.0086	0.00015	0.007	0.0061	0.0006	0.0061	0.023	0.009	-0.0017	-0.028	0.0081	
stays_in_weekend_nights	0.061	0.24	0.0053	0.027	-0.018	1	0.55	0.091	0.029	0.014	-0.11	-0.021	-0.057	0.035	0.16	-0.092	-0.032	0.044	-0.043	0.033	0.78	0.091	
stays_in_week_nights	0.084	0.31	0.0038	0.028	-0.028	0.55	1	0.099	0.031	0.016	-0.11	-0.019	-0.059	0.066	0.19	-0.067	0.0019	0.059	-0.044	0.038	0.95	0.098	
adults	0.08	0.14	0.039	0.025	-0.0012	0.091	0.099	1	0.022	0.016	-0.17	-0.042	-0.12	-0.036	0.028	-0.17	-0.015	0.26	0.007	0.11	0.11	0.8	
children	0.067	0.028	0.041	0.014	0.016	0.029	0.031	0.022	1	0.017	-0.045	-0.019	-0.029	0.033	0.042	-0.051	-0.02	0.35	0.036	0.044	0.034	0.6	
babies	-0.021	-0.0037	-0.023	0.014	-0.0004	0.014	0.016	0.016	0.017	1	-0.013	-0.0054	-0.0092	0.083	0.029	-0.011	-0.0068	0.025	0.031	0.095	0.017	0.17	
is_repeated_guest	-0.089	-0.15	0.025	-0.038	0.0045	-0.11	-0.11	-0.17	-0.045	-0.013	1	0.21	0.44	0.0072	-0.064	0.2	-0.013	-0.16	0.073	-0.0012	-0.12	-0.16	
previous_cancellations	0.052	0.0054	-0.054	0.007	-0.0086	-0.021	-0.019	-0.042	-0.019	0.0054	0.21	1	0.39	-0.01	-0.033	0.028	0.0037	-0.053	-0.0035	0.0017	-0.022	-0.045	
previous_bookings_not_canceled	-0.052	-0.079	0.027	-0.021	0.00015	-0.057	-0.059	-0.12	-0.029	-0.0092	0.44	0.39	1	0.0058	-0.058	0.13	-0.0063	-0.091	0.041	0.027	-0.065	-0.11	
booking_changes	-0.093	0.081	0.0089	0.013	0.007	0.035	0.066	-0.036	0.033	0.083	0.0072	-0.01	0.0058	1	0.025	0.088	0.024	0.0095	0.051	0.018	0.062	0.0026	
agent	0.0011	0.08	-0.0021	0.02	0.0061	0.16	0.19	0.028	0.042	0.029	-0.064	-0.033	-0.058	0.025	1	-0.13	-0.016	0.0084	0.12	0.033	0.2	0.05	
company	-0.075	-0.079	0.021	-0.023	0.0006	-0.092	-0.067	-0.17	-0.051	-0.011	0.2	0.028	0.13	0.088	-0.13	1	0.0076	-0.15	0.04	-0.11	-0.084	-0.17	
days_in_waiting_list	0.0047	0.13	-0.027	0.013	0.0061	-0.032	0.0019	-0.015	-0.02	-0.0068	0.013	0.0037	0.0063	0.024	-0.016	0.0076	1	-0.034	-0.016	-0.049	-0.011	-0.024	
adr	0.13	0.024	0.19	0.11	0.023	0.044	0.059	0.26	0.35	0.025	-0.16	-0.053	-0.091	0.0095	0.0084	-0.15	-0.034	1	0.041	0.15	0.06	0.4	
required_car_parking_spaces	-0.18	-0.087	-0.04	0.0091	0.009	-0.043	-0.044	0.007	0.036	0.031	0.073	-0.0035	0.041	0.051	0.12	0.04	-0.016	0.041	1	0.048	-0.049	0.031	
total_of_special_requests	-0.12	0.034	0.064	0.047	-0.0017	0.033	0.038	0.11	0.044	0.095	-0.0012	0.0017	0.027	0.018	0.033	-0.11	-0.049	0.15	0.048	1	0.041	0.13	
total_stay	0.085	0.32	0.0048	0.031	-0.028	0.78	0.95	0.11	0.034	0.017	-0.12	-0.022	-0.065	0.062	0.2	-0.084	-0.011	0.06	-0.049	0.041	1	0.11	
total_people	0.099	0.13	0.051	0.03	0.0081	0.091	0.098	0.8	0.6	0.17	-0.16	-0.045	-0.11	0.0026	0.05	-0.17	-0.024	0.4	0.031	0.13	0.11	1	
Same_room_alloted_or_not																							

Conclusions:

- is_canceled and same_room_alloted_or_not are negatively corelated. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.
- lead_time and total_stay is positively correlated. That means more is the stay of customer more will be the lead time.
- adults, childrens and babies are corelated to each other. That means more the people more will be adr.
- is_repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.

THANK YOU