Capstone Project 1

Hotel Booking Analysis

Team : Impact players

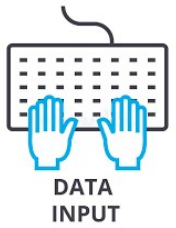
Rahul Kumar
Aparna Khale Anwar
Sujata Jadhav
Mohammad Aasif Malik

Submitted on 25th September 2022

Dataset input and problem Statement

- ✓ The hotel industry is one of the most important components of the wider service industry, catering for customers who require overnight accommodation. It is also closely associated with the travel industry and the hospitality industry,

- ✓ For this project we are analyzing hotel booking data of a city hotel and a resort hotel of few years. The information includes the booking time, check in and check out time, room and meal type, customers stay time, the visitors break up, available parking spaces, visitors' origin, cancellation cases.



- ✓ The key objective of this project to analyze and explore the given data to conclude the meaningful important factors which can help the hotel management to improve both revenue and quality. Also mainly root cause analysis for the cancellation cases can be scrutinize to take necessary preventive actions.

Objective – key questions

1. Hotel type analysis
2. Find peak business season of hotel booking
3. Hotel revenue from ADR (Average daily rate)
4. Type of rooms
5. Meal Consumption Analysis
6. Waiting time Analysis
7. Customer wise analysis
8. Country origin customer analysis
9. Distributor Channel Analysis
10. Agent wise bookings Analysis
11. Company wise bookings Analysis
12. Market Segment -Booking Analysis
13. Hotel booking cancellation on basis of days_in_waiting_list and required_car_parking_spaces
14. Analysis of is_repeated_guest column
15. Number of weekdays booked by distribution channel
16. Number of weekend nights booked by distribution channel

Activity Work flow

✓ **Data Collection and understanding the problem**.

We will be going through each variable and do a logical analysis about their meaning and importance for this problem.

✓ **Data cleaning and manipulation**

We'll clean the dataset and handle the missing data, outliers and categorical variables.
Test assumptions. We'll check if our data meets the assumptions required by most multivariate techniques.

✓ **EDA (Exploratory Data Analysis) and visualization**

Univariate analysis

The data we are analyzing is only one variable

Bivariate analysis

We are comparing two variables to study their relationship

Multivariate analysis

Same as Bivariate analysis but only we are comparing more than two variables

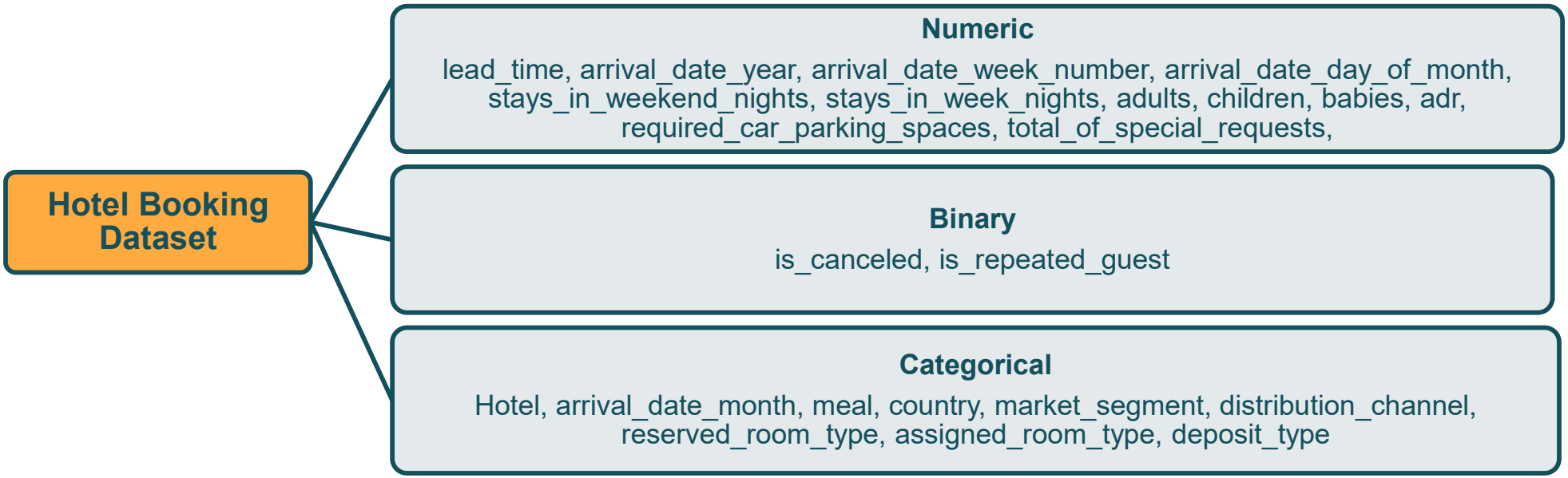
Data Collection and understanding dataset input

Data Input	Description
hotel	City and Resort hotel
is_canceled	indicating booking cancelled (1) or not canelled (0)
lead_time	the time difference between booking date and actual check in
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week no of year for arrival date
arrival_date_day_of_month	day of arrival date
stays_in_weekend_nights	no of weekends night
stays_in_week_nights	no of week nights
adults	no of adults
children	no of children
babies	no of babies
meal	type of meal
country	customers country of origin
market_segment	Market segment type
distribution_channel	booking description channel
is_repeated_guest	if repeated guest (1) or no(0)

Data Collection and understanding dataset input

Data Input	Description
previous_cancellations	no of previous bookings those are cancelled by the customer before the current booking
previous_bookings_not_canceled	no of previous bookings not cancelled by the customer before the current booking
reserved_room_type	Type of reserved room
assigned_room_type	Type of assigned room
booking_changes	no of changes made in the booking from the moment the booking was entered till check in or cancellation
deposit_type	no deposit or refundable or non refundable
agent	ID of travel agent
company	ID of the company that made the booking
days_in_waiting_list	no of days the booking was in waiting list
customer_type	type of customer contract,group
adr	Average daily rate
required_car_parking_spaces	required car parking spaces
total_of_special_requests	no of special request
reservation_status	reservation last status
reservation_status_date	check out date

Dataset Input data summary



Data Collection and understanding dataset input

Prerequisites

- ✓ Import Python libraries.
- ✓ Mount google drive to google colab
- ✓ Authorize notebook to access google drive files

Understanding dataset input

- ✓ Find out the total columns and rows of dataset
- ✓ Find the data type of each column.
- ✓ Find the continuous and categorical data
- ✓ Find individual distribution for some of the columns
- ✓ Also check the correlation between dependent columns

Data cleaning and manipulation

- ✓ Extract the unique values of each column content from the hotel booking dataset.

  Dataset size : 119390 rows × 32 columns

- ✓ Identify duplicated rows and remove the same.

  Dataset size : 87396 rows × 32 columns

- ✓ Calculate percentage values of null values of each column.
- ✓ Combine the null_value and null_value_percentage series in the data frame using ‘concat’ method.
- ✓ Replace NaN values with 0 for heading Agent & company
- ✓ Replace NaN values with their mean values for heading children
- ✓ Replace NaN values with 'others' for heading Country
- ✓ Modify datatype from float to int64 for heading Agent, Company, Children

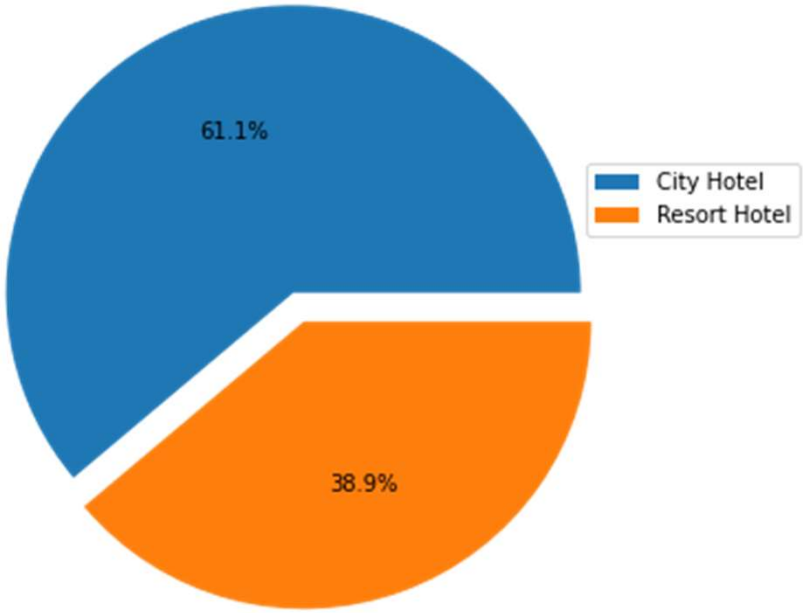
EDA (Exploratory Data Analysis)

1.Hotel type analysis

Q1.Find type of hotel
Output : array(['Resort Hotel', 'City Hotel'], dtype=object)

Q2.Find count of booking as per type of hotel
Output:

Sr	Hotel type	Number of booking
0	City Hotel	53428
1	Resort Hotel	33968

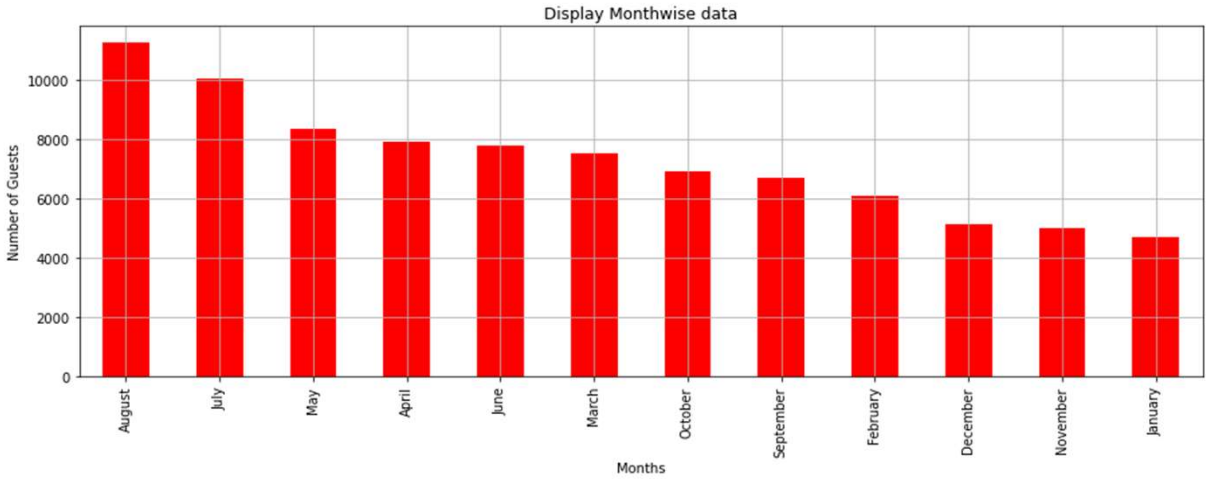


Conclusion :
Booking of city hotel[61.1%] is more than compared to resort hotel[38.9%]

EDA (Exploratory Data Analysis)

2. Find peak business season of hotel booking

Q1. Display the number of cumulative booking months wise



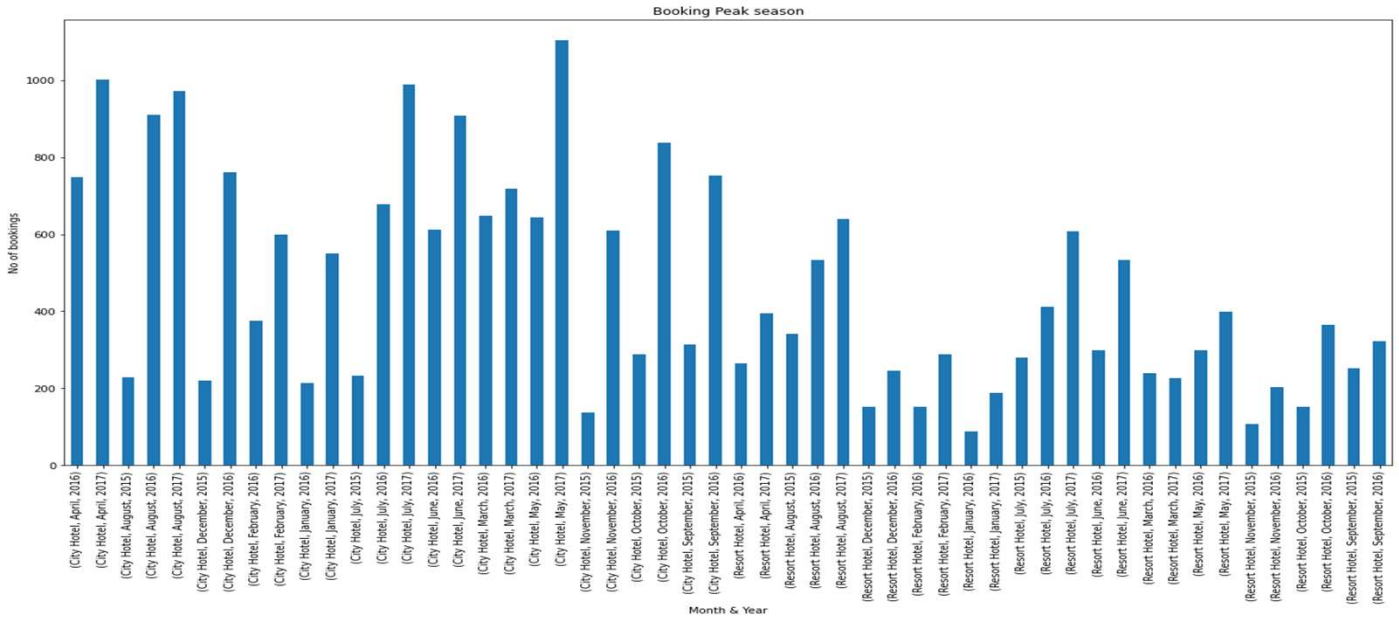
Conclusion:

As per bar plot shows, May: August month is a peak season for the hotel business whereas November and December is slack seasons.

EDA (Exploratory Data Analysis)

2. Find peak business season of hotel booking

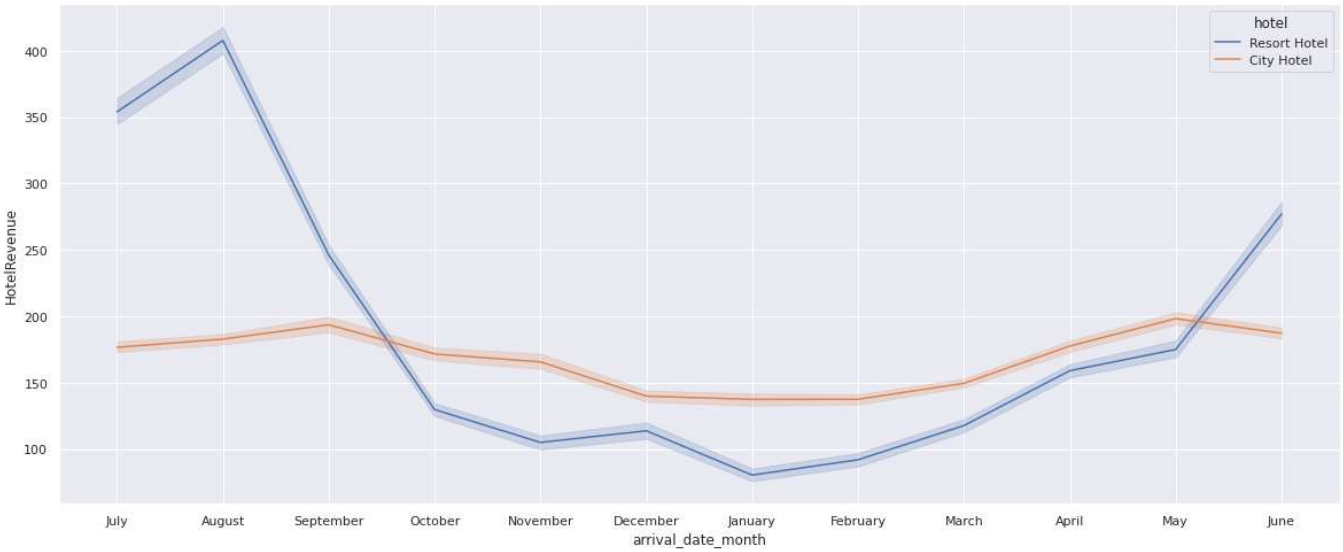
Q2. Display no. of booking as per month and year.



EDA (Exploratory Data Analysis)

3. ADR (Average daily rate)

- Q1. Find total no. of guests on basis of adults,children and babies.
- Q2. Find ADR per person.
- Q3. Find total no. of stay on basis of week nights and weekend nights.
- Q4. Find hotel revenue=ADR per person * no. of days



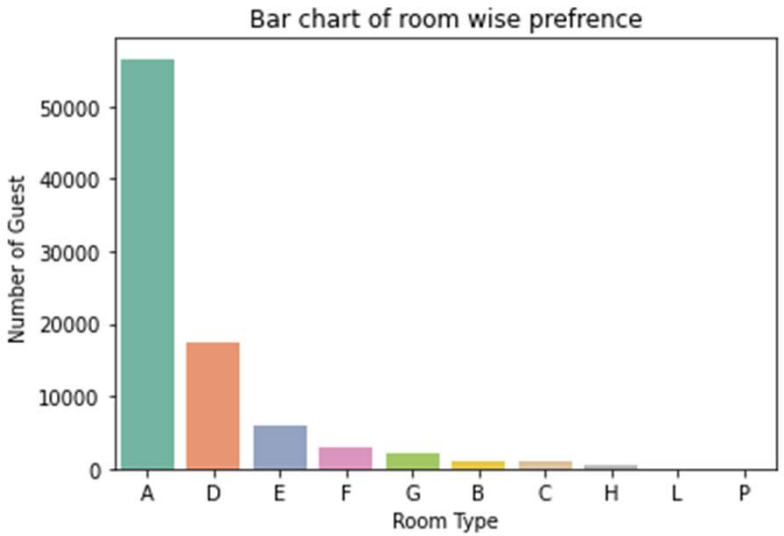
Conclusion : Resort hotel is getting more revenue in the month of August.

EDA (Exploratory Data Analysis)

4. Type of rooms

Q1.How many types of rooms are provided through Hotels?
Output : array(['C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B']

Q2.Which is the most preferred room by the guest?



Conclusion :" A" type room is more preferred by the guests.

EDA (Exploratory Data Analysis)

5. Meal Consumption Analysis

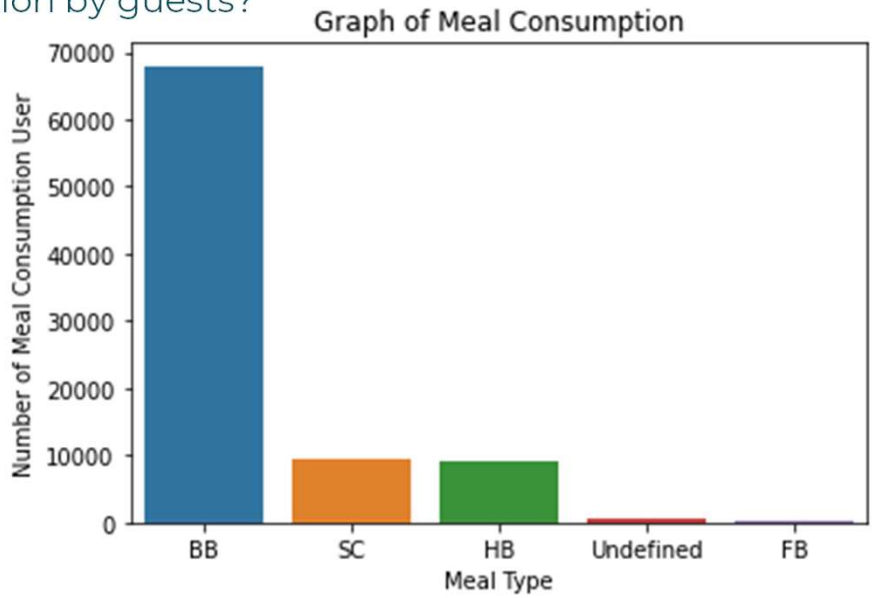
Q1. Which type of meal offer by hotel?

Output : array(['BB', 'FB', 'HB', 'SC', 'Undefined'], dtype=object)

Q2. Which is the most preferred meal consumption by guests?

Output :

	Meal_type	number_of_preference
0	BB	67978
1	SC	9481
2	HB	9085
3	Undefined	492
4	FB	360



Conclusion :"BB" meal type is mostly preferred by the guests.

EDA (Exploratory Data Analysis)

6. Waiting time Analysis

Q1. Analyze in which type of hotel there is more waiting time

Output:

Sr.	Hotel	Max_Waiting_Time
0	City Hotel	391
1	Resort Hotel	185

Conclusion :City Hotel having overall more waiting time which interprets that it is more crowded than Resort.

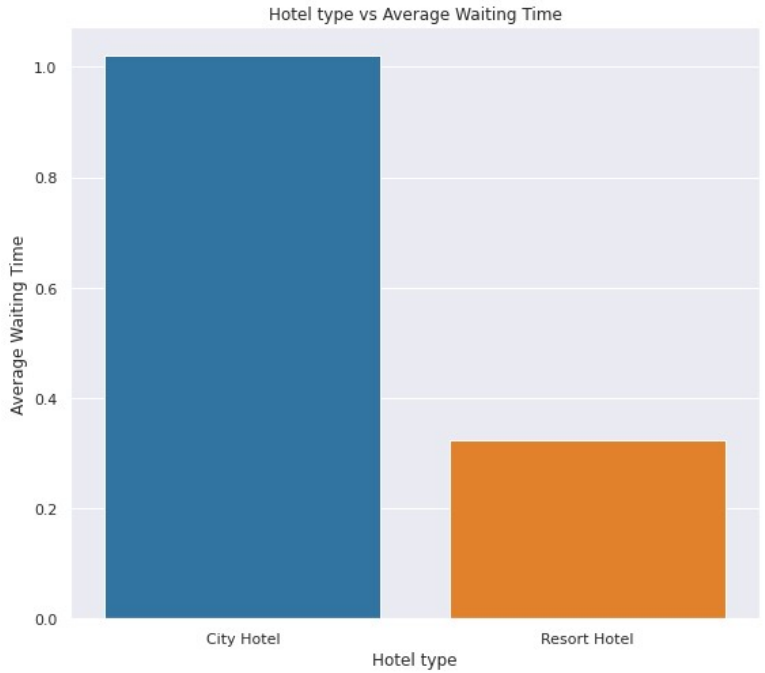
EDA (Exploratory Data Analysis)

6. Waiting time Analysis

Q2. Analyze the average of type of hotel there is more waiting time

Output:

Sr	Hotel	Avg_Waiting_Time
0	City Hotel	1.020233
1	Resort Hotel	0.323834



Conclusion :As city hotels are preferred most by guests , it's having more waiting period.

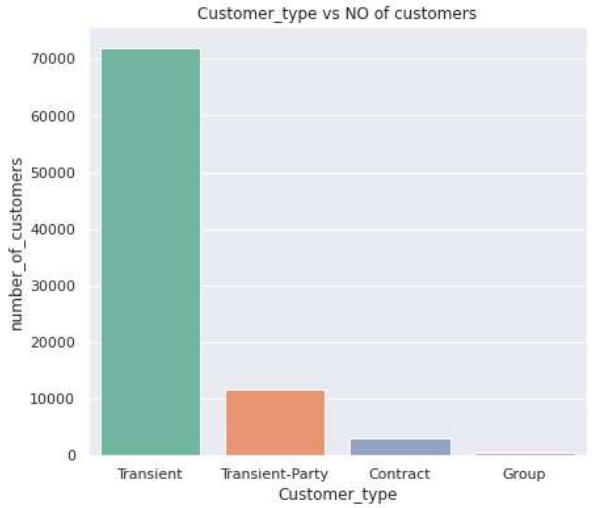
EDA (Exploratory Data Analysis)

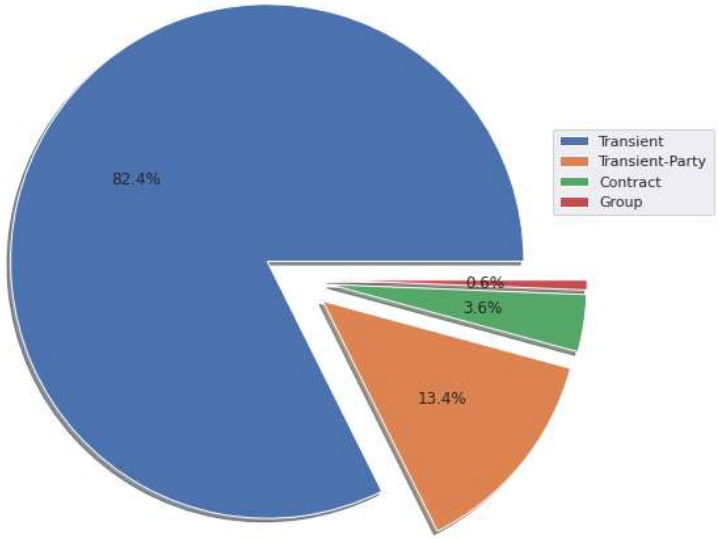
7. Customer wise analysis

Q1. Find out count of customers on basis of customer type.

Output:

Sr	Customer type	number_of_customers
0	Transient	71986
1	Transient-Party	11727
2	Contract	3139
3	Group	544





Conclusion :The maximum number of guest are from transient category which is near about 75.1%.

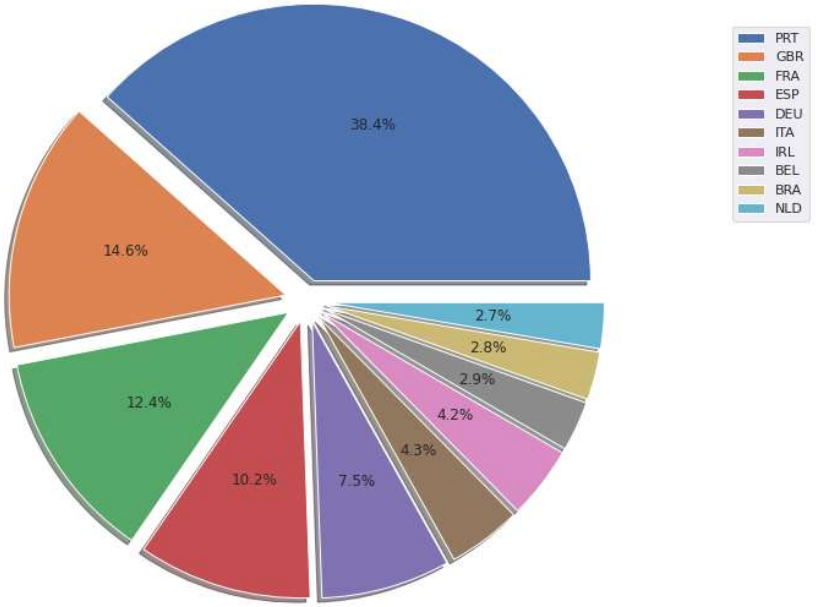
EDA (Exploratory Data Analysis)

8. Country origin customer analysis

Q1. Analyze the data from which country guests are visiting most.

Output:

Sr	country_name	number_of_guests
0	PRT	27453
1	GBR	10433
2	FRA	8837
3	ESP	7252
4	DEU	5387
5	ITA	3066
6	IRL	3016
7	BEL	2081
8	BRA	1995
9	NLD	1911



Conclusion:

Below top 5 countries from which most guests are visiting

1.PORTUGAL(PRT)-->38.4%
2.GREAT BRITAIN(GBR)-->14.6%
3.FRANCE(FRA)---->12.4%
4.SPAIN (ESP)--->10.2%
5.GERMANY (DEU)---->7.5%

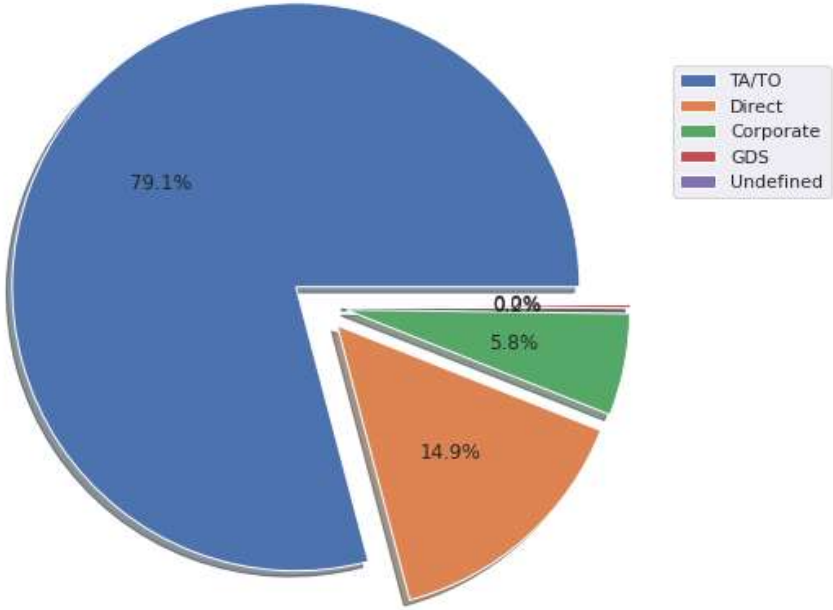
EDA (Exploratory Data Analysis)

9. Distributor Channel Analysis

Q1. Find out which distribution channel is giving the most booking business.

Output:

Sr	Distribution Type	NumberofBooking
0	TA/TO	69141
1	Direct	12988
2	Corporate	5081
3	GDS	181
4	Undefined	5



Conclusion : Distributor Channel Analysis TA/TO is giving the most booking business

EDA (Exploratory Data Analysis)

9. Distributor Channel Analysis

Q2.Which distribution channel is giving more business to respective hotels?

Output:

- 1.TA/TO -->69141 i.e. 79.1%
- 2.Direct -->12988 i.e. 14.9%
- 3.corporate -->5081 i.e. 5.8%
- 4.GDS -->181 i.e. 0.2%
- 5.undefined -->5 i.e. close to 0.001%

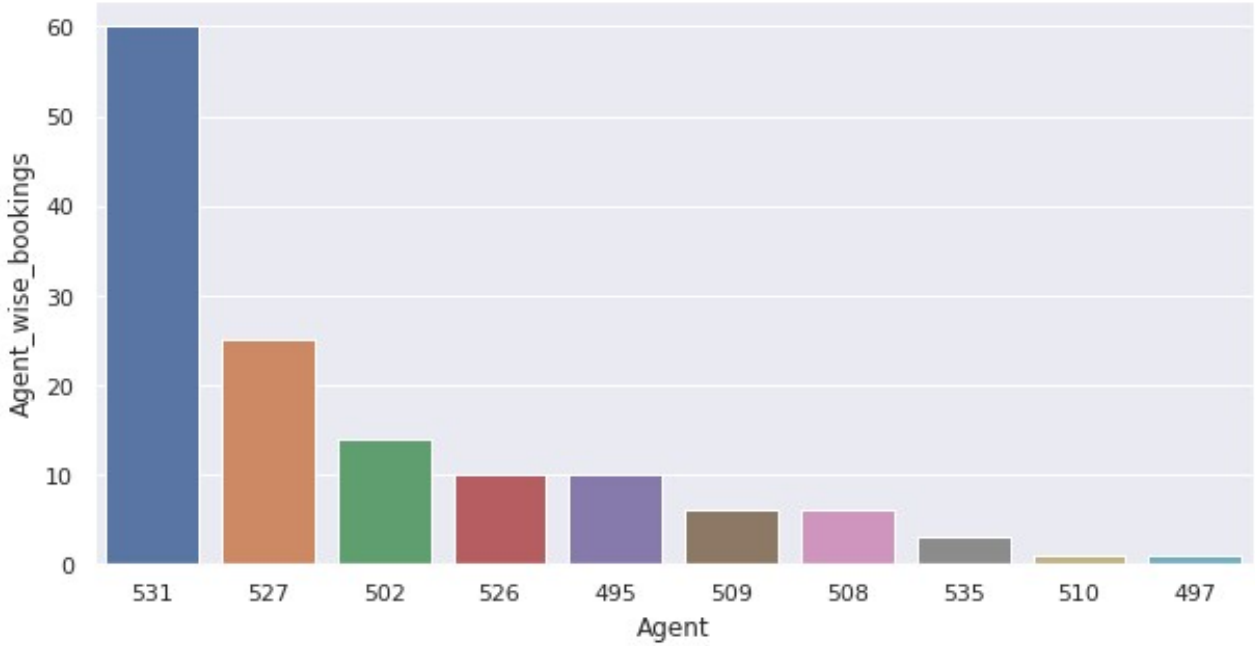
Conclusion : Distributor channel TA/TO is giving the most booking business

EDA (Exploratory Data Analysis)

10. Agent wise bookings Analysis

Q1.Which agent giving more business to respective hotels?

Output:



Conclusion :Agent ID 531 is giving maximum hotel bookings so this data can be utilized to decide commission % for the agent

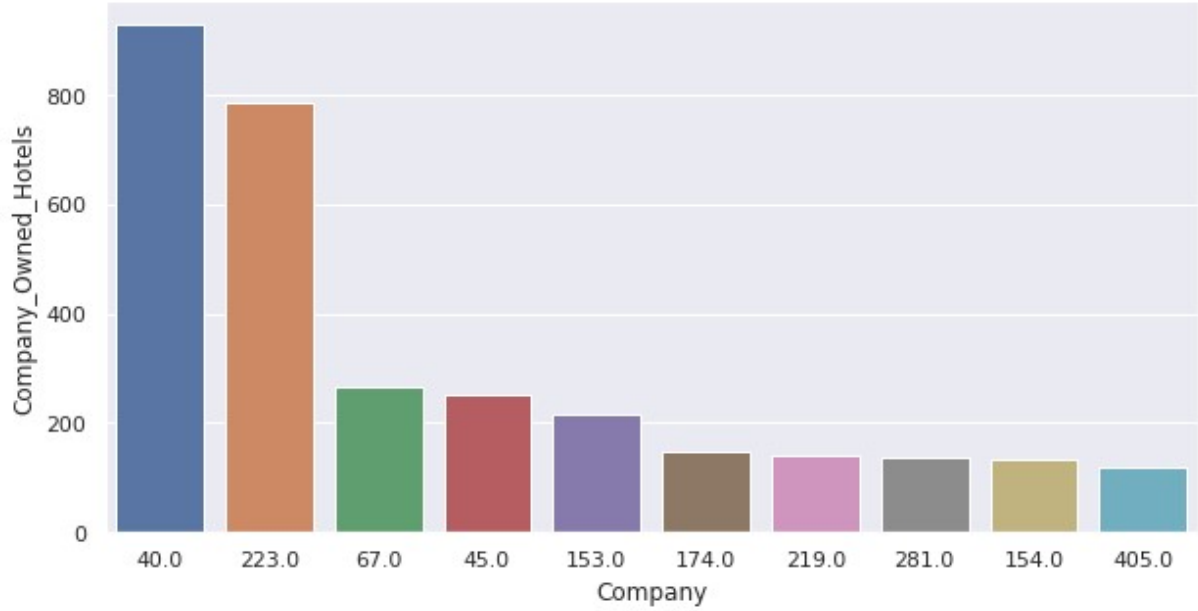
EDA (Exploratory Data Analysis)

11. Company wise bookings Analysis

Q1.Which company owning how many hotels?

Output:

Sr	Company	Company_Owned_Hotels
0	40.0	927
1	223.0	784
2	67.0	267
3	45.0	250
4	153.0	215
5	174.0	149
6	219.0	141
7	281.0	138
8	154.0	133
9	405.0	119



Conclusion : Above graph shows company wise owning maximum no of hotels

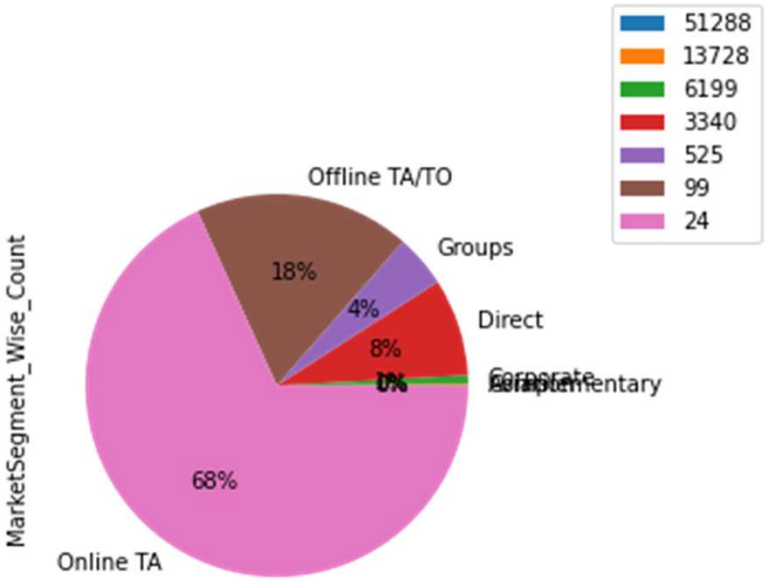
EDA (Exploratory Data Analysis)

12. Market Segment -Booking Analysis

Q1.Which market segment is giving more business?

Output:

Sr	MarketSegment	MarketSegment_Wise_Count
0	Online TA	51288
1	Offline TA/TO	13728
2	Direct	6199
3	Groups	3340
4	Corporate	525
5	Complementary	99
6	Aviation	24



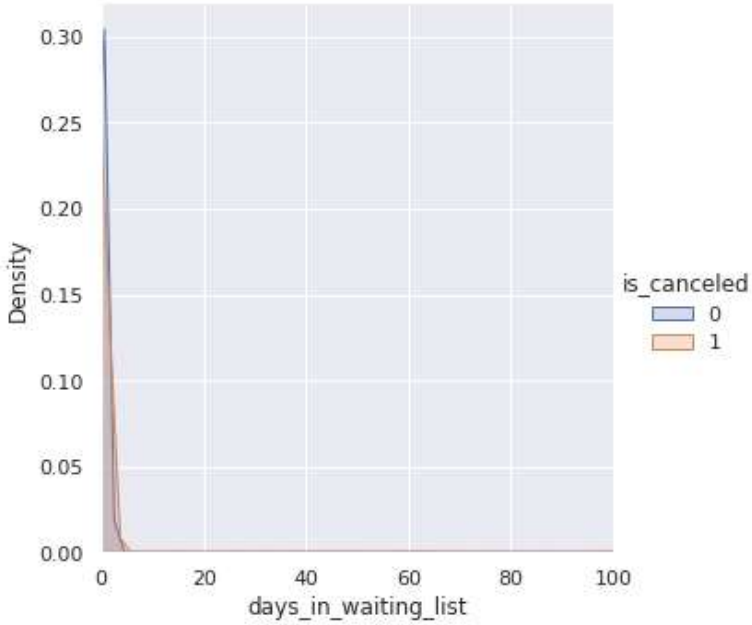
Conclusion : Online TA is most commonly used market segment for the booking purpose.

EDA (Exploratory Data Analysis)

13. Hotel booking cancellation on basis of days_in_waiting_list and required_car_parking_spaces

Q1. Waiting List Vs Cancellations

Output:



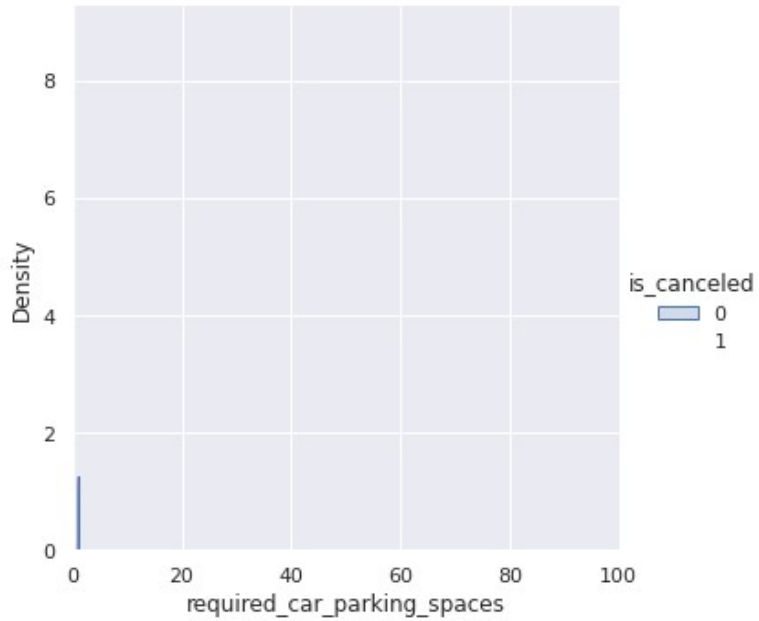
Conclusion : There are high chances of cancellation when the waiting period is high.

EDA (Exploratory Data Analysis)

13. Hotel booking cancellation on basis of days_in_waiting_list and required_car_parking_spaces

Q2. Car Parking Vs Cancellation

Output:



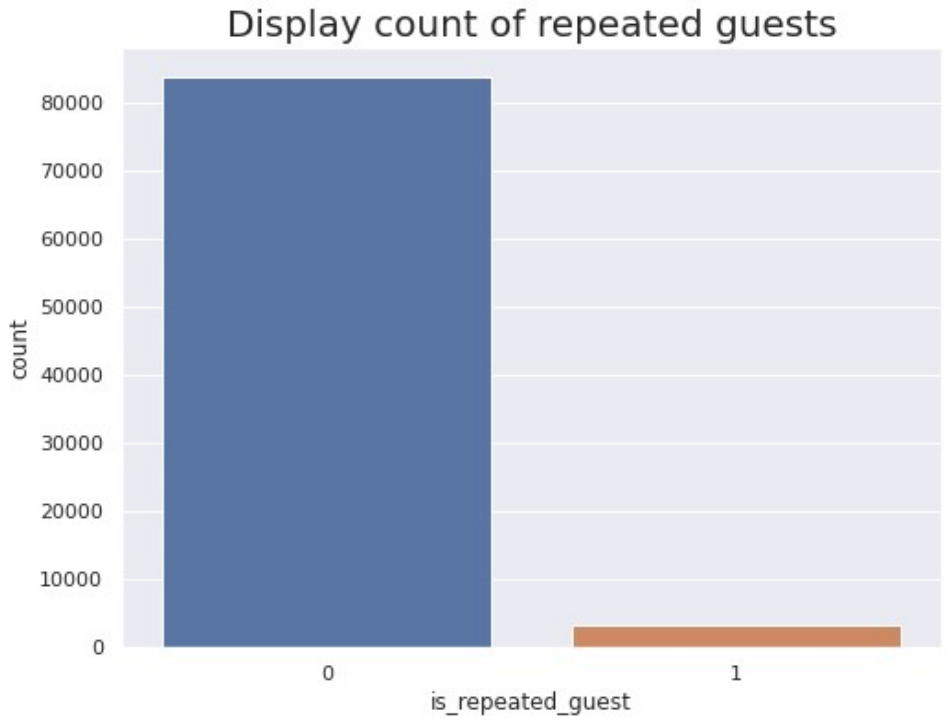
Conclusion : There is no impact on cancellation due to car parking space

EDA (Exploratory Data Analysis)

14. Analysis of is_repeated_guest column

Q1. Volume flow of repeated guest

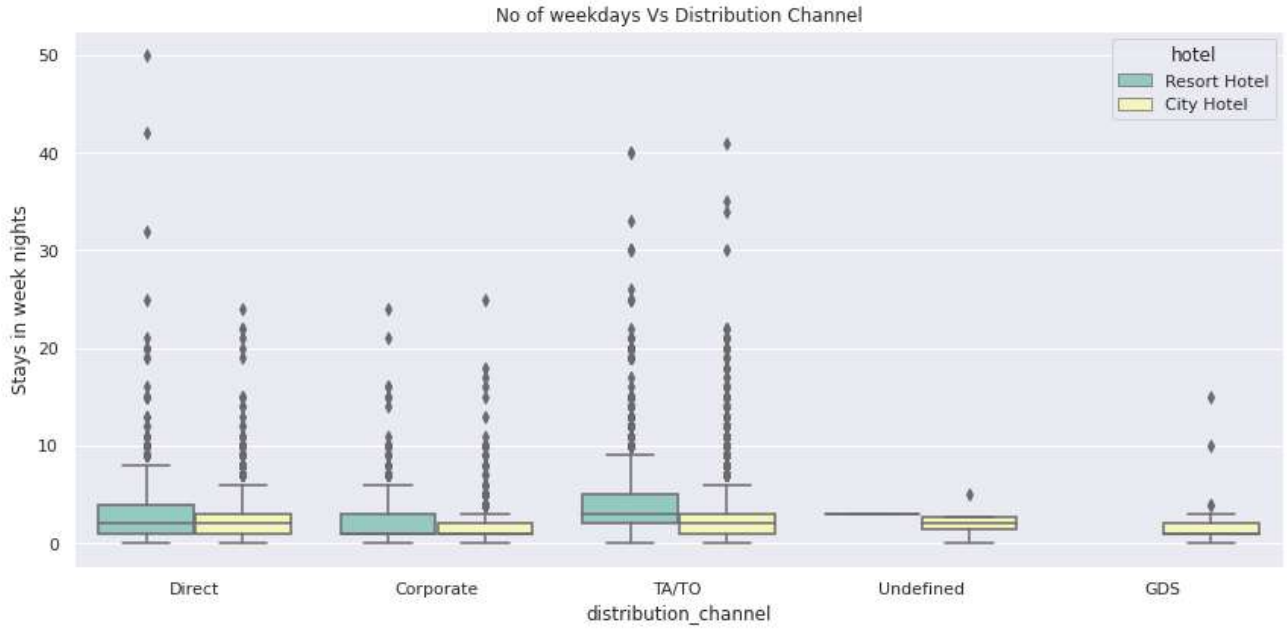
Output:



Conclusion : Above data shows that most of the guests are visiting first time.

EDA (Exploratory Data Analysis)

15. Number of weekdays booked by distribution channel



Conclusions :

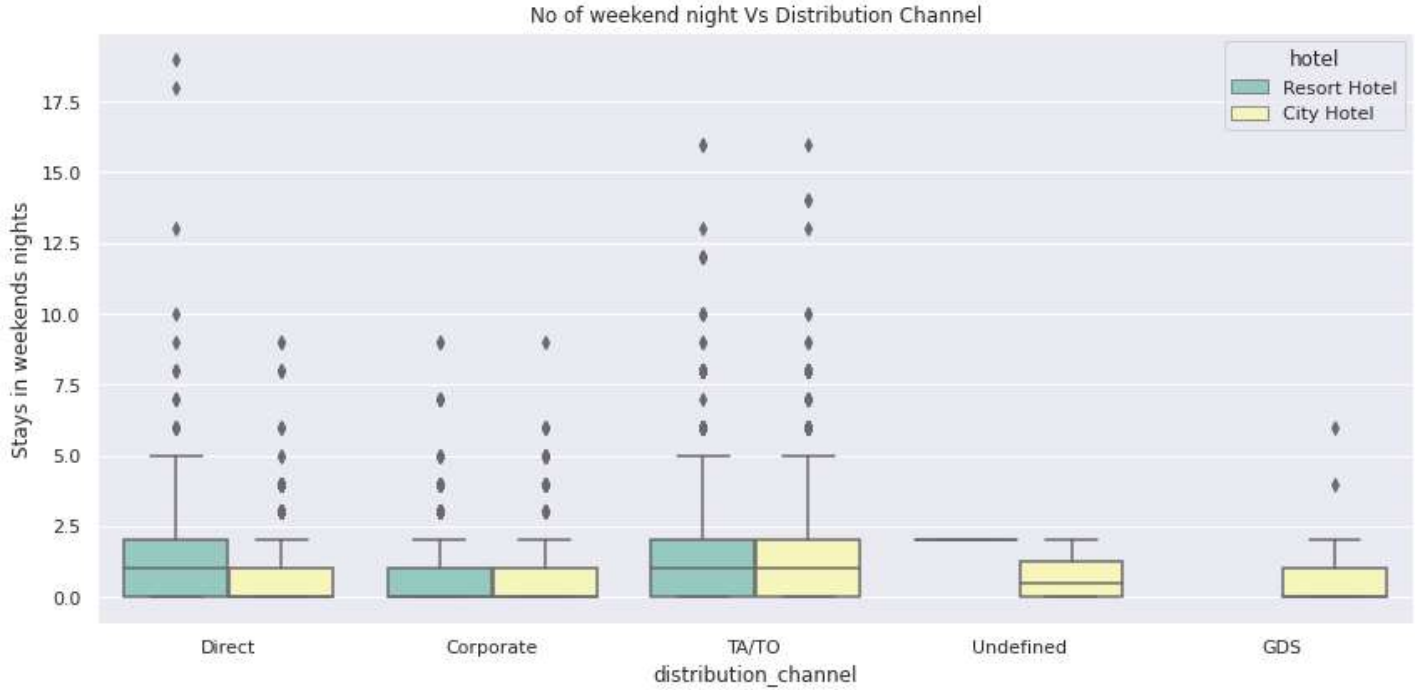
1.Visitors from direct and corporate distribution channel are staying almost in same range numbers of week nights.

2.TA/TO distribution channel has some deviation over stays week-nights between Resort and City hotels.

3.Undefined and GDS distribution channel visitors had not shown interest in the Resort Hotel.

EDA (Exploratory Data Analysis)

16. Number of weekend nights booked by distribution channel



Conclusion :

1.Direct distribution channel visitors prefer to stay more weekend nights in the Resort Hotel type.

2.Visitors through Corporate and TA/TO distribution channel are equally preferring between Resort and City hotels.

EDA (Exploratory Data Analysis)

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

- Bar Plot
- Pie Chart
- Line Plot
- Box Plot

Conclusions

1. 1.City hotel[61.1%] having more booking as compared to resort hotel[38.9%]."
2. 2.Overall from May: August month is a peak season for the hotel business whereas November and December is slack seasons.
3. 3.Resort hotel is getting more revenue in the month of August."
4. 4.City Hotel having overall more waiting time which interpret it is more crowded than Resort
5. 5. Here we can see the maximum number of customers are from transient category which is near about 75.1%.
6. 6.Distributor channel TA/TO is giving the most booking business
7. 7.Agent ID 531 is giving maximum hotel bookings so this data can be utilized to decide commission % for agent
8. 8.There are high chances of cancellation when waiting period is high.

Challenges

- Dataset contains a lot of duplications.
- Against few columns having a lot of Null values.
- Few dataset columns with wrong datatype format.

Thank You