CAPSTONE PROJECT

HOTEL BOOKING EDA ANALYSIS

Points to Discuss:

- Agenda
- Data Summary
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Correlation Analysis
- Hypothesized Question Analysis
- Conclusions and Summary

AGENDA

This project comprises a real-world data record of hotel bookings for a city and a resort hotel from 2015 to 2017, including details such as bookings, cancellations, and guest information. The project's main goal is to comprehend and visualize data from the hotel and customer perspectives to get the proper insights from the dataset and to make proper data driven decisions.

- **We have performed various kind of analysis to manage the data in terms of data wrangling.**

- **Wrangled our raw data to proper formatted data to perform the analysis to get desired results.**

DATA SUMMARY

- The “Hotel Booking” dataset contains very important and leading information regarding a customer prone to cancel his/her hotel booking or not.

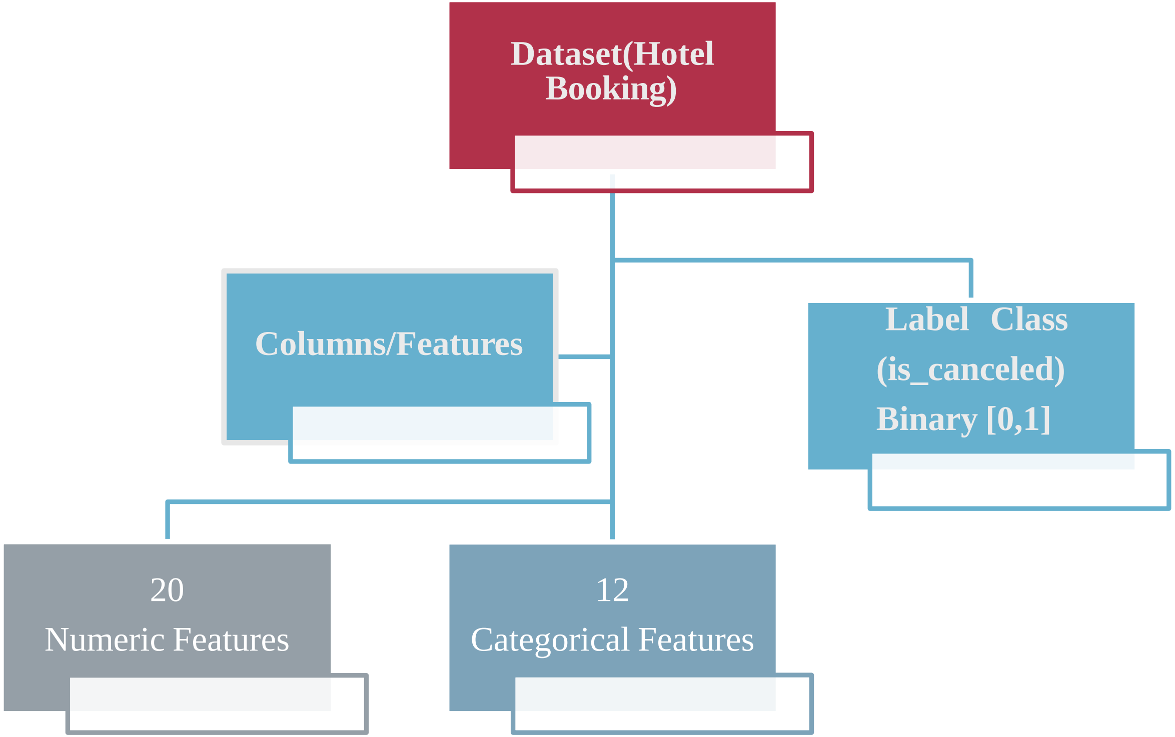
- There are many features which have very explanatory information which helps us to get to exact flow of hotel booking occurrence.

- Some of the data fields are defined in further slides to get the proper understanding of different kind of important feature/columns:

Features/Columns	Explanation
hotel	There are two types of hotels: resort hotels and city hotels.
is_cancelled	The cancellation type is indicated by the value 0 and 1 of the column. 1 -> Not Cancelled 2 -> Canceled
lead_time	The period between making a reservation and arriving.
stayed_in_weekend_nights	The number of weekend nights a reservation can stay for.
stayed_in_weekday_nights	Per reservation, the amount of weekday nights to stay.
market_segment	This column explains how reservations are made and why they are made.
distribution_channel	The medium of booking.

Features/Columns	Explanation
is_repeated_guest	Indicates if the guest was the one who arrived earlier or not. 1 -> Repeated 2 -> Not Repeated
days_in_waiting_list	The time between making a reservation and completing a transaction is measured in days.
customer_type	Type of customers(Transient, group, etc.)
reserved_room_type	The code representing the room type assigned to the reservation. Due to hotel operations, the assigned room type may differ from the reserved room type.
adr	column represents the average daily rate
assigned_room_type	Code for the type of room assigned to the booking.

DATA SUMMARY



Dataset(Hotel Booking)

Columns/Features

Label Class (is_canceled) Binary [0,1]

20
Numeric Features

12
Categorical Features

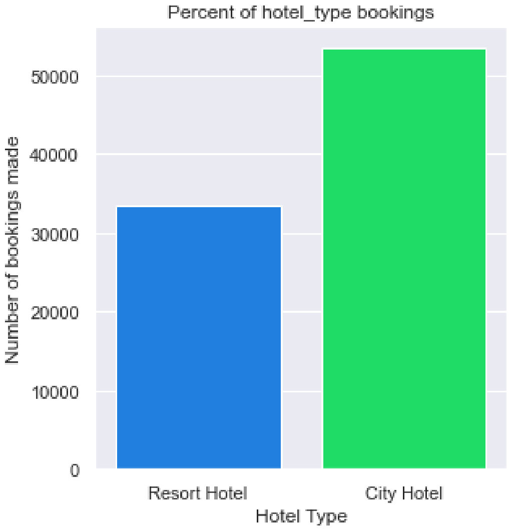
UNIVARIATE ANALYSIS

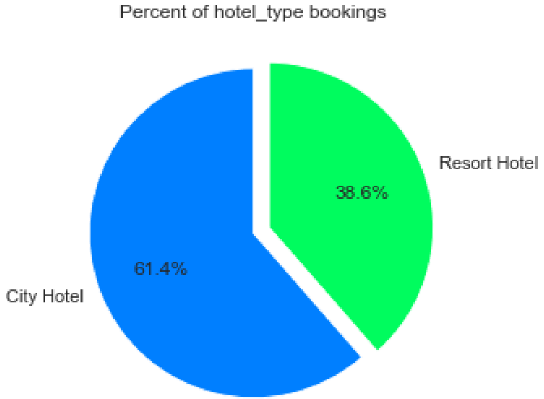
- Because “uni” means one and variate means variable, there is only one trustworthy variable in univariate analysis. The goal of univariate analysis is to derive data, characterize and summaries it, and examine any patterns that may exist. It investigates each variable separately in a dataset. There are two types of variables that can be used: categorical and numerical.

- We have performed a lot of univariate analysis on various features:

1. **Observation from hotel type that which type of hotel has how many bookings**
2. **The count of bookings is_canceled**
3. **Observation on Market Segment wise bookings**
4. **Observation on Distribution Channel wise bookings**

Hotel Type Booking Count

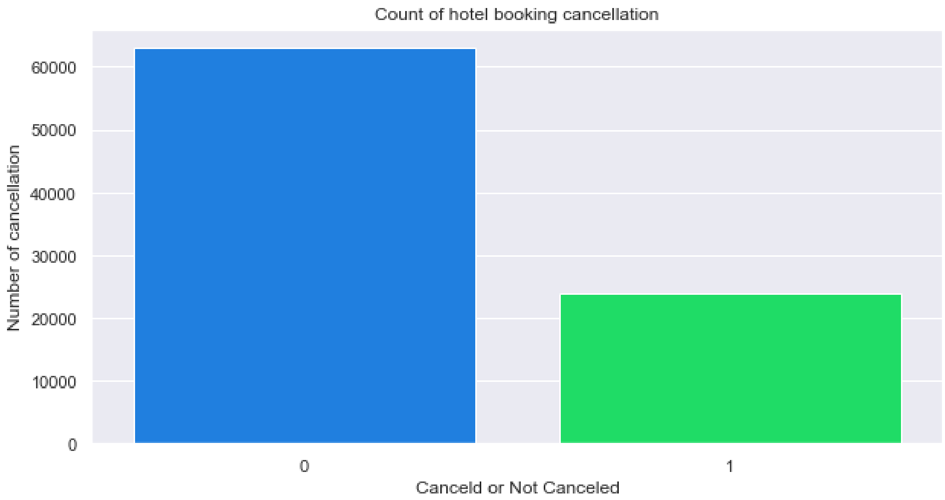




- City hotel has bookings above 50000 which is around 61.4% and Resort hotel has less than 35000 which is around 38.6% bookings

- Which implies that City Hotel has more number of bookings than Resort Hotel

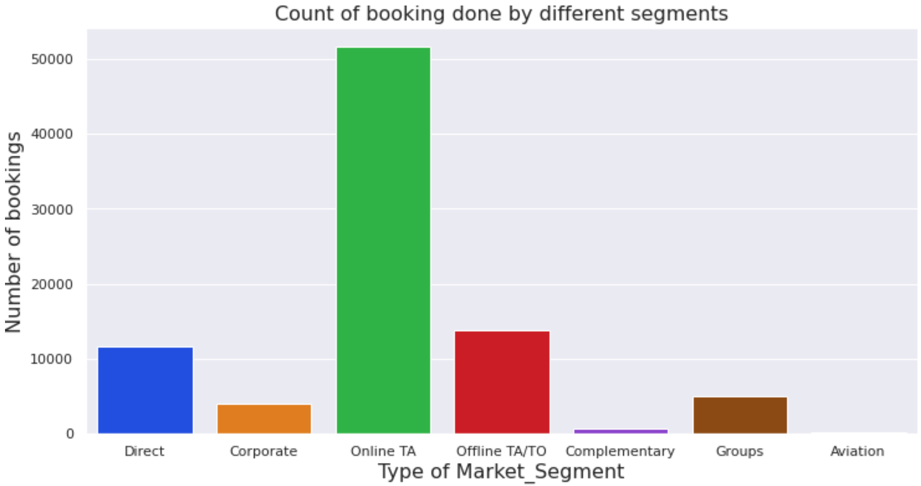
Counting ratio of cancelled and not cancelled bookings



- 0 ---> Non Cancelled Booking | 1 ---> Cancelled Booking

- Hotels are less cancelled as compare to cancellation bar

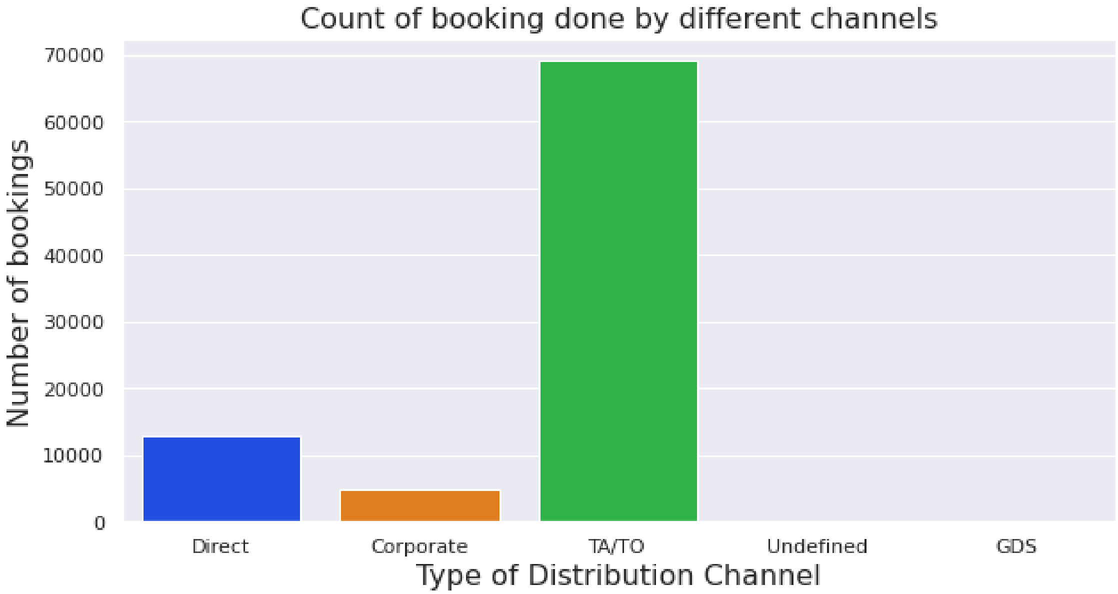
Market Segment Booking Count



- Online TA --> Online Travel Agent
- Offline TA/TO--> Offline Travel Agent and Travel Operator

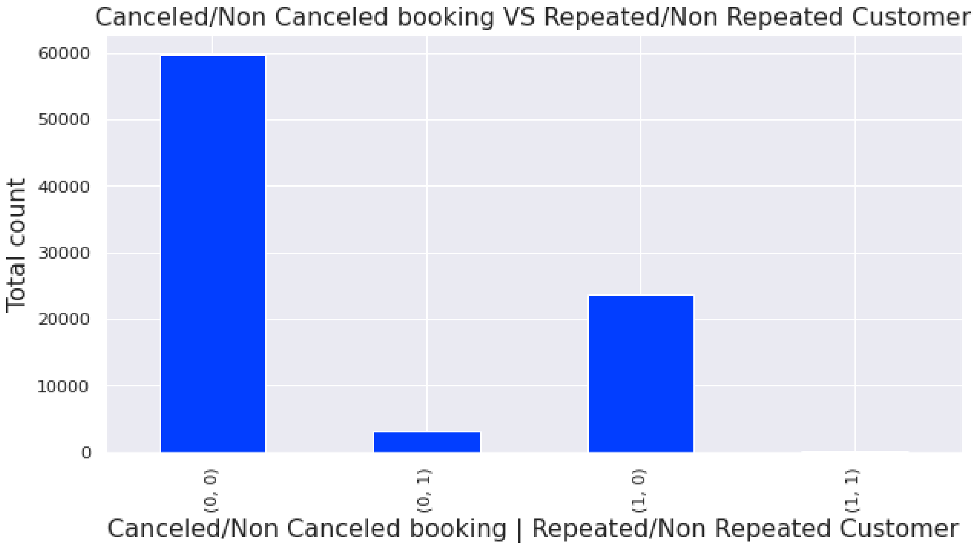
- Online TA and Offline TA/TO has made to maximum number of hotel bookings

Distribution Channel Booking Count



- Offline TA/TO has made to maximum number of hotel bookings

Repeated Guest count who cancelled the booking



- The lowest cancellation occurs when the customer is repeated
- That means the existing customers are very less likely to cancel the booking

BIVARIATE ANALYSIS

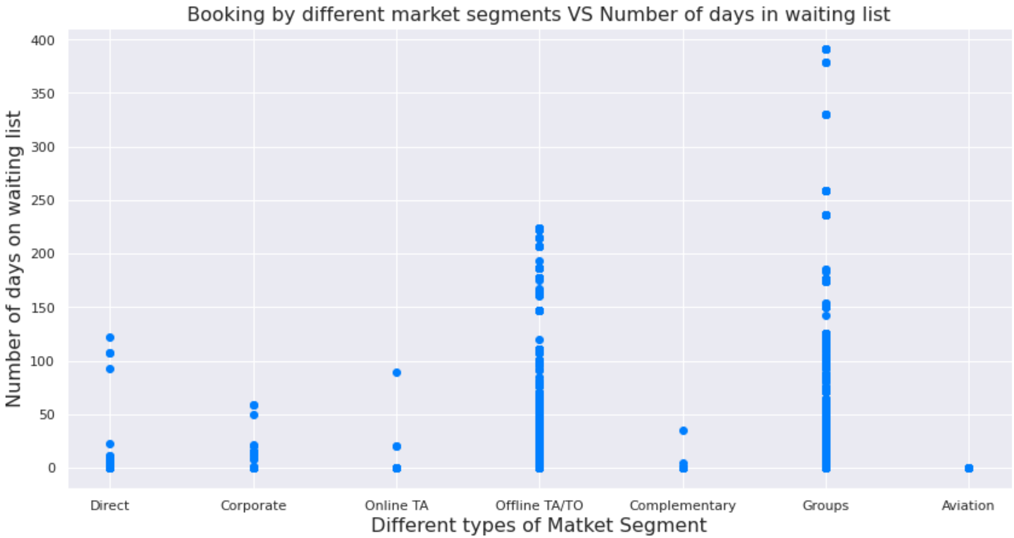
There are two variables since bi means two and variate implies variable. The investigation focuses on the root of the problem and the relationship between the two factors. Bivariate analysis can be divided into three categories.

Analysed different features together and we got interesting insights about the data.

In this data we have done the following steps mentioned below

- **Market segment and days on waiting list for each of them**
- **Comparing Distribution Channels and days on waiting list for each of them**
- **Per month arrivals hotel bookings**
- **Market Segment wise bookings for each Hotel Type**
- **Distribution Channel wise bookings for each Hotel Type**
- **Months of the year with the lead time**
- **Arrival of customers per day of the months to the hotels**
- **Demand of parking space by customers in each type of hotels**

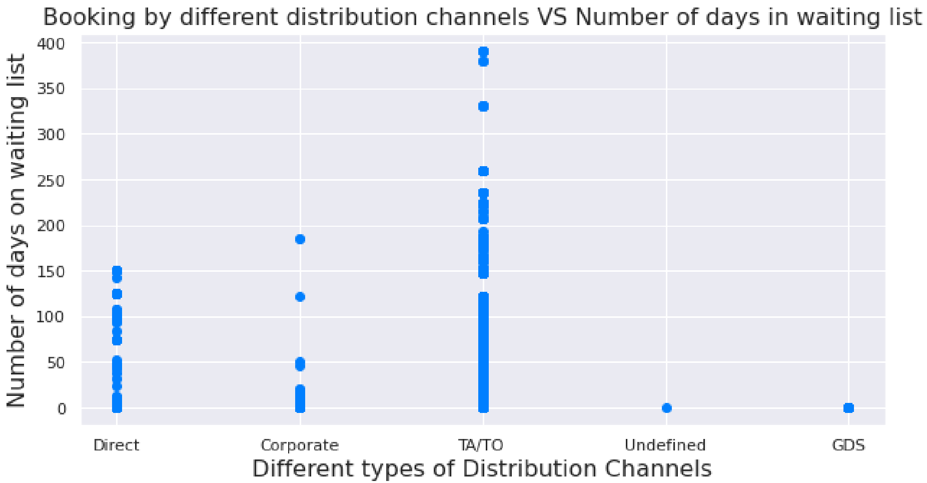
MARKET SEGMENT AND DAYS ON WAITING LIST



- The aviation industry has the minimum number of days in waiting list.

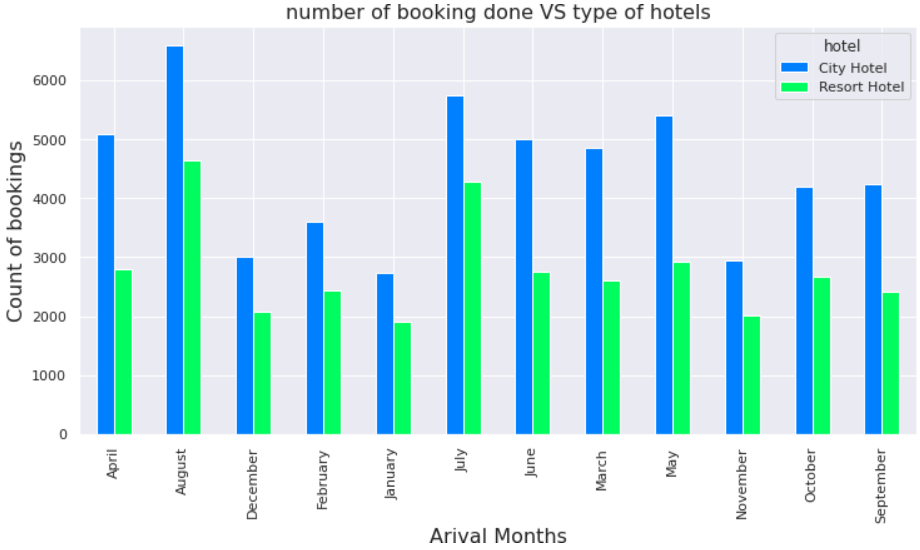
- It might due to when a flight landed they have to provide immediate accommodation for there working staff like pilots and air hostess and they do not go for that hotels which put them into higher days on waiting list.

DISTRIBUTION CHANNELS AND DAYS ON WAITING LIST



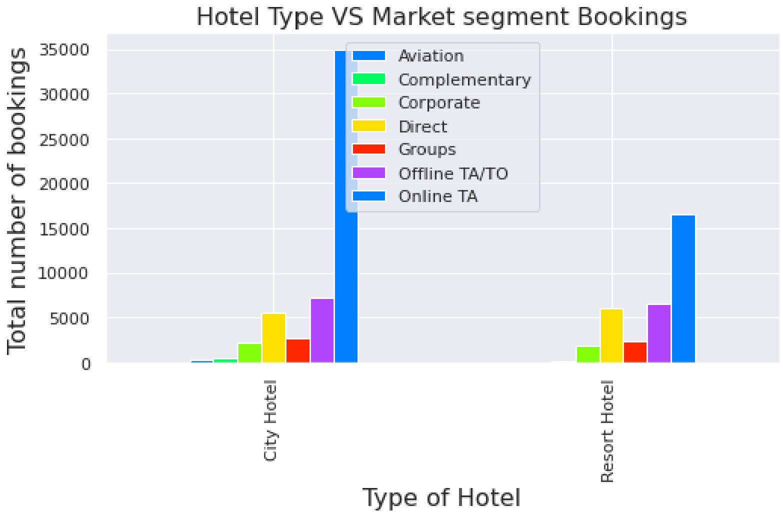
Here we can also see that the TA -> Travel Agent and TO -> Travel Operator have the higher number on days on waiting list

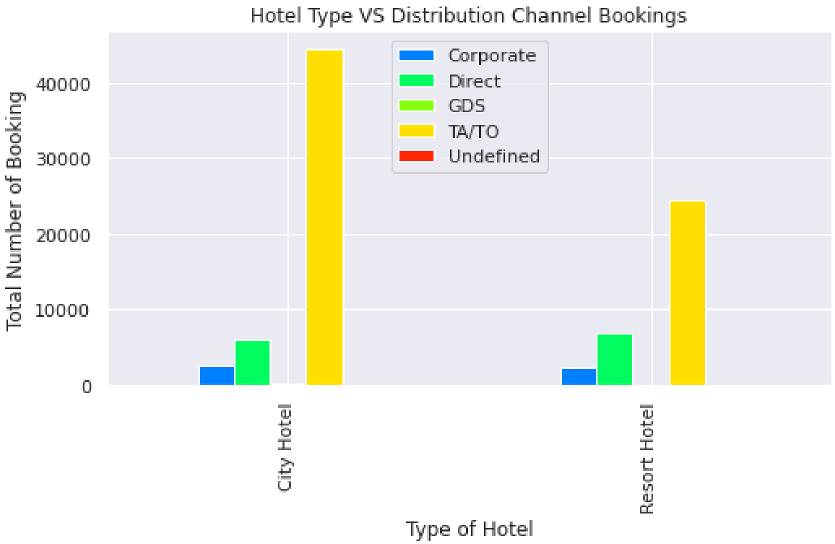
PER-MONTH ARRIVALS HOTEL BOOKINGS



- Number of bookings are highest for city hotels in the month of August, July and May.
- And the Resort hotel has also number booking higher in the month of July and August.

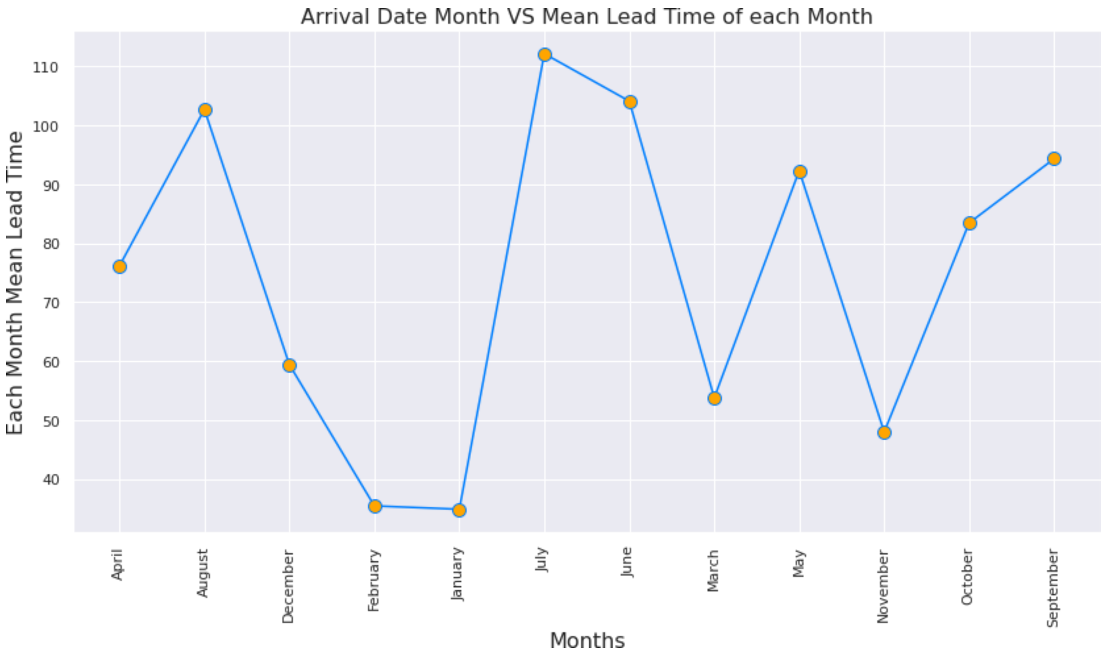
MARKET SEGMENT/DISTRIBUTION CHANNEL WISE BOOKINGS FOR EACH HOTEL TYPE





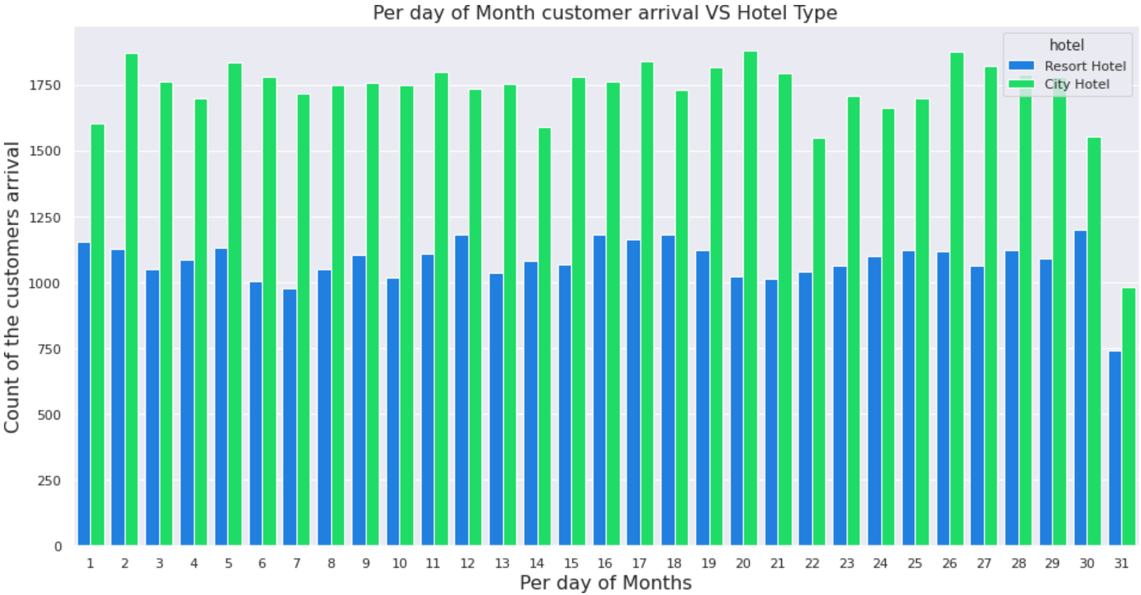
- Maximum number of booking is done by the Online TA(Travel Agent) for the City Hotels as well as for the Resort Hotels.
- The most number of booking is done by the TA/TO Distribution Channels for both City and Resort Hotel

MONTHS OF THE YEAR WITH THE LEAD TIME



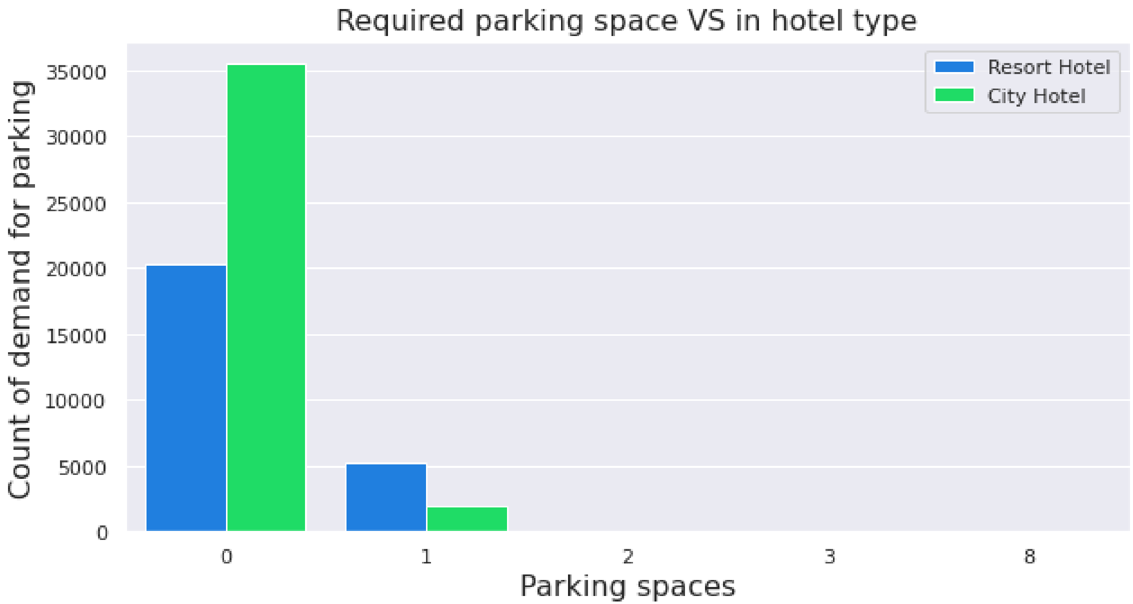
- The lead time for the month of July is very high and for January and February Month is the least.

ARRIVAL OF CUSTOMERS PER DAY OF THE MONTHS TO THE HOTELS



- The least number of bookings are occurred in the month ends.

DEMAND OF PARKING SPACE BY CUSTOMERS IN EACH TYPE OF HOTELS



- This chart tells us that most number of the guest has no demand for parking space in both the hotel type.

MULTIVARIATE ANALYSIS

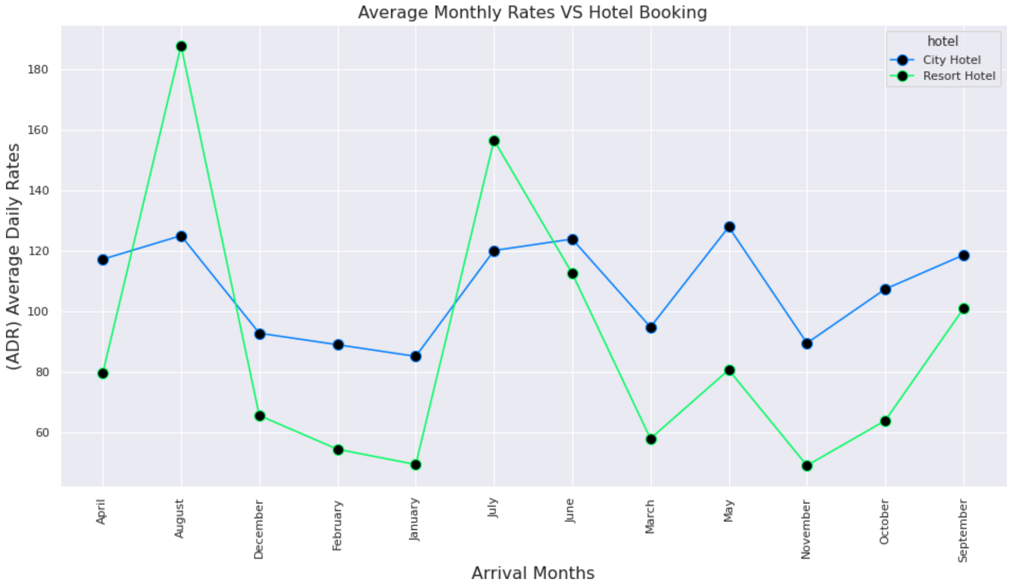
- Multivariate indicates that numerous dependent variables are combined to produce a single result. This explains why the vast majority of real-world situations are multivariate.

- We have done Multivariate analysis to check hotel bookings occurred according to the average daily rates per month.

- By performing this multivariate analysis we came to know about how different hotel type has impacted in each month in terms of average daily rates.

- It explains us that how customers are prone to book hotels which month they prefer and when not.

HOTEL BOOKINGS OCCURRED ACCORDING TO THE AVERAGE DAILY RATES PER MONTH



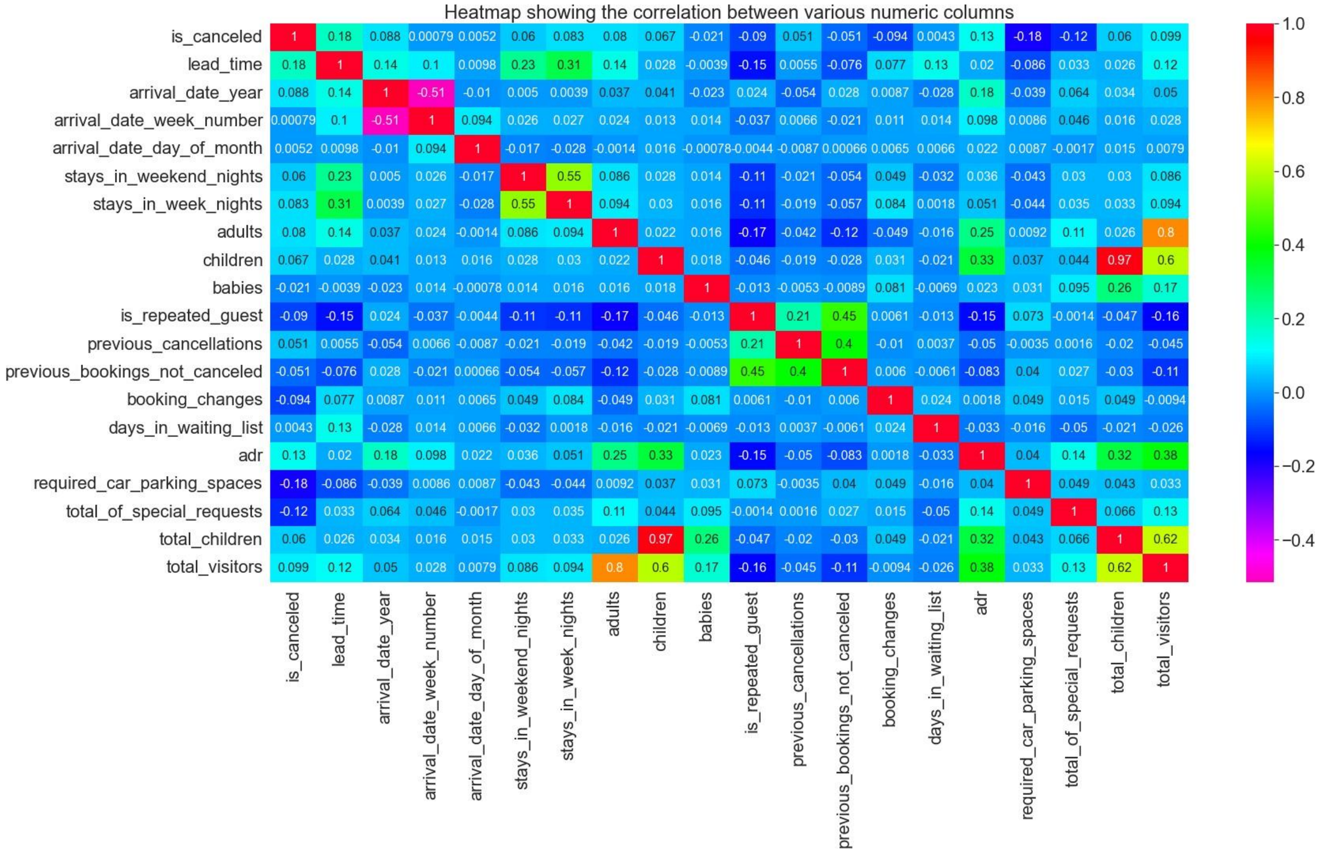
- The average daily rate is more expensive during June, July, August and September for Resort Hotels
- The average daily rate is more expensive during June, July and May for City Hotels

Correlation is a statistical term that expresses how closely two variables are related in a linear fashion (meaning they change together at a constant rate). It's a typical way of describing simple relationships without stating a cause and effect relationship.

- The negative correlation coefficient between the data pair LEAD TIME versus ADR indicates that as the time between room booking and check-in rises, the ADR for that hotel room will fall by a modest amount..

- is_canceled is not influenced but little bit by lead_time feature.

Correlation Matrix:



Heatmap showing the correlation between various numeric columns

	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	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_children	total_visitors
is_canceled	1	0.18	0.088	0.00079	0.0052	0.06	0.083	0.08	0.067	-0.021	-0.09	0.051	-0.051	-0.094	0.0043	0.13	-0.18	-0.12	0.06	0.099
lead_time	0.18	1	0.14	0.1	0.0098	0.23	0.31	0.14	0.028	-0.0039	-0.15	0.0055	-0.076	0.077	0.13	0.02	-0.086	0.033	0.026	0.12
arrival_date_year	0.088	0.14	1	-0.51	-0.01	0.005	0.0039	0.037	0.041	-0.023	0.024	-0.054	0.028	0.0087	-0.028	0.18	-0.039	0.064	0.034	0.05
arrival_date_week_number	0.00079	0.1	-0.51	1	0.094	0.026	0.027	0.024	0.013	0.014	-0.037	0.0066	-0.021	0.011	0.014	0.098	0.0086	0.046	0.016	0.028
arrival_date_day_of_month	0.0052	0.0098	-0.01	0.094	1	-0.017	-0.028	-0.0014	0.016	-0.00078	-0.0044	-0.0087	0.00066	0.0065	0.0066	0.022	0.0087	-0.0017	0.015	0.0079
stays_in_weekend_nights	0.06	0.23	0.005	0.026	-0.017	1	0.55	0.086	0.028	0.014	-0.11	-0.021	-0.054	0.049	-0.032	0.036	-0.043	0.03	0.03	0.086
stays_in_week_nights	0.083	0.31	0.0039	0.027	-0.028	0.55	1	0.094	0.03	0.016	-0.11	-0.019	-0.057	0.084	0.0018	0.051	-0.044	0.035	0.033	0.094
adults	0.08	0.14	0.037	0.024	-0.0014	0.086	0.094	1	0.022	0.016	-0.17	-0.042	-0.12	-0.049	-0.016	0.25	0.0092	0.11	0.026	0.8
children	0.067	0.028	0.041	0.013	0.016	0.028	0.03	0.022	1	0.018	-0.046	-0.019	-0.028	0.031	-0.021	0.33	0.037	0.044	0.97	0.6
babies	-0.021	-0.0039	-0.023	0.014	-0.00078	0.014	0.016	0.016	0.018	1	-0.013	-0.0053	-0.0089	0.081	-0.0069	0.023	0.031	0.095	0.26	0.17
is_repeated_guest	-0.09	-0.15	0.024	-0.037	-0.0044	-0.11	-0.11	-0.17	-0.046	-0.013	1	0.21	0.45	0.0061	-0.013	-0.15	0.073	-0.0014	-0.047	-0.16
previous_cancellations	0.051	0.0055	-0.054	0.0066	-0.0087	-0.021	-0.019	-0.042	-0.019	-0.0053	0.21	1	0.4	-0.01	0.0037	-0.05	-0.0035	0.0016	-0.02	-0.045
previous_bookings_not_canceled	-0.051	-0.076	0.028	-0.021	0.00066	-0.054	-0.057	-0.12	-0.028	-0.0089	0.45	0.4	1	0.006	-0.0061	-0.083	0.04	0.027	-0.03	-0.11
booking_changes	-0.094	0.077	0.0087	0.011	0.0065	0.049	0.084	-0.049	0.031	0.081	0.0061	-0.01	0.006	1	0.024	0.0018	0.049	0.015	0.049	-0.0094
days_in_waiting_list	0.0043	0.13	-0.028	0.014	0.0066	-0.032	0.0018	-0.016	-0.021	-0.0069	-0.013	0.0037	-0.0061	0.024	1	-0.033	-0.016	-0.05	-0.021	-0.026
adr	0.13	0.02	0.18	0.098	0.022	0.036	0.051	0.25	0.33	0.023	-0.15	-0.05	-0.083	0.0018	-0.033	1	0.04	0.14	0.32	0.38
required_car_parking_spaces	-0.18	-0.086	-0.039	0.0086	0.0087	-0.043	-0.044	0.0092	0.037	0.031	0.073	-0.0035	0.04	0.049	-0.016	0.04	1	0.049	0.043	0.033
total_of_special_requests	-0.12	0.033	0.064	0.046	-0.0017	0.03	0.035	0.11	0.044	0.095	-0.0014	0.0016	0.027	0.015	-0.05	0.14	0.049	1	0.066	0.13
total_children	0.06	0.026	0.034	0.016	0.015	0.03	0.033	0.026	0.97	0.26	-0.047	-0.02	-0.03	0.049	-0.021	0.32	0.043	0.066	1	0.62
total_visitors	0.099	0.12	0.05	0.028	0.0079	0.086	0.094	0.8	0.6	0.17	-0.16	-0.045	-0.11	-0.0094	-0.026	0.38	0.033	0.13	0.62	1

HYPOTHESIZED QUESTIONS TO CROSS-CHECK SOME ANOTHER ASPECTS OF OUR ANALYSIS:

1. Do customers have any preference with or without children’s.

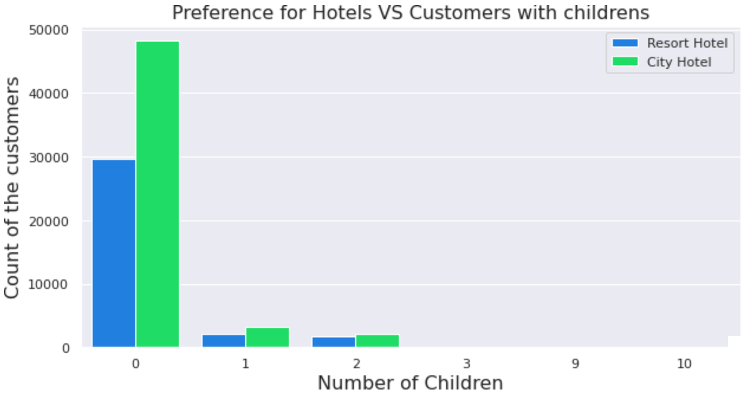
2. How daily average rate is impacting the reserved room type in hotels.

3. Does total stay per month impacted by customers with and without children's.

4. Do reservation status impacted by type of customers.

5. Do customers cancel their booking if they are allotted with different room type.

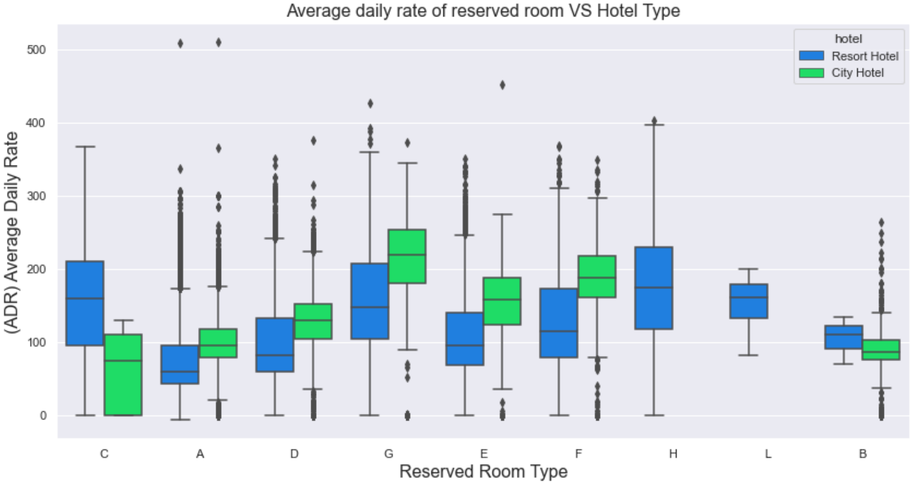
1. DO CUSTOMERS HAVE ANY PREFERENCE WITH OR WITHOUT CHILDREN'S.
2. HOW DAILY AVERAGE RATE IS IMPACTING THE RESERVED ROOM TYPE HOTELS



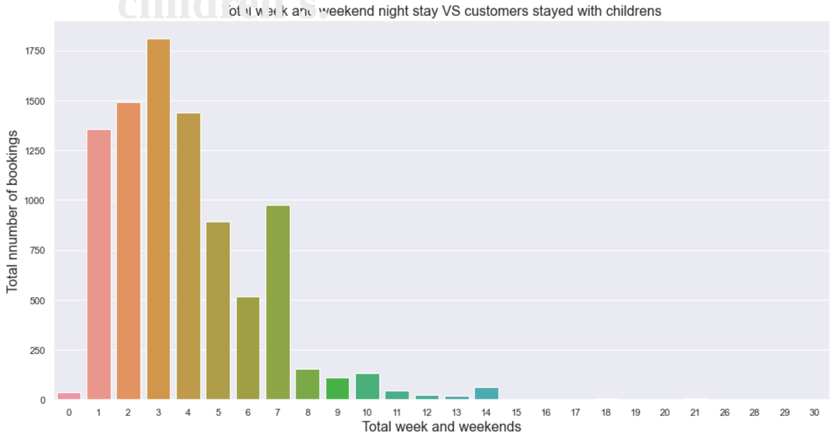
- Customers with children's doesn't have that much preference for the hotel type.

- Room type 'G' has the highest and Room type 'C' has the lowest Average Daily Rates (ADR) in City Hotels.

- Room type 'C' and 'H' has the highest and Room type 'A' has the lowest Average Daily Rates (ADR) in Resort Hotels.

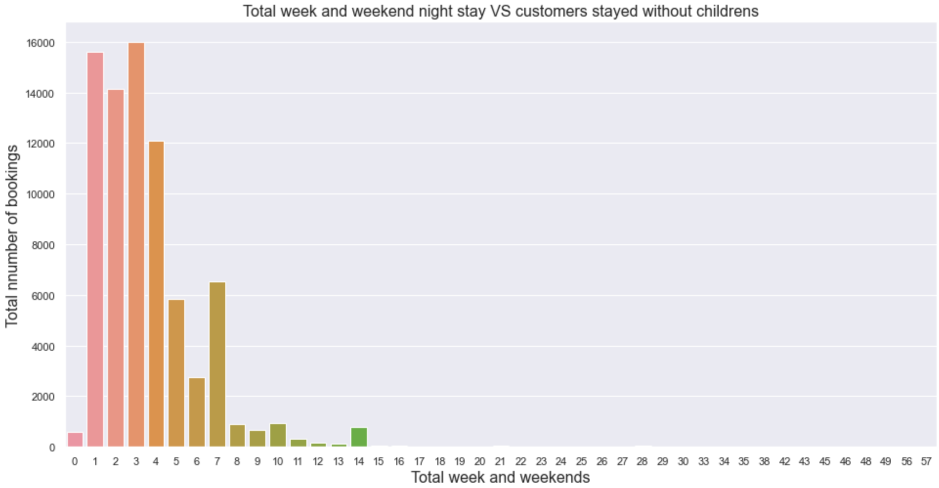


3. Does total stay per month impacted by customers with and without children's

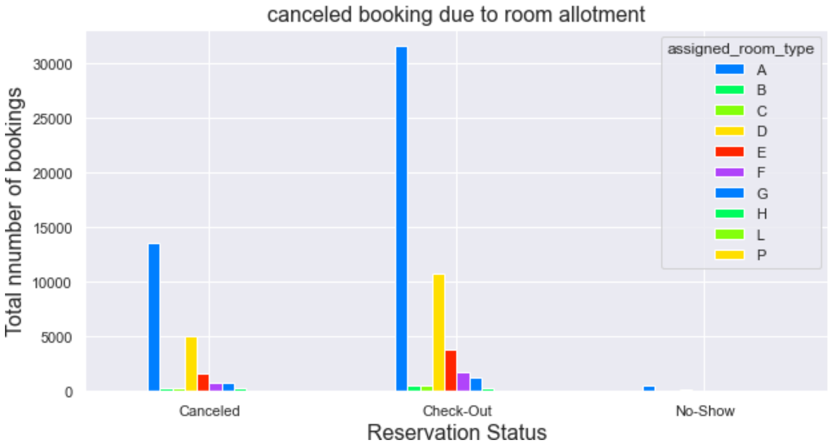


- There is very much similarity in both cases customers stayed with and without children's about 1-7 days have the count for stays with and without children.

- If we go for the average stay then there would be 3-4 days of stay in which people have there children with them otherwise there is only adults staying

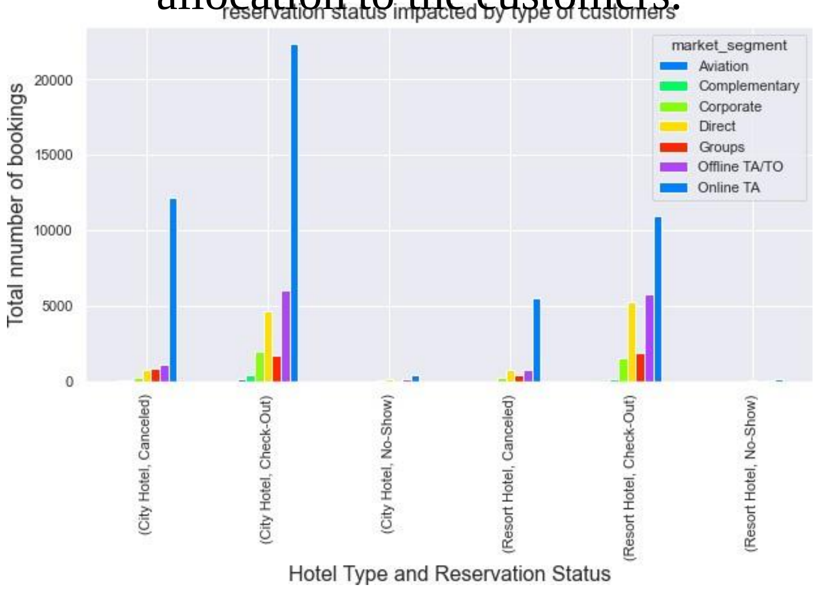


4. Do reservation status impacted by type of customers.
5. Do customers cancel their booking if they are allotted with different room type.



- Customer group from Online TA has the highest number of booking for city hotels.
- Customer group that booked directly is highest at resort hotels.
- But We can Clearly see that most of the bookings are provided by Online

- Hence there is not that much impact of room type allocation to the customers.



CONCLUSIONS AND SUMMARY

- The vast majority of reservations are for hotels in cities. Resort hotels have fewer cancellations than city hotels. The aviation industry has the shortest wait time.

- The months of August, July, and May saw the most hotel bookings in the city. The lowest cancellation rate occurs when a consumer is repeated. The lead time for July is quite long, whereas the lead time for January and February is extremely short. In July, August, and September, the average daily rate for Resort Hotels is higher.

- In June, July, and May, the average daily rate for City Hotels is higher. Customers travelling with children have little preference for the type of hotel they stay in.

Conclusion and Summary(contd..)

- In both cases, the number of stays with and without children is rather equal. Customers who stayed for 1-7 days with or without children had the same count.

- If we take the average stay, three to four days will be spent with youngsters, while the rest of the time will be spent with adults solely.

- Online TA customer group has the highest number of hotel reservations in the city and at resort hotels. As a result, the impact of room type allocation on clients is minimal. The majority of appointments are made by consumers of the Online TA group who book directly through the website.

THANK YOU