Capstone Project

Project Title

HOTEL BOOKING ANALYSIS

by

Sudhanshu Chouhan
Kapil Narayan Singh
Nimisha Nooti

PROBLEM STATEMENT

- Hotel Industry is the backbone of tourism sector. Revenue of the hotel industry depend upon multiple factors.
- Some of the factors influence the hotel business are Location, Quality Management, Flexibility, Global Outlook etc.
- In order to understand how guest select hotels to stay in and what are decision making factors that prevail, this analysis will help us to target the important area and generating more revenue for the hotel industry.

Road Map of Data Analysis

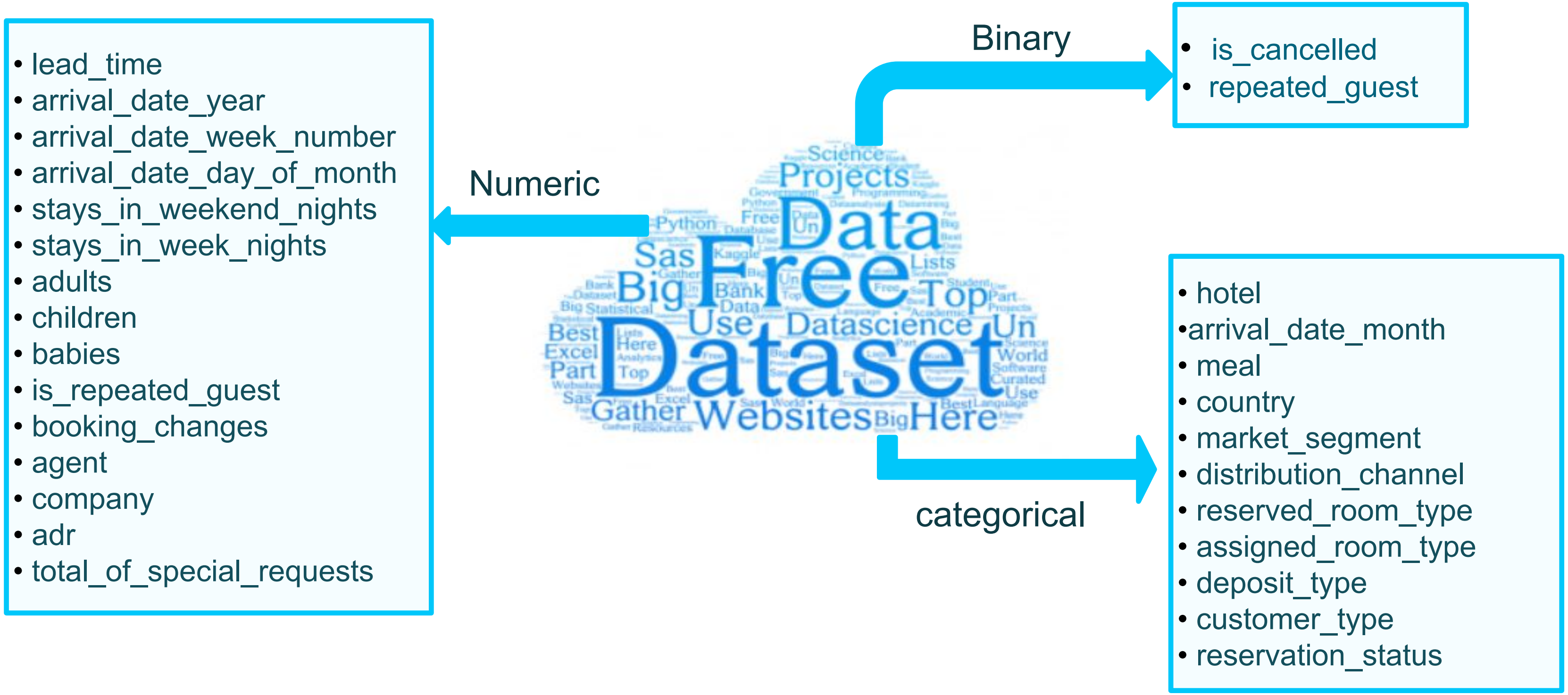
- **Basic cleaning：** Separating the required data from the data set to use it for analyzing.
- **Understanding and analyzing the factors effecting the booking：** Factors like seasons, festivals and holidays effect the data.
- **Data processing :** We’ll go through each and every feature and encoded the categorical features. Changed the columns according to requirement of analysis.
- **visualizing the test assumptions：** We'll check if our data meets the assumptions required by most multivariate techniques, and represent them in the understandable-form of visualization using bar, pie, line, box etc plots

Data Description

There are 119390 rows and 32 columns in the given data set.
From the variables we will evaluate the dependency of hotel booking as follows:

- Hotel wise analysis
- Booking cancellation analysis
- Bookings based on Meals, Countries, Room Type
- Distribution channel wise analysis
- Lead time analysis
- Revenue analysis
- Bookings over time analysis

Data Summary



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
- is_repeated_guest
- booking_changes
- agent
- company
- adr
- total_of_special_requests

Binary

- is_cancelled
- repeated_guest

categorical

- hotel
- arrival_date_month
- meal
- country
- market_segment
- distribution_channel
- reserved_room_type
- assigned_room_type
- deposit_type
- customer_type
- reservation_status

Let's first describe the Variables.

1. Hotel : Two category of hotels are Resort Hotel or City Hotel.
2. is_canceled : Value indicating if the booking was canceled ()1 or not(0)
3. lead_time : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. arrival_date_year : Year of booking arrival date .
5. arrival_date_month : Month of booking arrival date.
6. arrival_date_week_number : Week number of the booking arrival date.
7. arrival_date_day_of_month : Day of booking arrival date.
8. stays_in_weekend_nights : Number of weekend nights(Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. stays_in_week_nights : Number of week nights(Monday or Friday) the guest stayed or booked to stay at the hotel.
10. adults : Number of adults reserved the hotel stay.

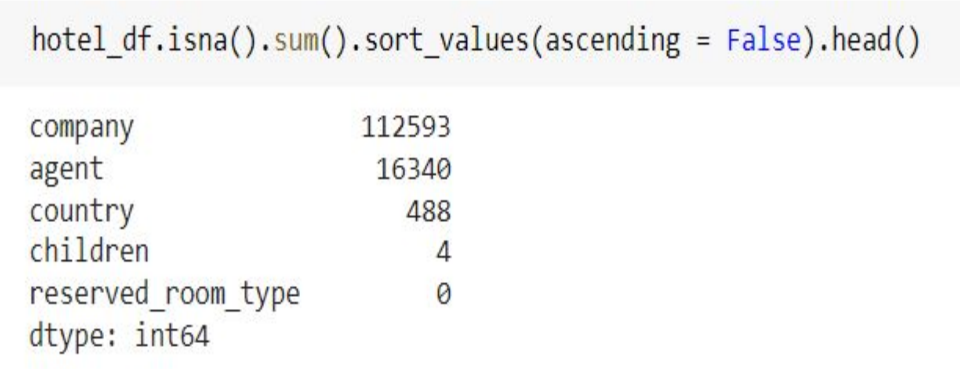
Describing Variables continued….

11. children : Number of children.
12. babies : Number Of babies.
13. meal : kind of meal opted for.
14. country : Country code.
15. market_segment : Market segment designation .
16. distribution_channel : Booking distribution channel.
17. is_repeated_guest : Is a repeated guest is binary info (1) yes or (0) no.
18. previous_cancellations : Number of previous cancellation not cancelled by current booking.
19. previous_bookings_not_canceled : Number of previous booking not cancelled by current booking.
20. reserved_room_type : Code of room type reserved.
21. assigned_room_type : Code for the type of room assigned.
22. deposit_type : No deposit, Non Refund, Refundable

Describing Variables continued….

22. booking_changes : Number of changes made to the booking from the moment PMS until the moment of check in/cancellation.
24. Agent : ID of the travel agency that made the booking.
25. Company : ID of the company/entity that made the booking.
26. days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer.
27. customer_type : Type of customer. contact, group, transient, transient party.
28. adr : Average daily rate as defined by dividing the sum of all lodging transaction by the total number of staying nights.
29. required_car_parking_spaces : Is parking required.
30. total_of_special_requests : Number of additional special requirement.
31. reservation status : status of reservation
32. reservation_status_date : Date of the specific status.

Data Cleaning and Manipulation



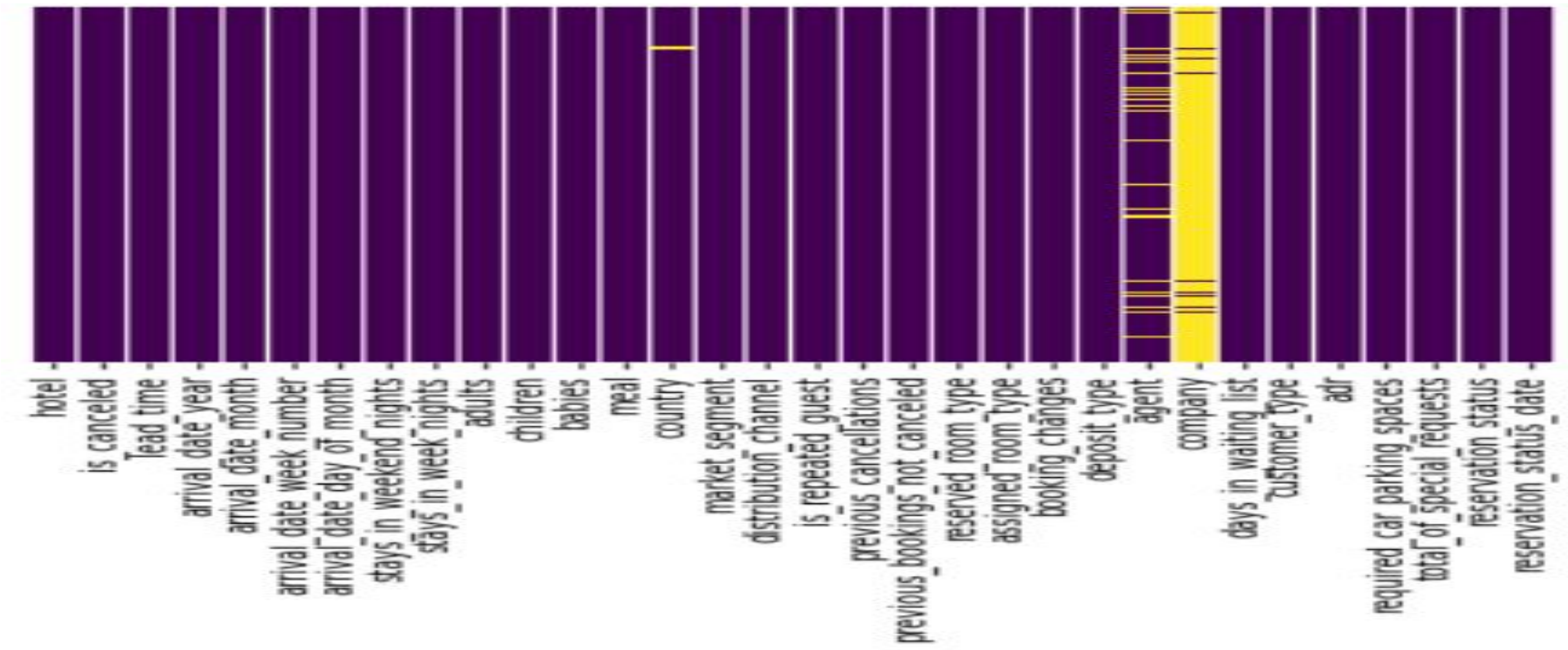
```
hotel_df.isna().sum().sort_values(ascending = False).head()
```

```
company              112593
agent                 16340
country                 488
children                  4
reserved_room_type        0
dtype: int64
```

Finding the null values in column:

To Identify the null values in the columns of data set we are using ‘isna()’ method. We can see that the columns – [company , agent, country and children] have null values in decreasing order respectively rest of the data don't have any missing value.

Heat map showing null values:

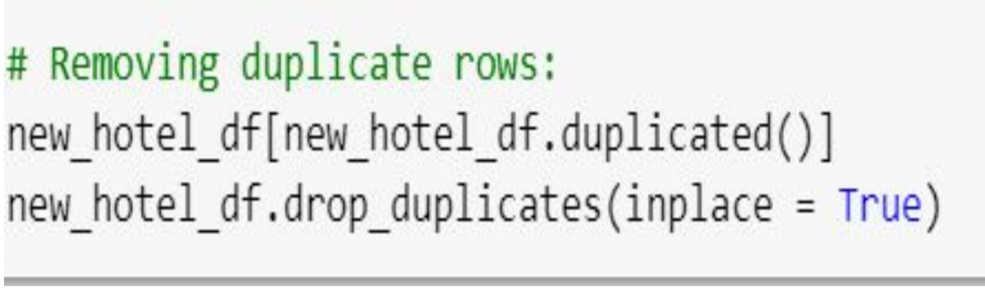


This heat map shows the column company has highest null values. The color difference (yellow color) shows the null values in that column.

❑ Removing the duplicate values in the data set:

Now we will find out duplicate values in the dataset using ‘drop()’ method

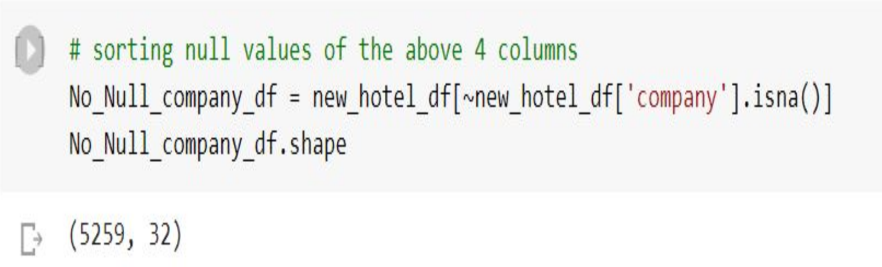
There are 31994 duplicate values found in the data set.



```
# Removing duplicate rows:
new_hotel_df[new_hotel_df.duplicated()]
new_hotel_df.drop_duplicates(inplace = True)
```

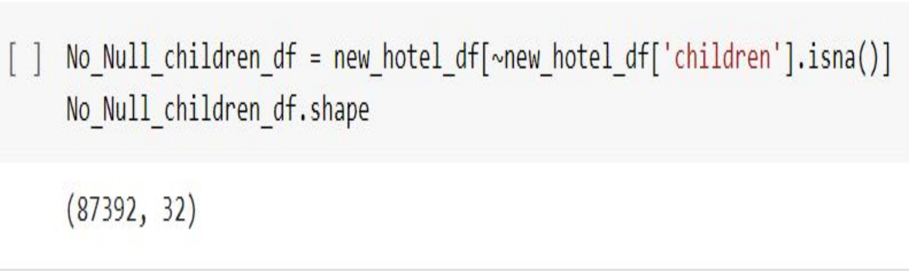
❑ Removing the null values in the columns of data set:

When ever we are working with above 4 columns of data set we will use 'no null df' to eliminate the null values.



```
# sorting null values of the above 4 columns
No_Null_company_df = new_hotel_df[~new_hotel_df['company'].isna()]
No_Null_company_df.shape
```

```
(5259, 32)
```

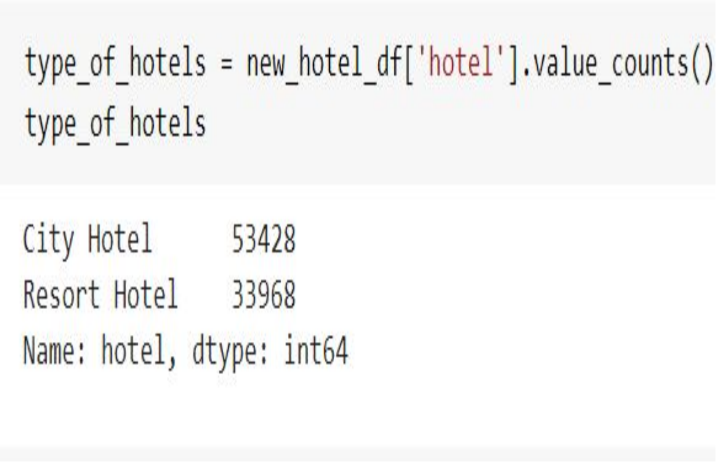


```
[ ]  No_Null_children_df = new_hotel_df[~new_hotel_df['children'].isna()]
     No_Null_children_df.shape
```

```
(87392, 32)
```

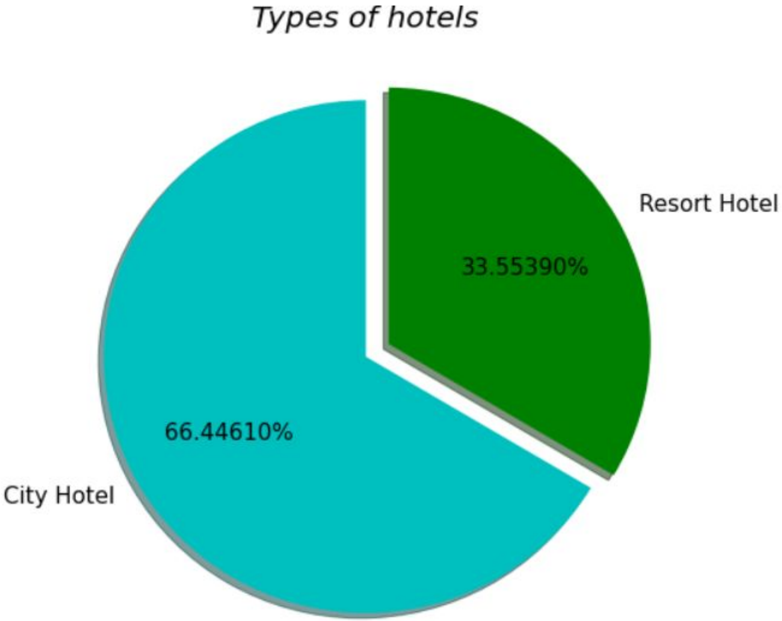
Data Visualization

After data cleaning and manipulation we will visualize our data set through various visualization techniques and discuss about various conclusion we will get from mentioned techniques.



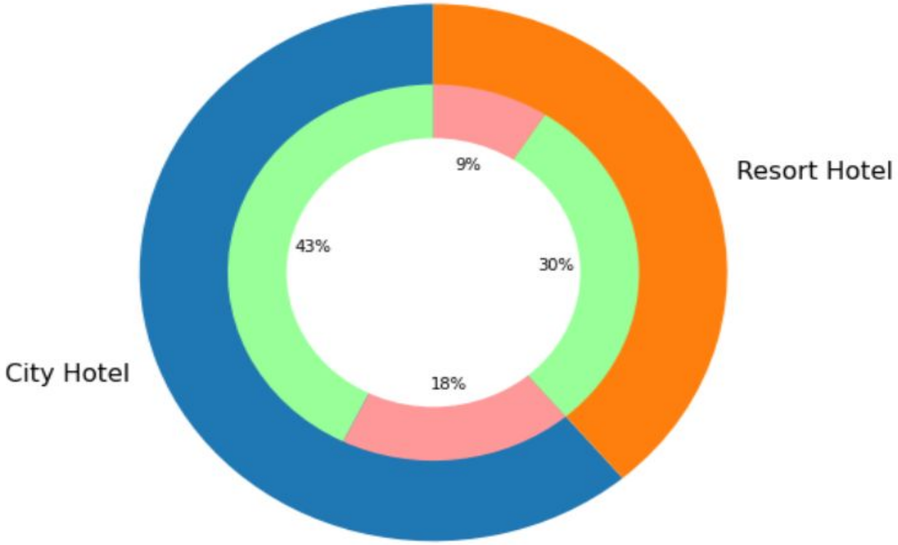
```
type_of_hotels = new_hotel_df['hotel'].value_counts()
type_of_hotels
```

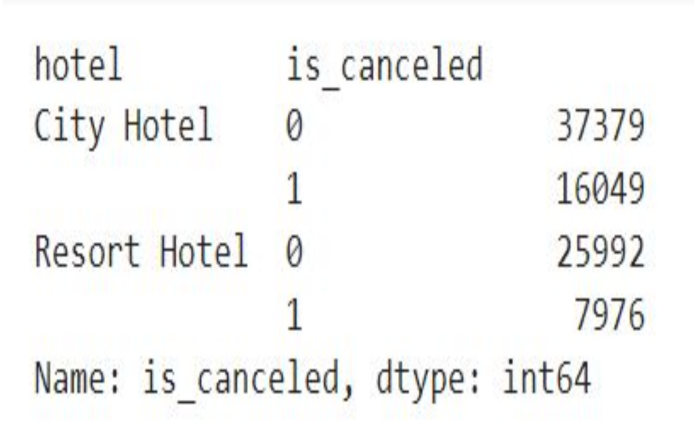
```
City Hotel      53428
Resort Hotel    33968
Name: hotel, dtype: int64
```



There are 66.4% of City Hotels and 33.5% of Resort Hotels were booked. Therefore City Hotels are more preferred by guests compared to Resort hotels.

Percentage of bookings that are canceled at diffrent type of hotels





```
hotel         is_canceled
City Hotel    0              37379
              1              16049
Resort Hotel  0              25992
              1               7976
Name: is_canceled, dtype: int64
```

we can observe from the above pie visualization that the max bookings and cancellations are happening in city hotel.

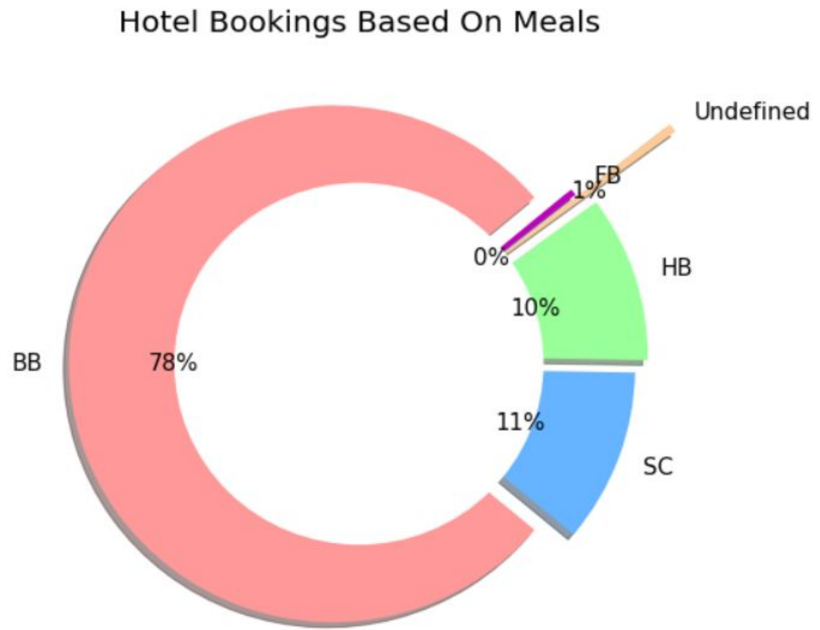
total bookings that are cancelled = (18%+9%) = 27% (66.6% of cancellation happening in city hotel)

total bookings that are not cancelled = (43%+30%) = 73%

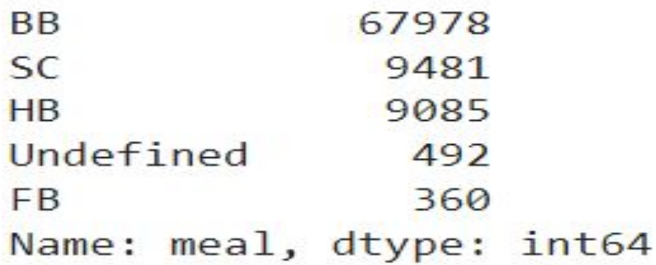
- green = bookings not cancelled
- pink = bookings cancelled

Hotel Bookings based on Meals

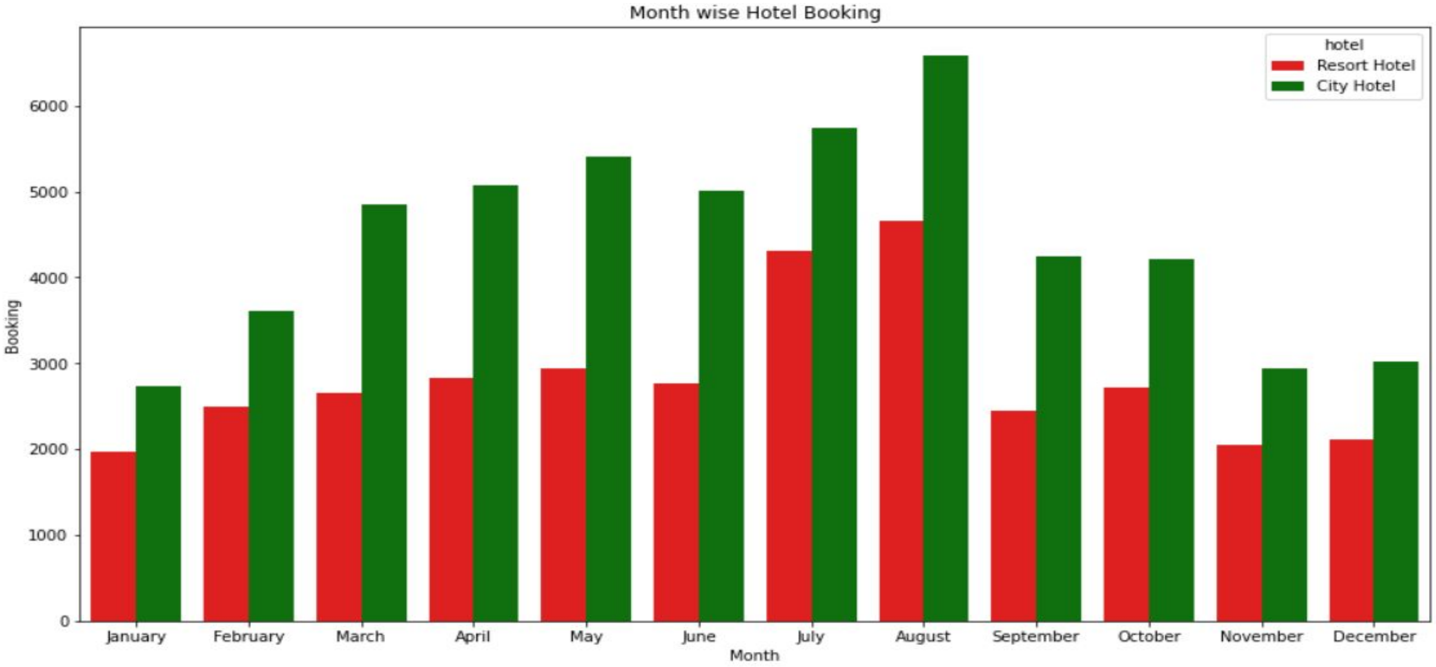
- RO: Room only
- BB: Bed & Breakfast
- HB: Half Board (Breakfast and Dinner normally)
- FB: Full Board (Breakfast, Lunch and Dinner)
- AI: All Inclusive (all services of full board plus any others specified in each case)



From the above pie visualization we can conclude that 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast‘.

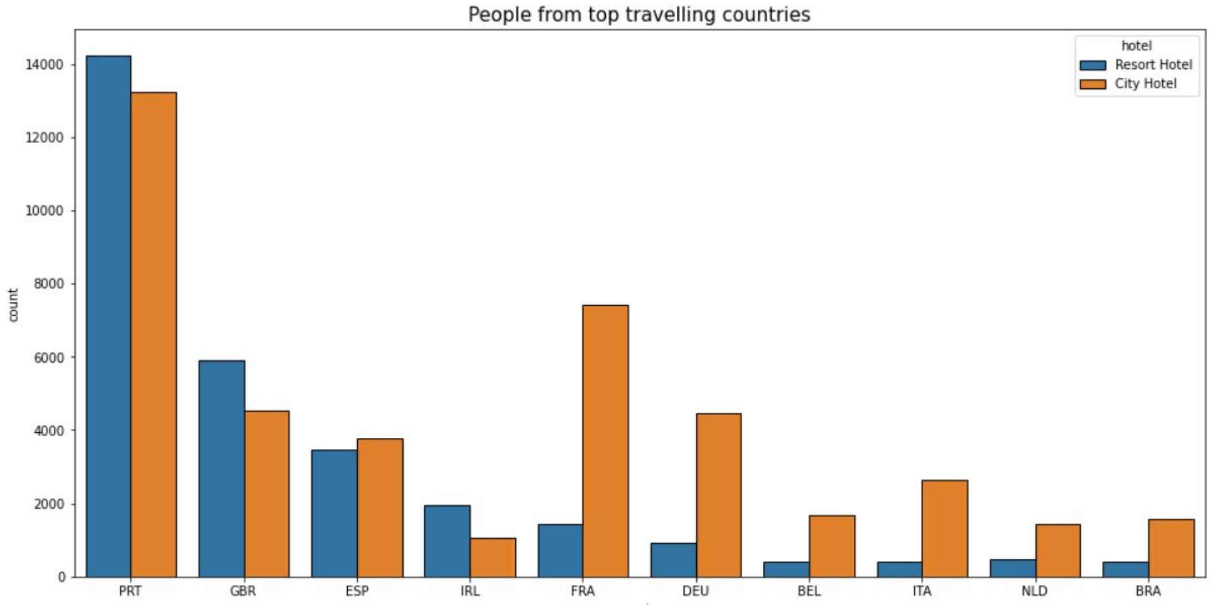


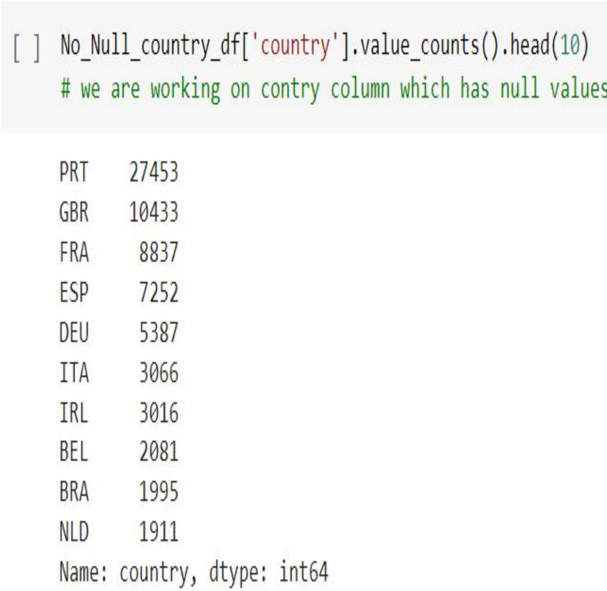
```
BB           67978
SC            9481
HB            9085
Undefined      492
FB             360
Name: meal, dtype: int64
```



Most of the city and resort bookings are happening in the month of **August** followed by July. Least bookings are happening in the month of January, November and December.

Booking analysis based on countries



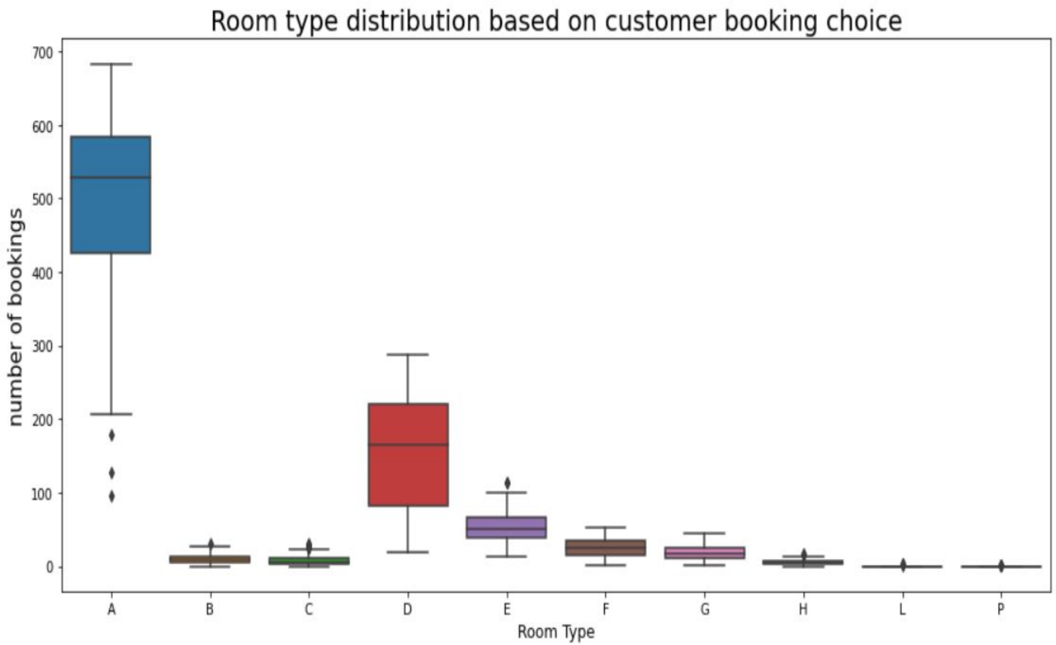


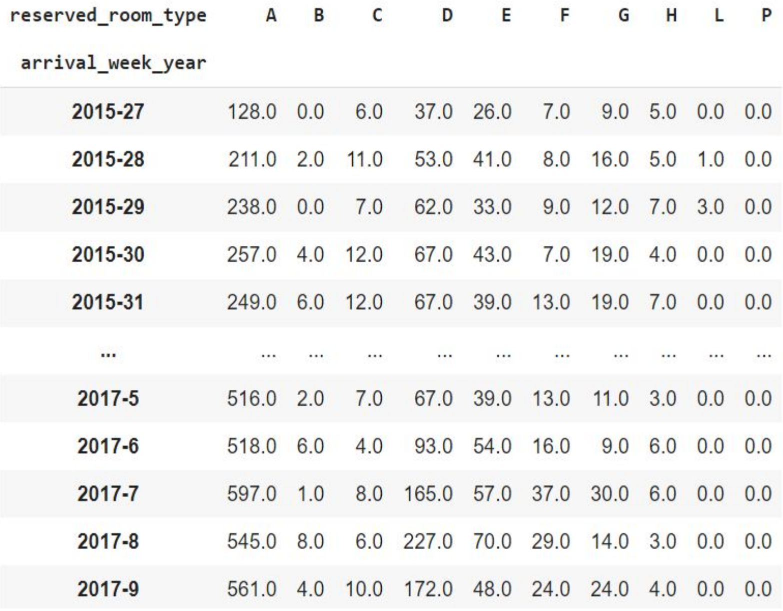
```
[ ]  No_Null_country_df['country'].value_counts().head(10)
     # we are working on contry column which has null values
```

```
PRT    27453
GBR    10433
FRA     8837
ESP     7252
DEU     5387
ITA     3066
IRL     3016
BEL     2081
BRA     1995
NLD     1911
Name: country, dtype: int64
```

From the above bar chart visualization we can notice that most of the hotel bookings are happening in **"PTR(Portugal)"** country. we can also observe that the maximum people are preferring city hotels compared to Resort Hotels.

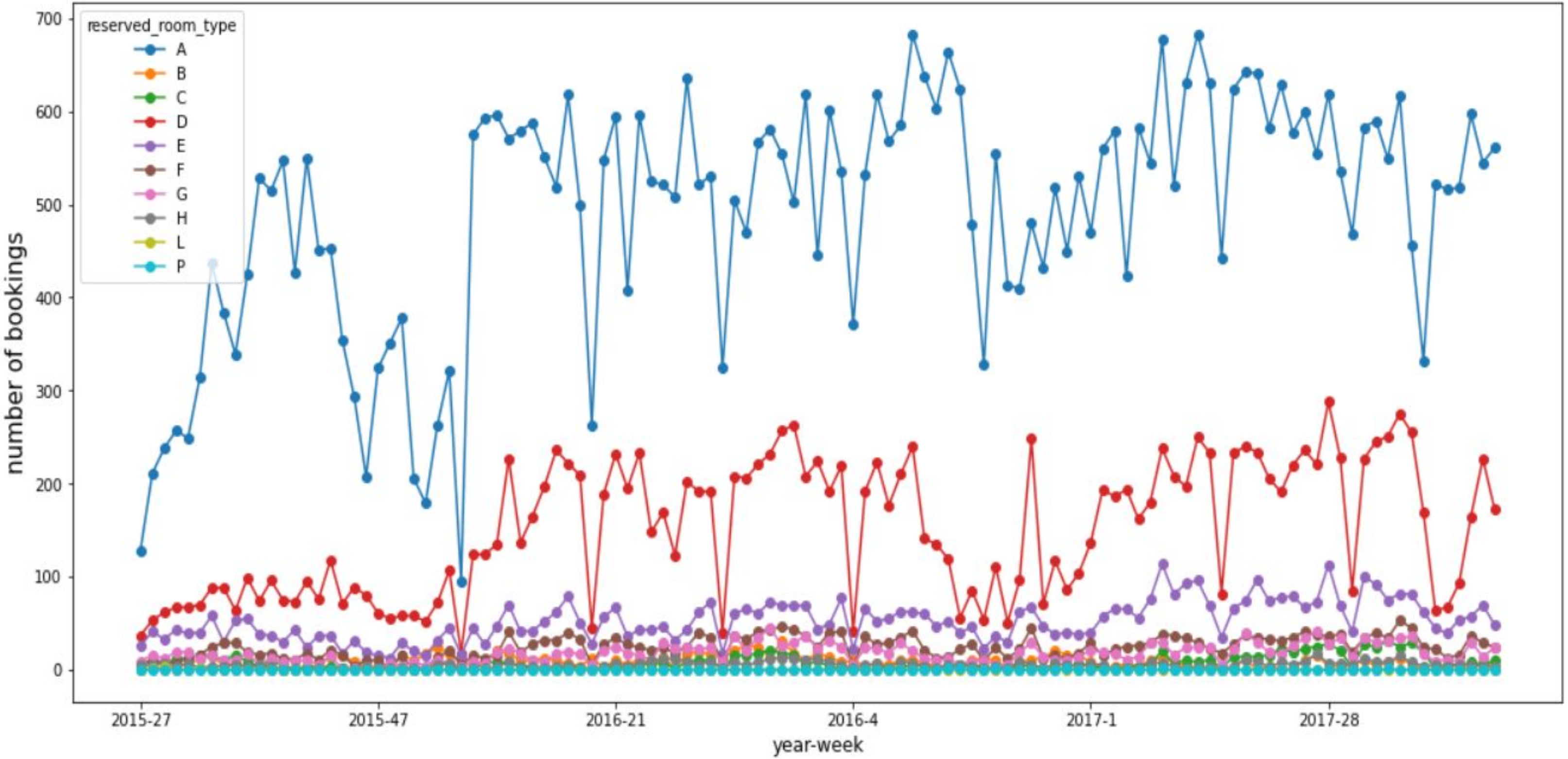
Demand of Room Types with respect to weeks of years(2015-17):





reserved_room_type	A	B	C	D	E	F	G	H	L	P
arrival_week_year										
2015-27	128.0	0.0	6.0	37.0	26.0	7.0	9.0	5.0	0.0	0.0
2015-28	211.0	2.0	11.0	53.0	41.0	8.0	16.0	5.0	1.0	0.0
2015-29	238.0	0.0	7.0	62.0	33.0	9.0	12.0	7.0	3.0	0.0
2015-30	257.0	4.0	12.0	67.0	43.0	7.0	19.0	4.0	0.0	0.0
2015-31	249.0	6.0	12.0	67.0	39.0	13.0	19.0	7.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
2017-5	516.0	2.0	7.0	67.0	39.0	13.0	11.0	3.0	0.0	0.0
2017-6	518.0	6.0	4.0	93.0	54.0	16.0	9.0	6.0	0.0	0.0
2017-7	597.0	1.0	8.0	165.0	57.0	37.0	30.0	6.0	0.0	0.0
2017-8	545.0	8.0	6.0	227.0	70.0	29.0	14.0	3.0	0.0	0.0
2017-9	561.0	4.0	10.0	172.0	48.0	24.0	24.0	4.0	0.0	0.0

Most demanded room types are A next comes D and least bookings are done for room type P and L.

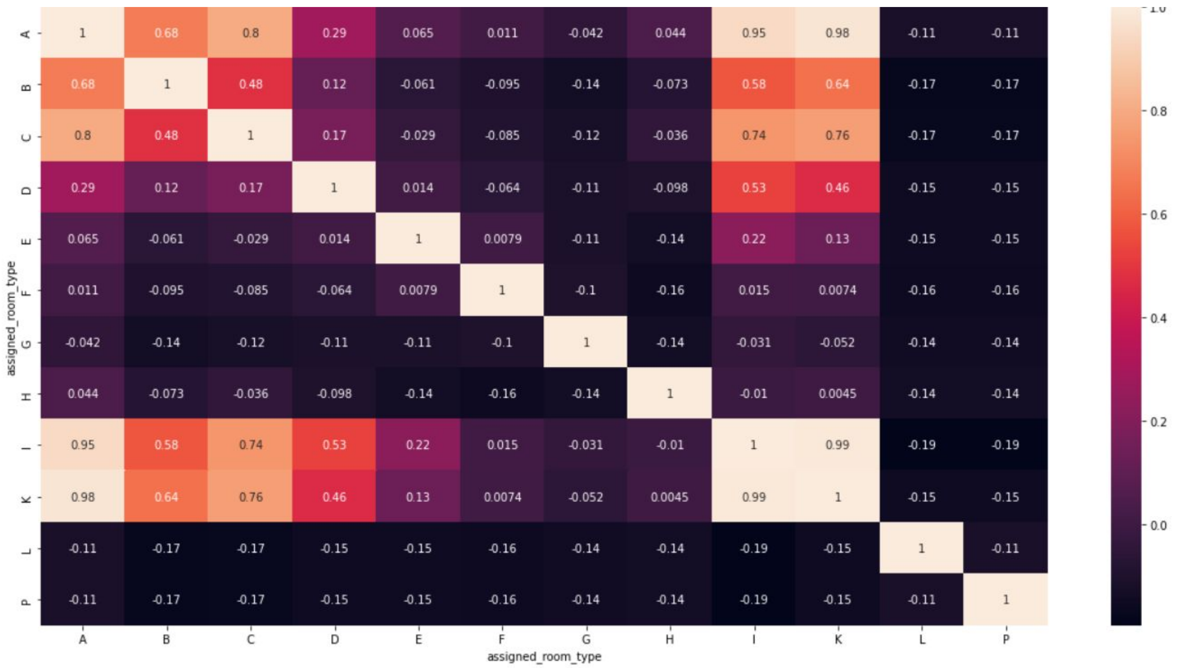


Maximum bookings were happened in 9th and 12th week of every year.

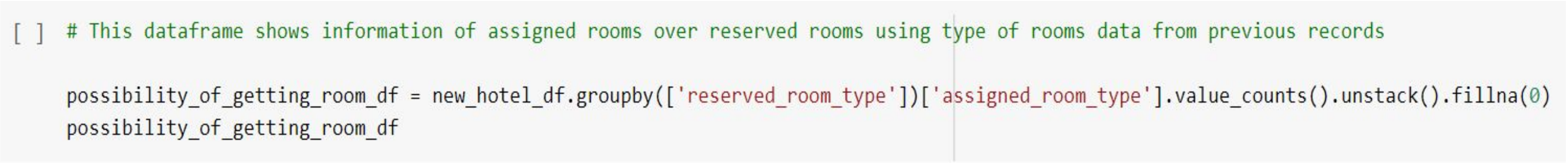
The possibility of getting the reserved room type

The lighter color indicates the more probability of getting the reserved type of room and the darker color indicates the less/no probability of getting the room of customer choice.

Probability of room allocation for the customer choice reserved type A is in the order - A, K, I, C, B.

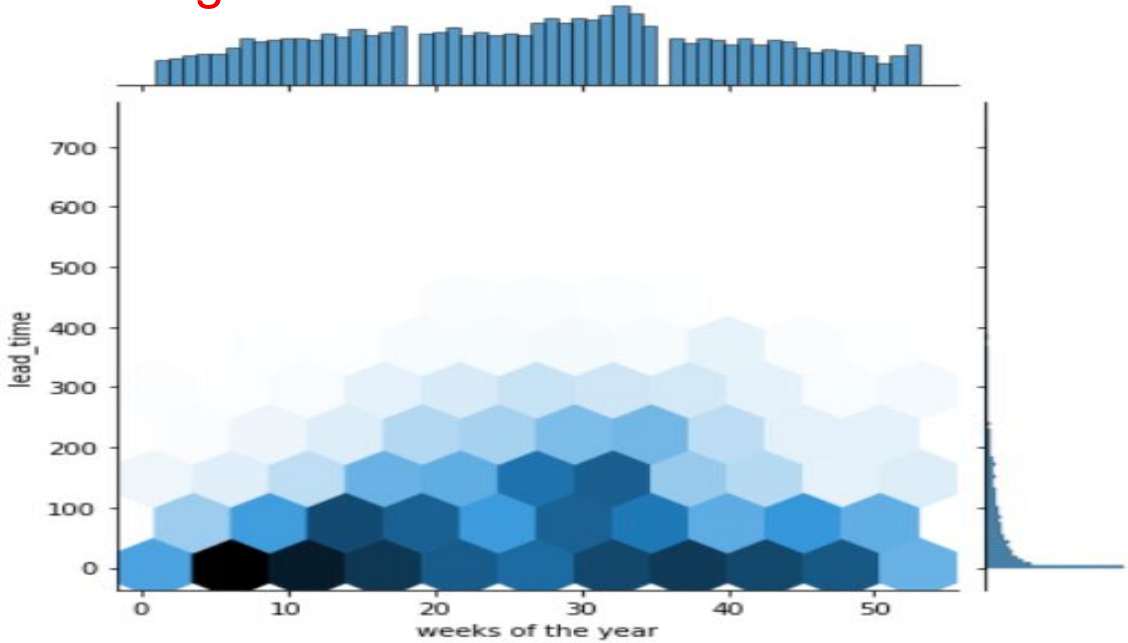


	A	B	C	D	E	F	G	H	I	K	L	P
A	1	0.68	0.8	0.29	0.065	0.011	-0.042	0.044	0.95	0.98	-0.11	-0.11
B	0.68	1	0.48	0.12	-0.061	-0.095	-0.14	-0.073	0.58	0.64	-0.17	-0.17
C	0.8	0.48	1	0.17	-0.029	-0.085	-0.12	-0.036	0.74	0.76	-0.17	-0.17
D	0.29	0.12	0.17	1	0.014	-0.064	-0.11	-0.098	0.53	0.46	-0.15	-0.15
E	0.065	-0.061	-0.029	0.014	1	0.0079	-0.11	-0.14	0.22	0.13	-0.15	-0.15
F	0.011	-0.095	-0.085	-0.064	0.0079	1	-0.1	-0.16	0.015	0.0074	-0.16	-0.16
G	-0.042	-0.14	-0.12	-0.11	-0.11	-0.1	1	-0.14	-0.031	-0.052	-0.14	-0.14
H	0.044	-0.073	-0.036	-0.098	-0.14	-0.16	-0.14	1	-0.01	0.0045	-0.14	-0.14
I	0.95	0.58	0.74	0.53	0.22	0.015	-0.031	-0.01	1	0.99	-0.19	-0.19
K	0.98	0.64	0.76	0.46	0.13	0.0074	-0.052	0.0045	0.99	1	-0.15	-0.15
L	-0.11	-0.17	-0.17	-0.15	-0.15	-0.16	-0.14	-0.14	-0.19	-0.15	1	-0.11
P	-0.11	-0.17	-0.17	-0.15	-0.15	-0.16	-0.14	-0.14	-0.19	-0.15	-0.11	1



```
[ ]  # This dataframe shows information of assigned rooms over reserved rooms using type of rooms data from previous records

     possibility_of_getting_room_df = new_hotel_df.groupby(['reserved_room_type'])['assigned_room_type'].value_counts().unstack().fillna(0)
     possibility_of_getting_room_df
```

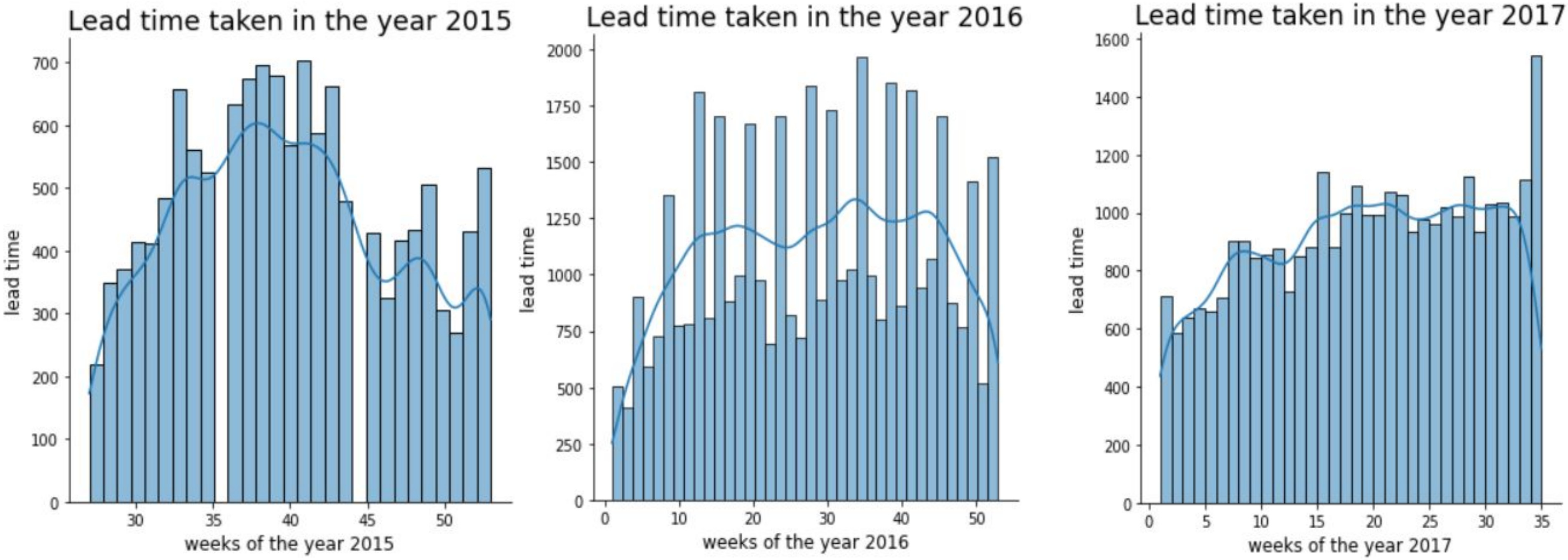
Analyzing Lead Time Of Bookings



The bar plots represents max bookings done in the weeks in X-axis i.e., (max bookings done b/w week 30 to 35) and max lead_time taken for booking in Y-axis i.e., (max lead time taken is 0-immediate booking).

Hex plot represents max lead time taken in the b/w 11th - 18th weeks of the year.

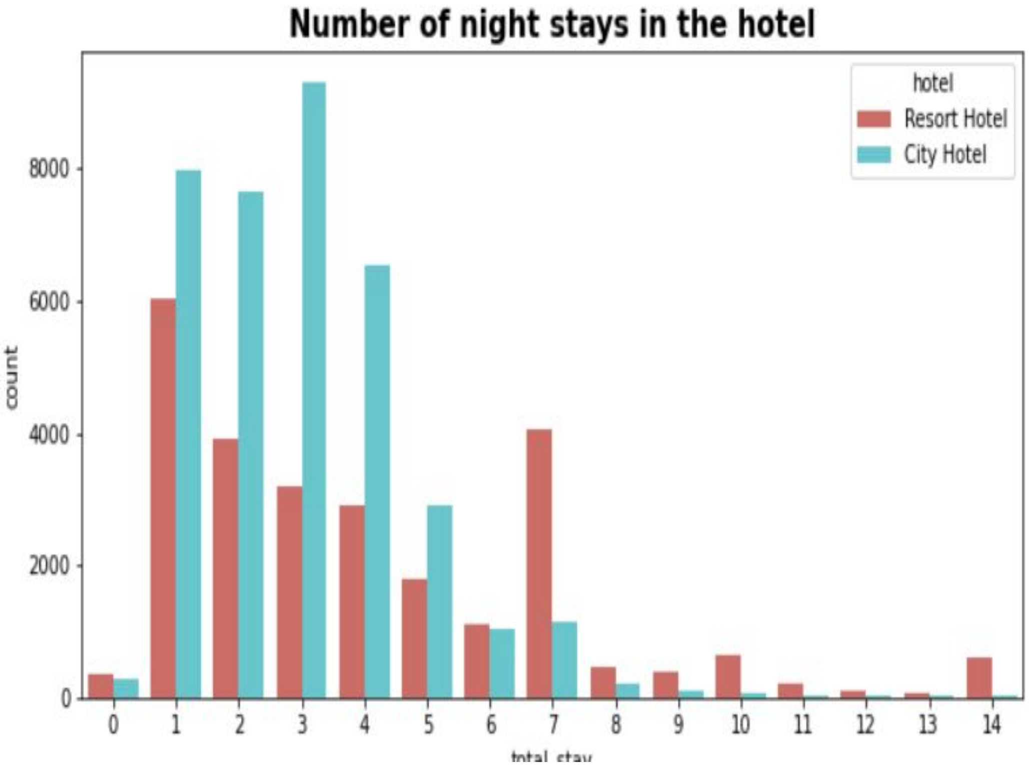
Analysis of lead time year by year

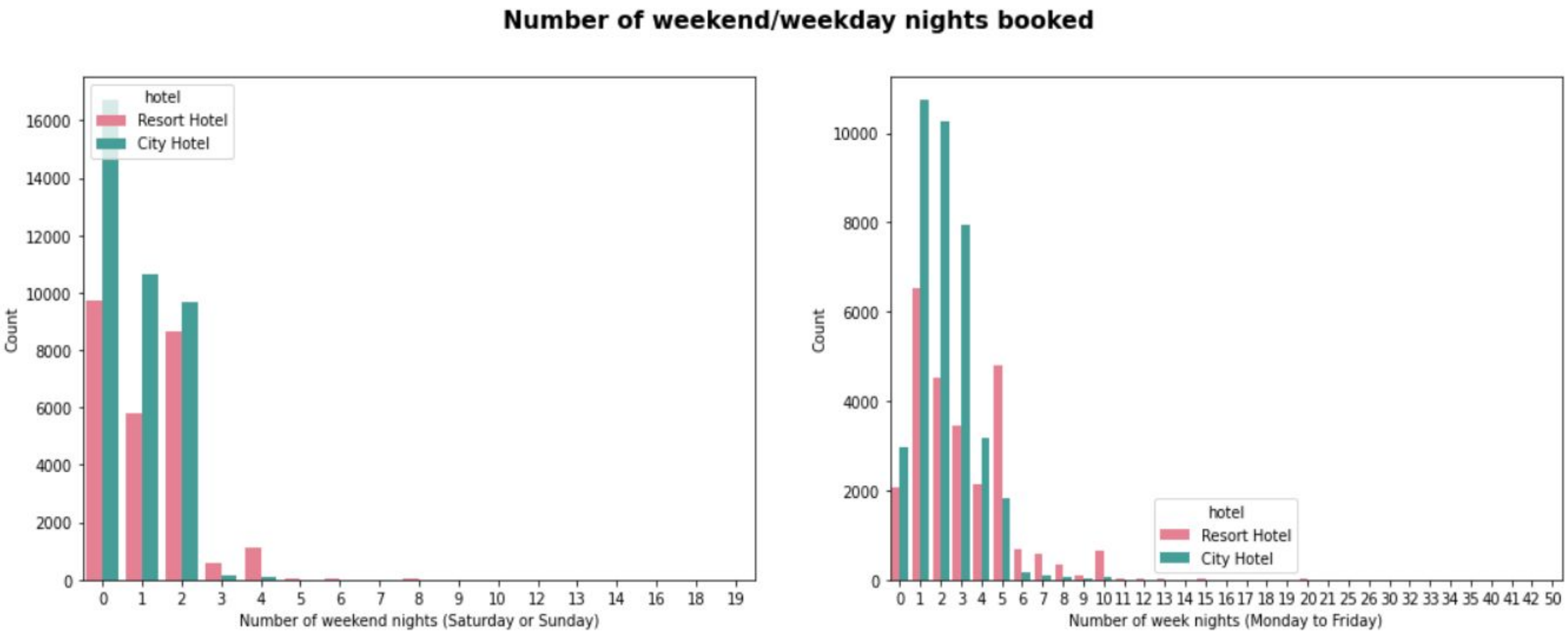


From the above displot analysis we can conclude that maximum lead time taken in bookings are in the year 2016. ‘UEFA Euro 2016 Final’ held in France may be the one of the reason of prior booking of hotels happened with high Lead Time.

Number of days people stay in the hotel

- We can notice that majority of people stay or do a booking of ‘7’ or less than ‘7’ days.
- Maximum night bookings are happening in the city hotels and the max length of stay is ‘3’ days.
- Maximum night bookings length of stay happening in the Resort Hotel are for one night stay.
- We can also observe that if the stay is longer than 7 days then guests are preferring to book Resort Hotels only.





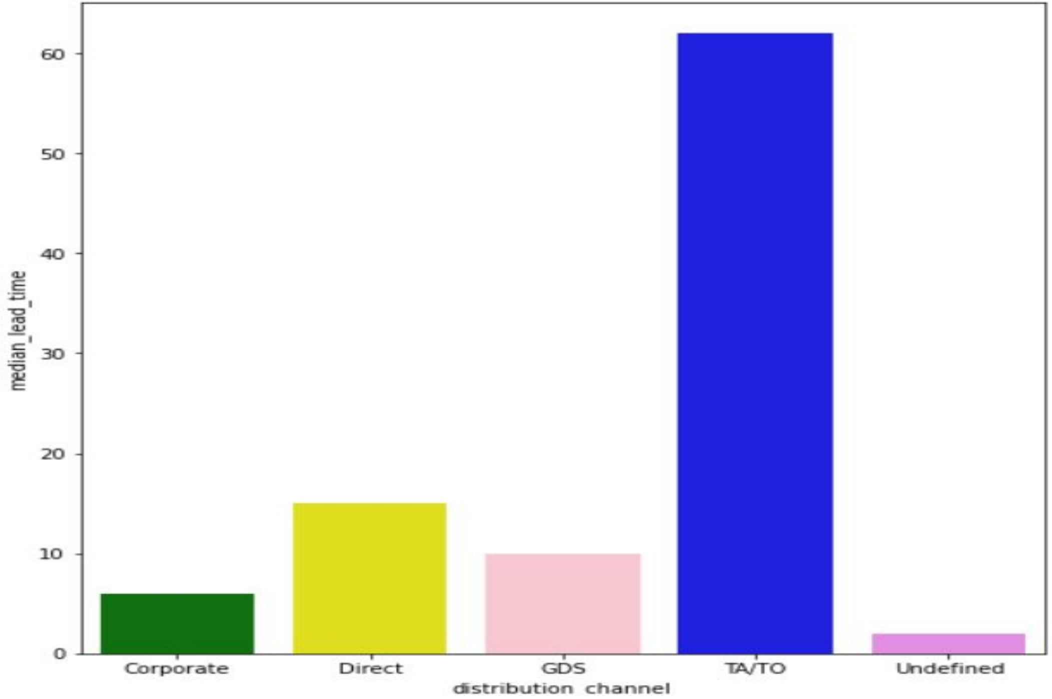
We can see that majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get best daily rate is '5' for week nights and '2' for weekend nights.

max night bookings are happening in the city hotels in weekdays and the max length of stay is 1 to 2 days.

Analyzing on the basis of distribution channel

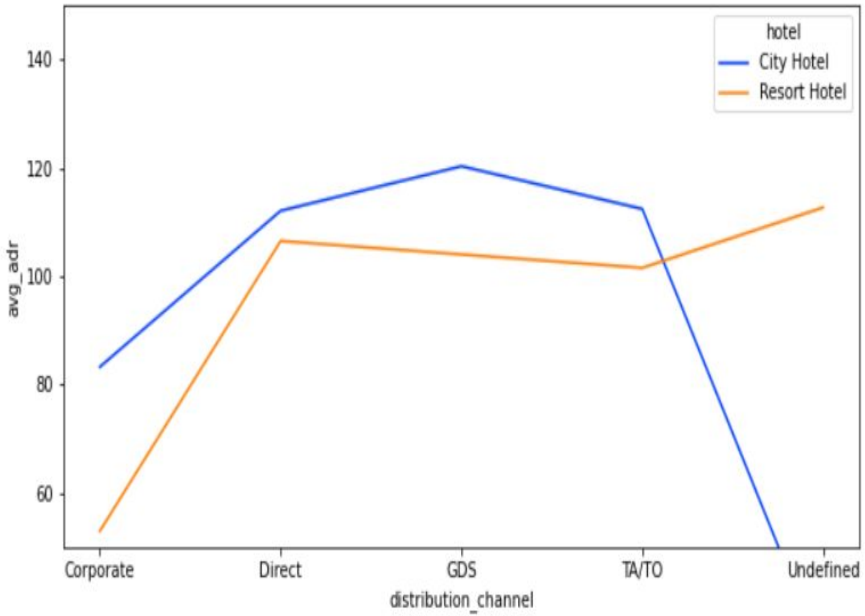
Distribution channel v/s median lead time:

Distribution channel is the costumer accessed by corporate booking/Direct/Travel agent(TA).Travel operator(TO) and Median lead time is the median of number of days that elapsed between the entering date of the booking into the PMS and arriving date.



Through TA/TO distribution channels, bookings were with high lead time i.e., they are booking early compare to other distribution channels.

ADR(Average daily rate) generated through various distribution channels



Average daily rate (ADR), one of the three key hotel performance indicators (along with occupancy and Revenue per available room(RevPAR)), is the measure of the average paid for rooms sold in a given time period. The metric covers only revenue-generating guestrooms.

How to calculate ADR:

ADR is calculated by dividing room revenue by rooms sold. The metric is of course applicable for any currency.

ADR = Room Revenue/Rooms Sold

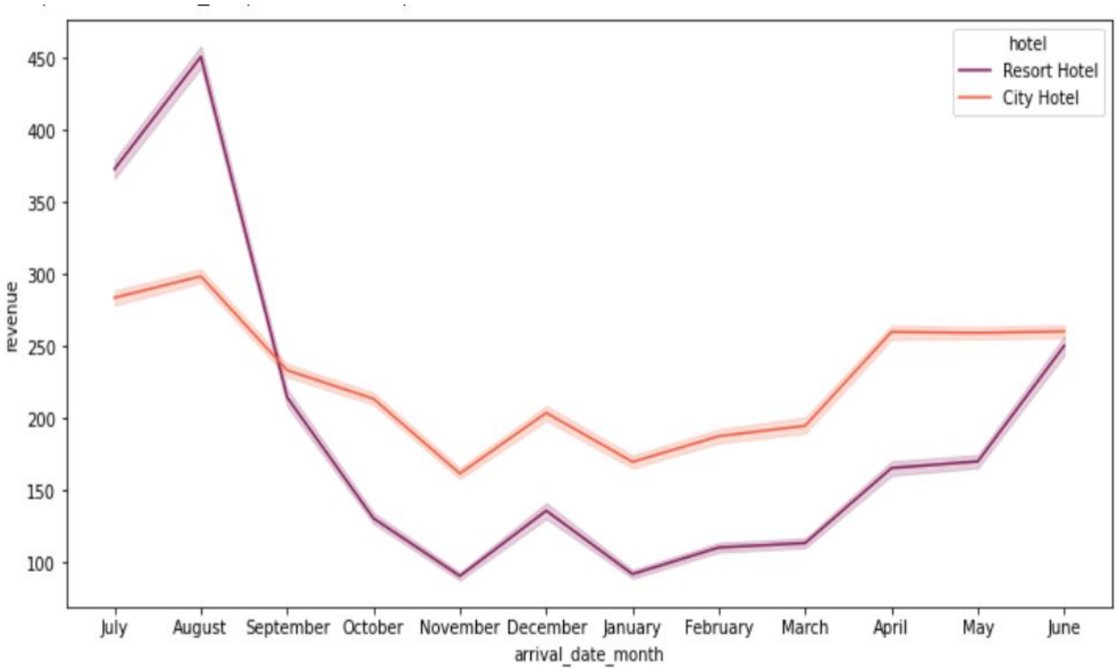
As ADR is the revenue determining factor 'GDS distribution channel' of city hotel bookings are achieving high adr (revenue).

Average Revenue of the Hotel

ADR = Room Revenue/Rooms Sold

Room Revenue = ADR * Rooms Sold

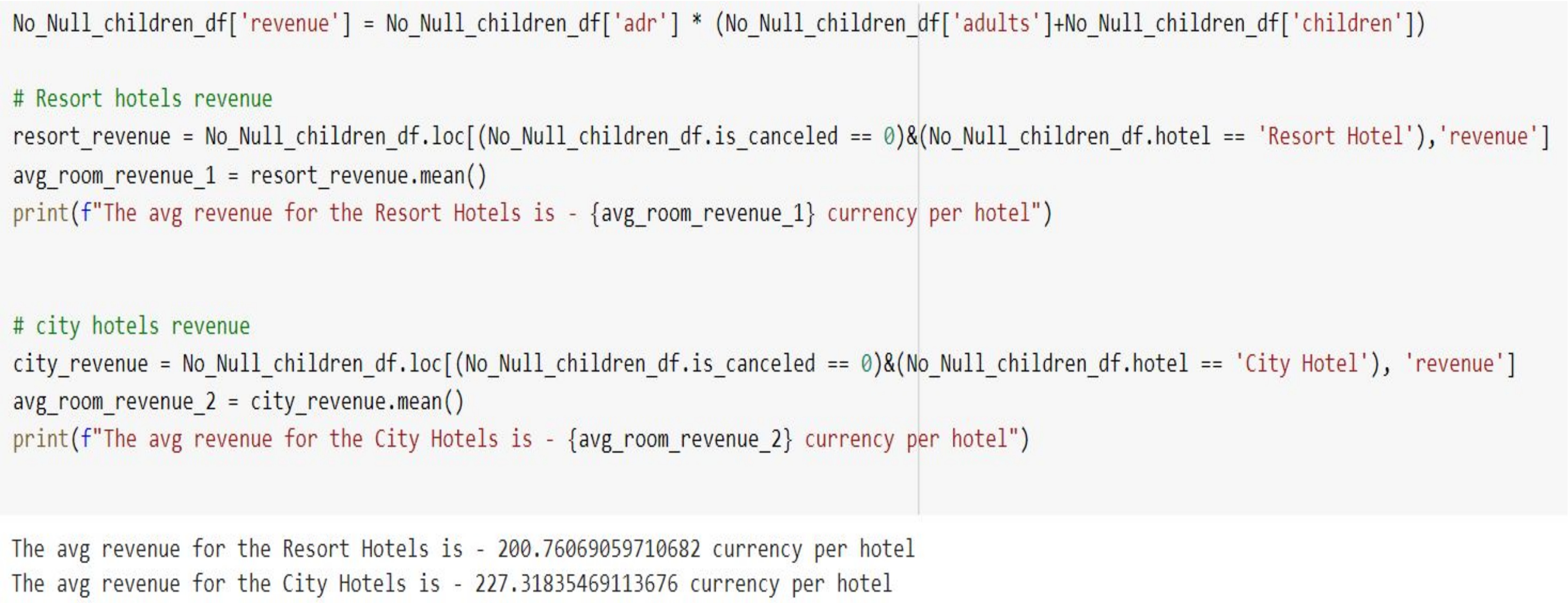
Avg Room Revenue = mean ADR * mean Rooms Sold



From the above analysis we can notice that the Resort hotels are getting highest revenue in the month of 'august', 'july' and then decreases drastically. City hotel's revenue is almost constant all over year.

Average Revenue of the hotels from 2015-17 is calculated as follows:

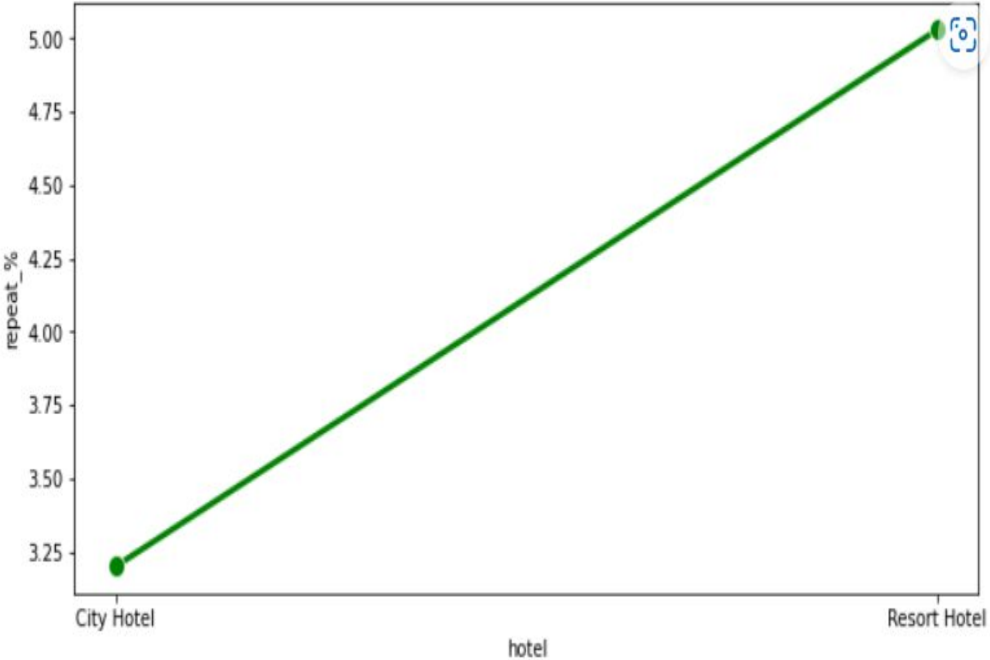
Rooms sold are calculated based on no. of booking i.e., no. of adult bookings+ no. of children bookings



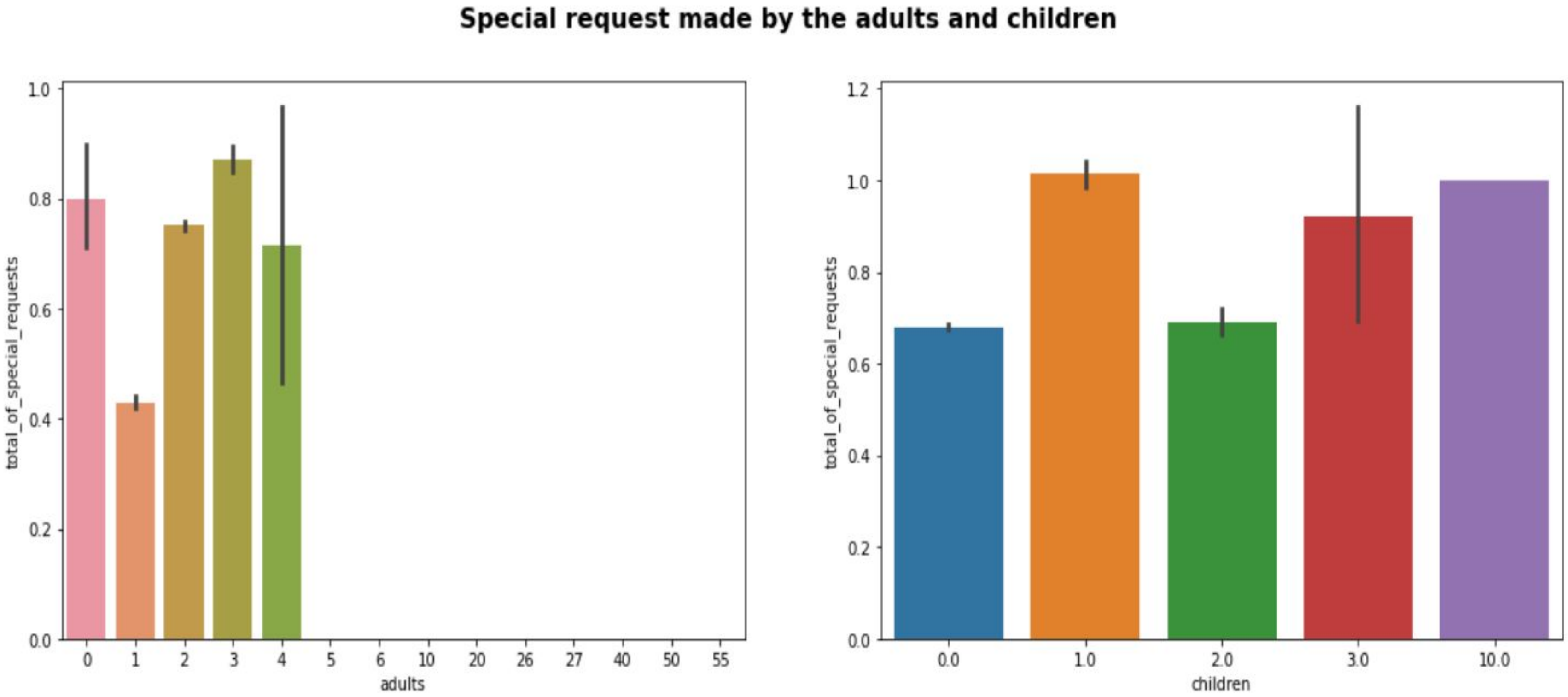
```
No_Null_children_df['revenue'] = No_Null_children_df['adr'] * (No_Null_children_df['adults']+No_Null_children_df['children'])

# Resort hotels revenue
resort_revenue = No_Null_children_df.loc[(No_Null_children_df.is_canceled == 0)&(No_Null_children_df.hotel == 'Resort Hotel'),'revenue']
avg_room_revenue_1 = resort_revenue.mean()
print(f"The avg revenue for the Resort Hotels is - {avg_room_revenue_1} currency per hotel")


# city hotels revenue
city_revenue = No_Null_children_df.loc[(No_Null_children_df.is_canceled == 0)&(No_Null_children_df.hotel == 'City Hotel'), 'revenue']
avg_room_revenue_2 = city_revenue.mean()
print(f"The avg revenue for the City Hotels is - {avg_room_revenue_2} currency per hotel")
```

```
The avg revenue for the Resort Hotels is - 200.76069059710682 currency per hotel
The avg revenue for the City Hotels is - 227.31835469113676 currency per hotel
```

Chances of customer will return



There is a very less probability that the customer will repeat. But the return percentage of resort is slightly greater than that of city hotel.



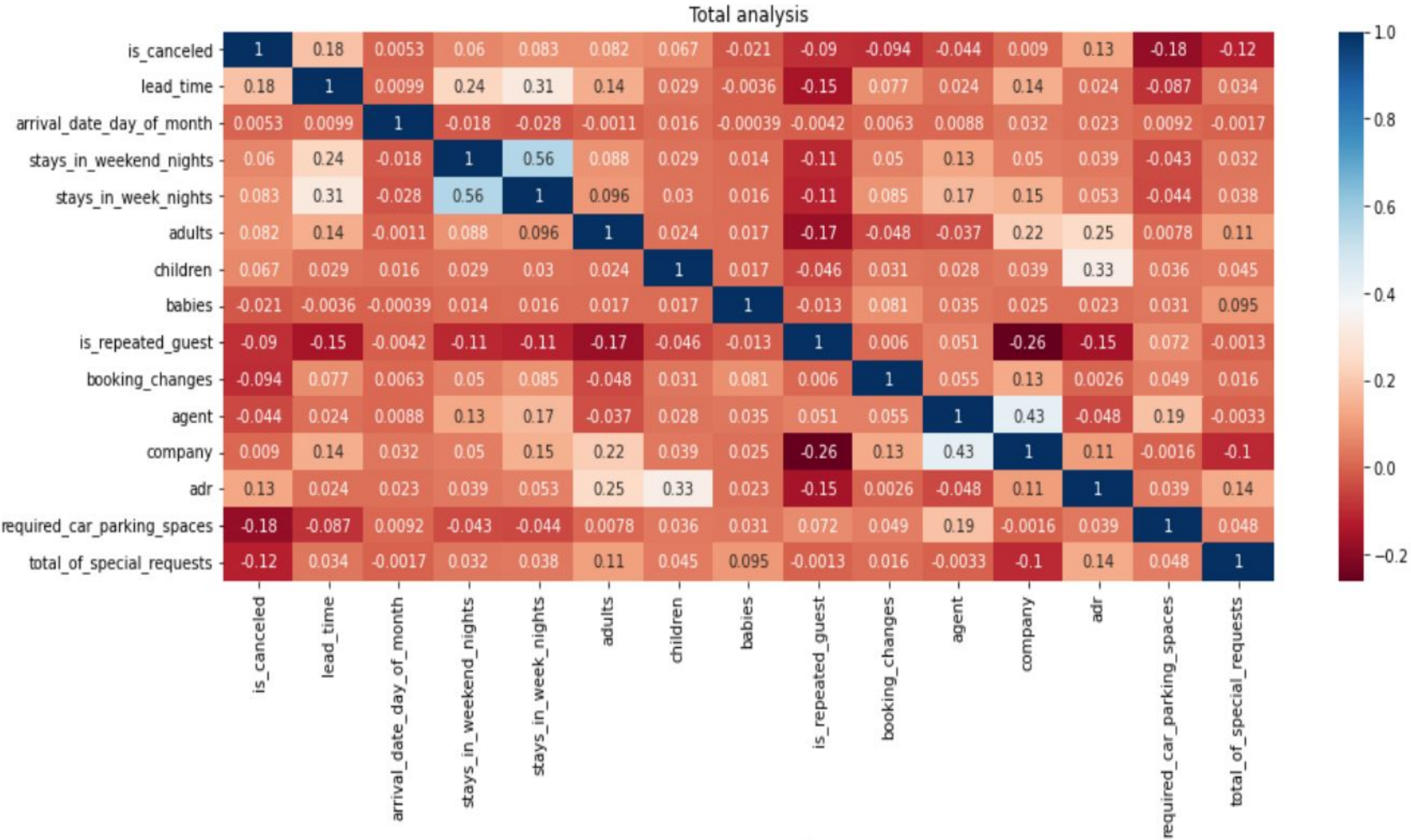
Special request made by the adults and children

We can see that if the adults are more than 2, there are high chances that the hotel receives more special requests and the no. of special requests for children has no much variation.

EDA of all Total numerical data of given data set:

The blue colour shows the highest probability of the column feature's influence with respect to that of the row feature in correlation matrix.

E.g.: cancellation of bookings mainly influenced by 'Lead time' and 'adr'(Average daily rate)



Total analysis

	is_canceled	lead_time	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	booking_changes	agent	company	adr	required_car_parking_spaces	total_of_special_requests
is_canceled	1	0.18	0.0053	0.06	0.083	0.082	0.067	-0.021	-0.09	-0.094	-0.044	0.009	0.13	-0.18	-0.12
lead_time	0.18	1	0.0099	0.24	0.31	0.14	0.029	-0.0036	-0.15	0.077	0.024	0.14	0.024	-0.087	0.034
arrival_date_day_of_month	0.0053	0.0099	1	-0.018	-0.028	-0.0011	0.016	-0.00039	-0.0042	0.0063	0.0088	0.032	0.023	0.0092	-0.0017
stays_in_weekend_nights	0.06	0.24	-0.018	1	0.56	0.088	0.029	0.014	-0.11	0.05	0.13	0.05	0.039	-0.043	0.032
stays_in_week_nights	0.083	0.31	-0.028	0.56	1	0.096	0.03	0.016	-0.11	0.085	0.17	0.15	0.053	-0.044	0.038
adults	0.082	0.14	-0.0011	0.088	0.096	1	0.024	0.017	-0.17	-0.048	-0.037	0.22	0.25	0.0078	0.11
children	0.067	0.029	0.016	0.029	0.03	0.024	1	0.017	-0.046	0.031	0.028	0.039	0.33	0.036	0.045
babies	-0.021	-0.0036	-0.00039	0.014	0.016	0.017	0.017	1	-0.013	0.081	0.035	0.025	0.023	0.031	0.095
is_repeated_guest	-0.09	-0.15	-0.0042	-0.11	-0.11	-0.17	-0.046	-0.013	1	0.006	0.051	-0.26	-0.15	0.072	-0.0013
booking_changes	-0.094	0.077	0.0063	0.05	0.085	-0.048	0.031	0.081	0.006	1	0.055	0.13	0.0026	0.049	0.016
agent	-0.044	0.024	0.0088	0.13	0.17	-0.037	0.028	0.035	0.051	0.055	1	0.43	-0.048	0.19	-0.0033
company	0.009	0.14	0.032	0.05	0.15	0.22	0.039	0.025	-0.26	0.13	0.43	1	0.11	-0.0016	-0.1
adr	0.13	0.024	0.023	0.039	0.053	0.25	0.33	0.023	-0.15	0.0026	-0.048	0.11	1	0.039	0.14
required_car_parking_spaces	-0.18	-0.087	0.0092	-0.043	-0.044	0.0078	0.036	0.031	0.072	0.049	0.19	-0.0016	0.039	1	0.048
total_of_special_requests	-0.12	0.034	-0.0017	0.032	0.038	0.11	0.045	0.095	-0.0013	0.016	-0.0033	-0.1	0.14	0.048	1

Conclusion:

- Around 61% bookings are of City hotel and 39% bookings areof Resort hotel, therefore City hotels are busier than the Resort Hotels.
- Around 27% of total bookings are cancelled, in that 66.6% cancellations are happening in City hotels.
  In both resort and city hotels most of the bookings are happening in "PTR(Portugal)" country. we can also observe that the maximum people are preferring city hotels compared to Resort Hotels.
- Most of the city and resort bookings are happening in the month of August. Followed by July. Least bookings are happening in the month of January, November and December.
- The Resort hotels are getting highest revenue in the month of 'august', 'july' and then decreasing drastically. City hotel's revenue is almost constant all over year.
- Maximum bookings are happening in 9th and 12th week of every year.
- 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast'.
- Most demanded room types are A next comes D and least demanded are of room type P and L.
- Probability of room allocation for the customer choice reserved type A is in the order - A, K, I, C, B.

Conclusion (continued.....):

- High probability of lead time taken is '0' i.e., immediate booking are happening but high leadtime taken is in the b/w 11th - 18th weeks of the year. Hence, this time is the busiest time of the year.
- The maximum lead time taken in bookings is in the year 2016.
- Majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get best daily rate is '5' for week nights and '2' for weekend nights.
- Max night bookings are happening in the city hotels in weekdays and the max length of stay is 1 to 2 days.
- Through TA/TO distribution channels, bookings happened with high lead time i.e, they are booking early compare to other distribution channels.
- As (Average daily rate) ADR is the revenue determining factor 'GDS distribution channel' of city hotel bookings are achieving high adr (revenue).

STRENGTHENING TOURISM SECTOR

- During covid 19, tourism sector has hit its lowest .So the revival of tourism sector not only help in revenue generation in that sector only but also help the economy overall.
- In this analysis each and every aspect has been analyzed properly

- For this various things has been taken into consideration to improve the revenue generation in the hotel industry.

thank you