Capstone Project

Hotel Booking Analysis

Individual Project:

Anupam Mishra

➢ Let's Analyse Hotel Booking

Data Exploration

- Observe the Data
- Find Missing

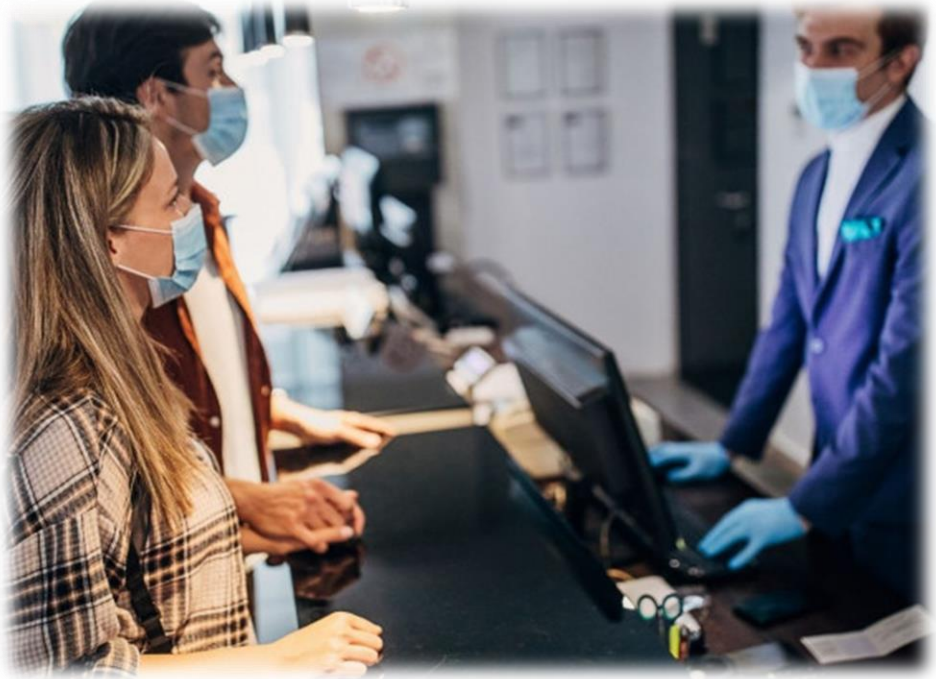
Data Cleaning

- Replace the Null Values
- Drop un-necessary columns

Analyse the data

Visualise the data

Conclusion



➢ Fast Growing Ever Green Hotel Business

- **A hotel is an establishment that provides lodging, meals and other services for travelers and other paying guests.**

➢ First Five Rows Data

```
# Checking first 5 rows in dataset
df.head(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_w
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

5 rows × 34 columns

➢ Last Five Rows Data

```
# Checking the last 5 rows in dataset
df.tail(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays
119385	City Hotel	0	23	2017	August	35	30	2	
119386	City Hotel	0	102	2017	August	35	31	2	
119387	City Hotel	0	34	2017	August	35	31	2	
119388	City Hotel	0	109	2017	August	35	31	2	
119389	City Hotel	0	205	2017	August	35	29	2	

5 rows × 34 columns

➢ Explore The Dataset

Go Through The Dataset

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
City Hotel	0	44	2017	August	35
City Hotel	0	188	2017	August	35
City Hotel	0	135	2017	August	35
City Hotel	0	164	2017	August	35
City Hotel	0	21	2017	August	35
City Hotel	0	23	2017	August	35
City Hotel	0	102	2017	August	35
City Hotel	0	34	2017	August	35
City Hotel	0	109	2017	August	35

Checking Null Values In Dataset

```
country                          488
market_segment                     0
distribution_channel               0
is_repeated_guest                  0
previous_cancellations             0
previous_bookings_not_canceled     0
reserved_room_type                 0
assigned_room_type                 0
booking_changes                    0
deposit_type                       0
agent                          16340
company                       112593
days_in_waiting_list               0
customer_type                      0
```

Replacing The Null Values With Their Mean

```
country                          0
market_segment                   0
distribution_channel             0
is_repeated_guest                0
previous_cancellations           0
previous_bookings_not_canceled   0
reserved_room_type               0
assigned_room_type               0
booking_changes                  0
deposit_type                     0
agent                            0
company                          0
days_in_waiting_list             0
customer_type                    0
```

➢ Description of the Data

```
[8]  # Exploring descriptive statistical parameter

     df.describe()
```

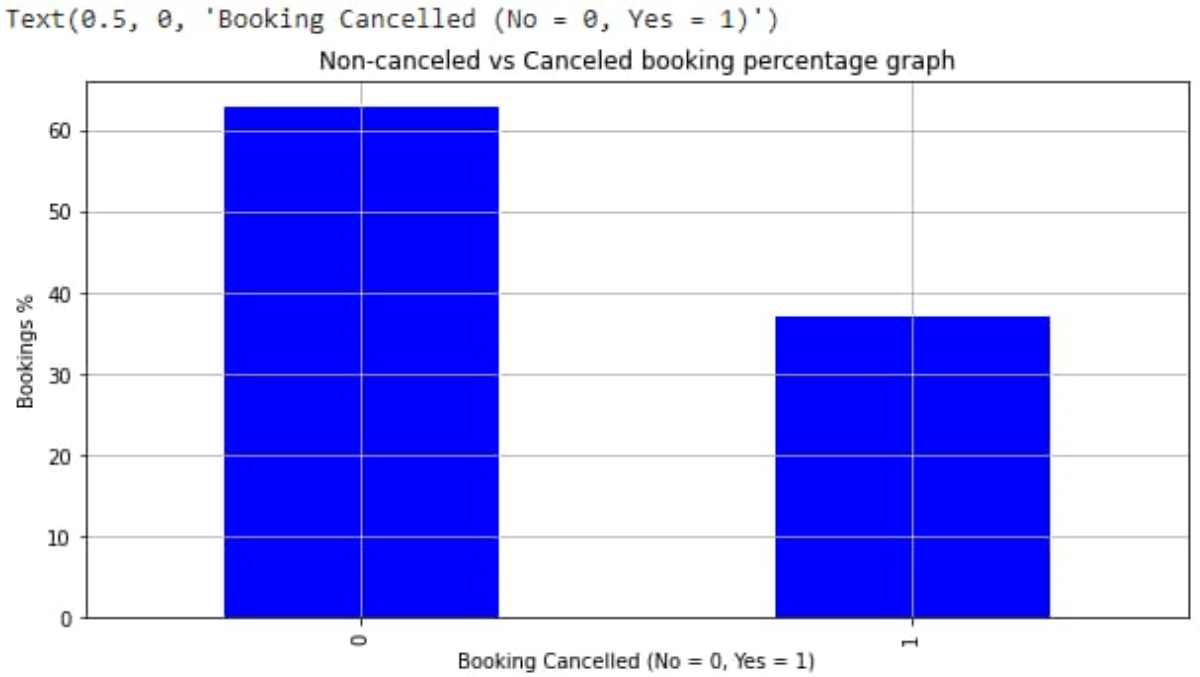
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.0
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	1.8
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.5
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.0
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.0
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.0
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.0
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.0

➢ Canceled & Check-Out Booking

- 75166 customers are actually checked-out to the hotel
- 43017 customers are cancelling their bookings.
- Due to this canceled bookings there will be an adverse effect on hotel business which means hotels are not able to make more profit, they are losing their customers.

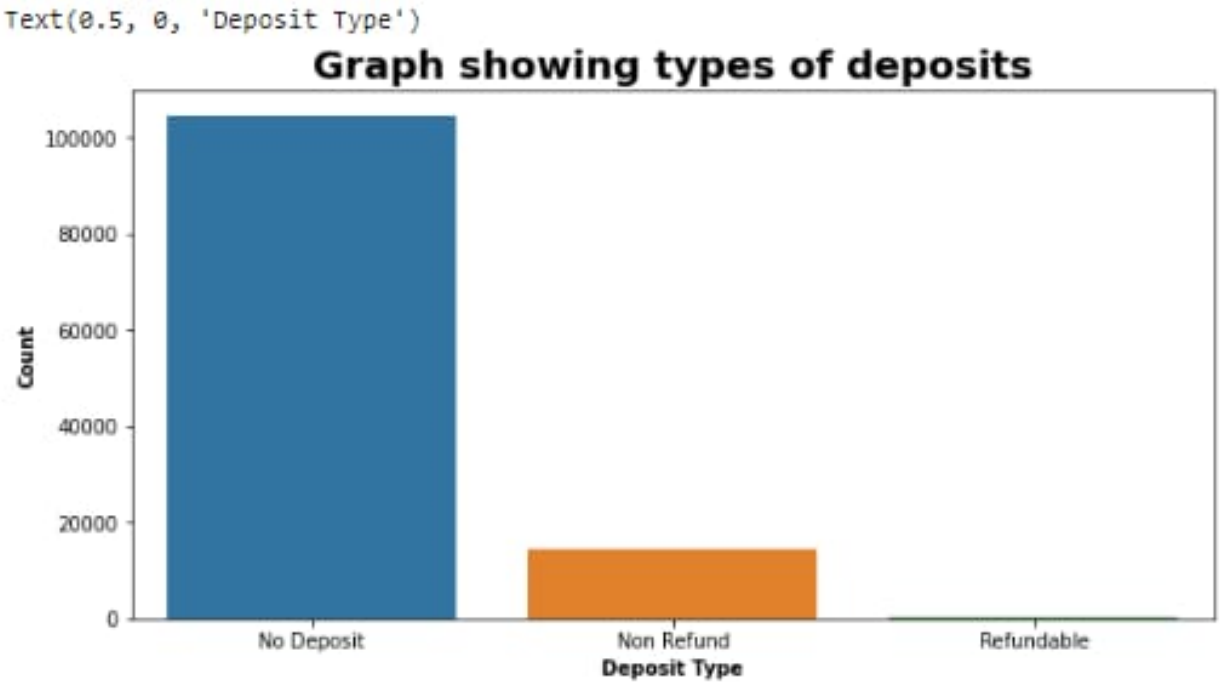
Non-Canceled vs Canceled Booking Percentage

As we saw total number of booking from our last slide, here we are going to see the same but in terms of percentage. This bar graph representing that 63% of customers are check-in to hotels where 37% of customers canceled their bookings.



➢ Deposit Policies Of Hotels

Very large amount of hotels have “No Deposit” Policy. And this may be the reason for cancellation of high amount of bookings. To avoid this booking cancellation, in account to collect more profit and customers- “No Deposit” policy should be change.



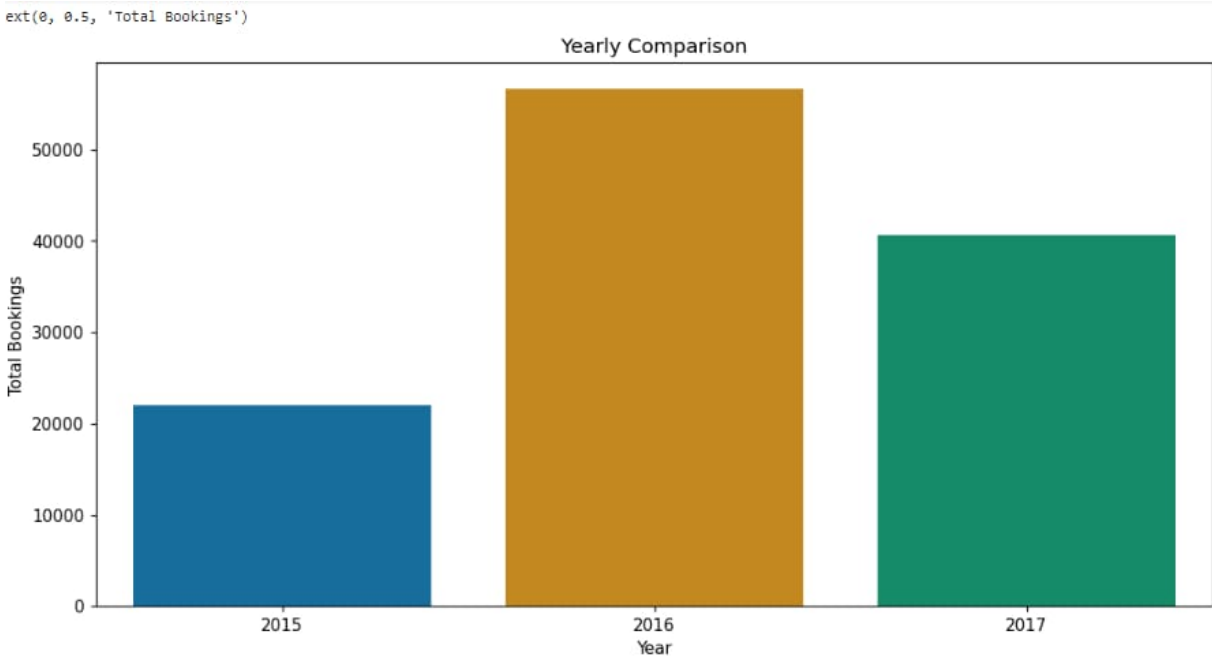
➢ Total Number Of Bookings Across Different Years

How many customers actually checked-in to the hotel across different years?

Let us find out with simple bar chart, in 2015 there are 18.5% of customers checking in.

Whereas in 2016 we can see that there is increase in bookings up to 47.4%.

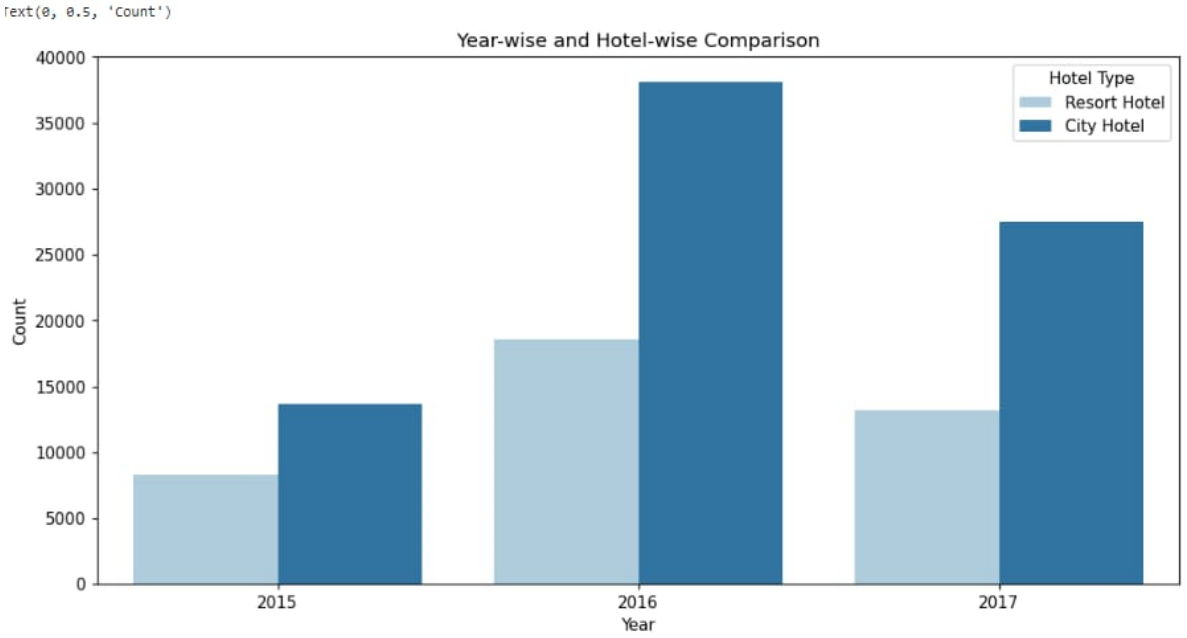
This increase in trend did not sustain for more time, going downward in 2017 with only 34.1% bookings.



➢ Demand Trend Of Hotels Year wise

Which type of hotels customer preferred to stay in different years?

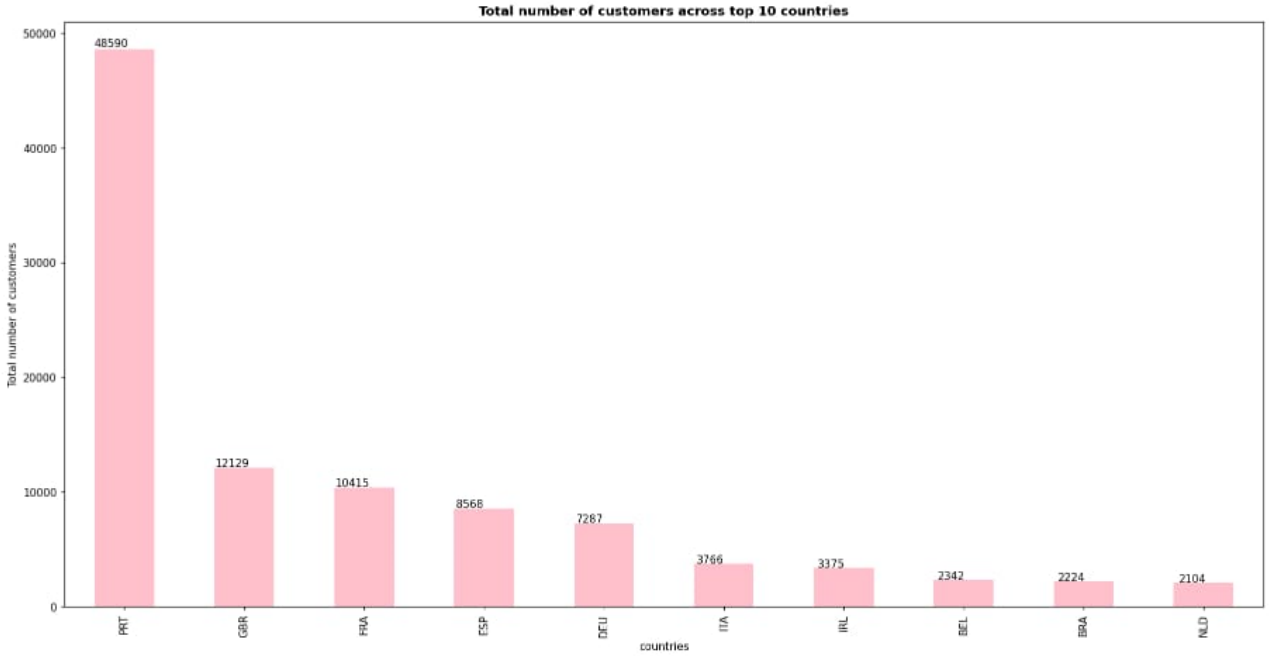
Here, we plotted a subplot for Resort hotel and City hotel. From these columns we can conclude that there is always demand of City hotels as compared to Resort hotels across three different years 2015, 2016 and 2017. As we discussed early, after increasing the booking trend it got decreased again. This happened in both cases – Resort as well as for City hotels.



Top 10 Countries With Maximum Customers

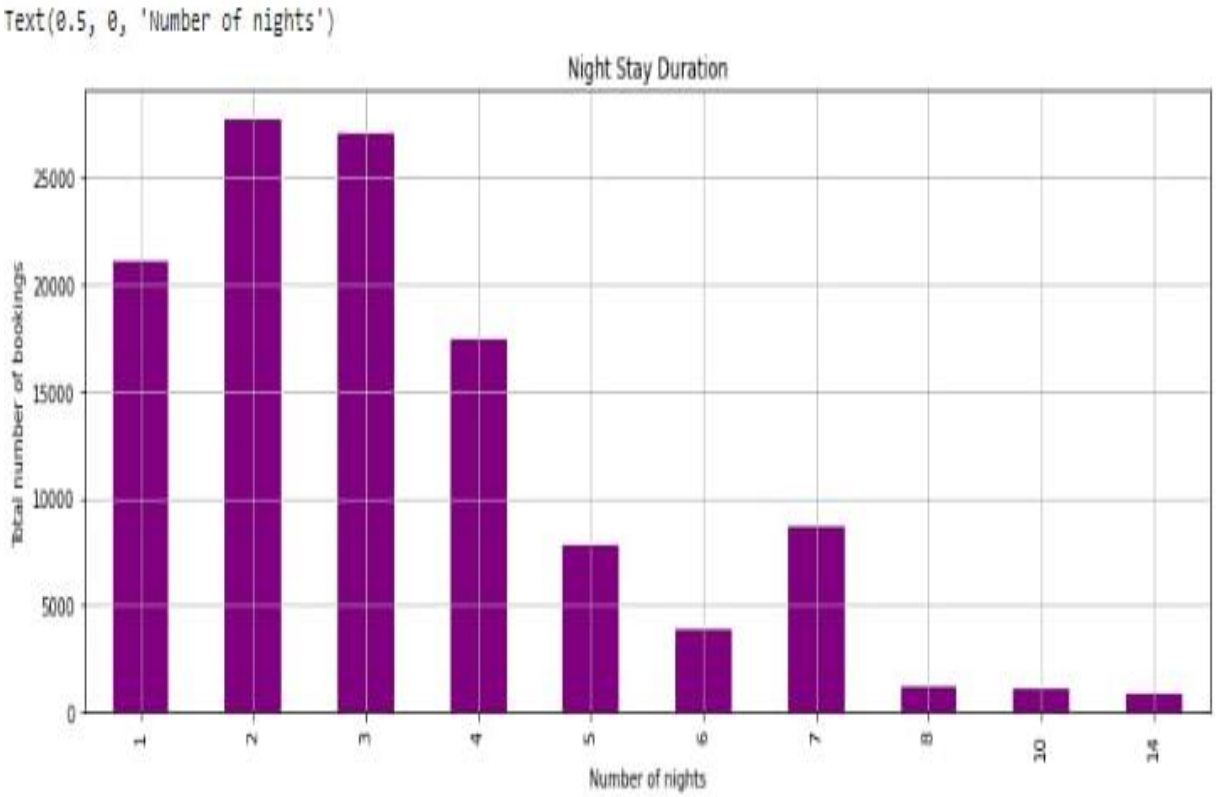
Which are those countries giving maximum customers?

So, here is the result – after analysing the dataset we found that Portugal is on Rank 1 with 48590 customers followed by UK, France, Spain, Germany with 12129, 10415, 8568, 7287 customers accordingly. After these 5 topmost countries Italy, Ireland, Belgium, Brazil and Netherlands has 3766, 3375, 2342, 2224 and 2104 customers i.e., Netherlands sits back with lowest number of customers.



➢ Night Stay Duration

By combining the two columns of stays_in_week_nights and stays_in weekend_nights we got total number of nights. Hence, we can say that more customers like to spend 2 – 3 nights where some customer prefer to stay for 1 – 4 nights. Very few customers are there who are interested to stay for more than 5 days.



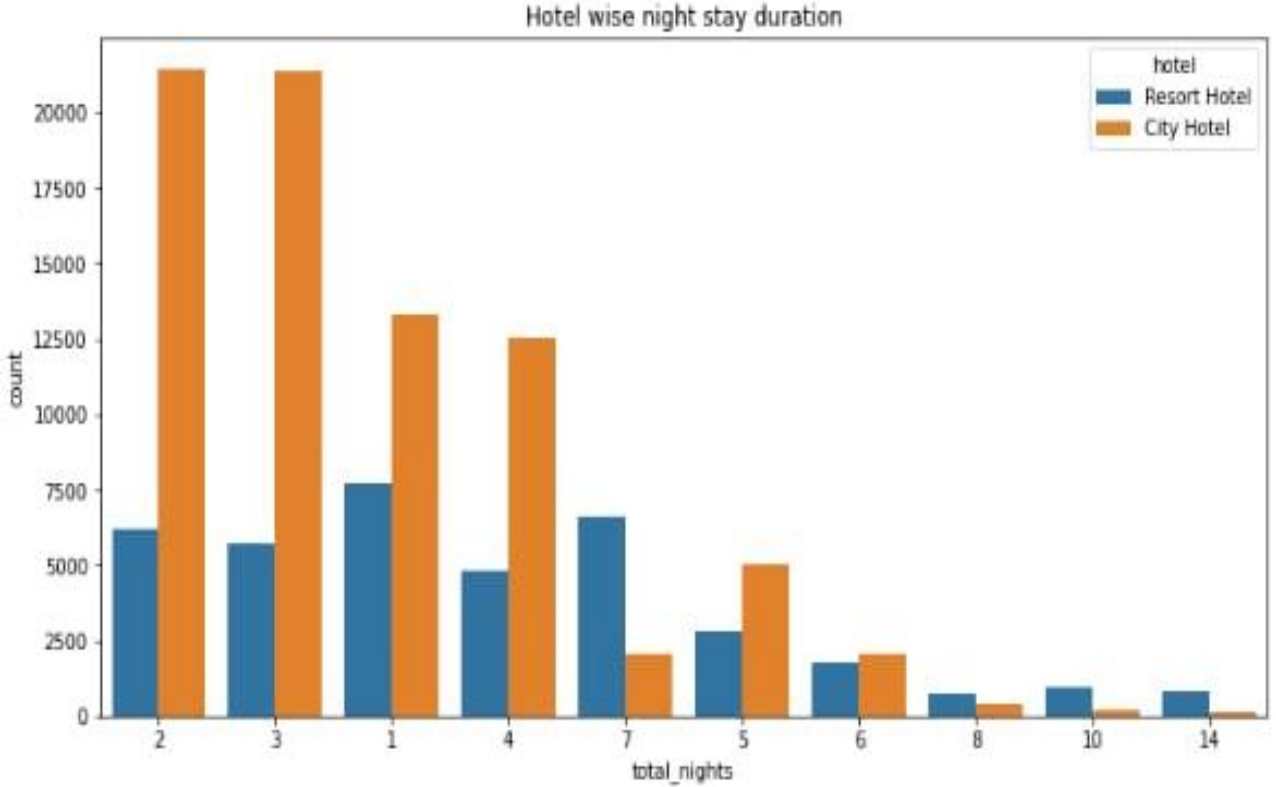
➢ Hotel wise Night Stay Duration

Now we are going to track night stay duration of customers according to Resort hotels and City hotels.

As we already aware that customers loves to stay in City hotels, here also

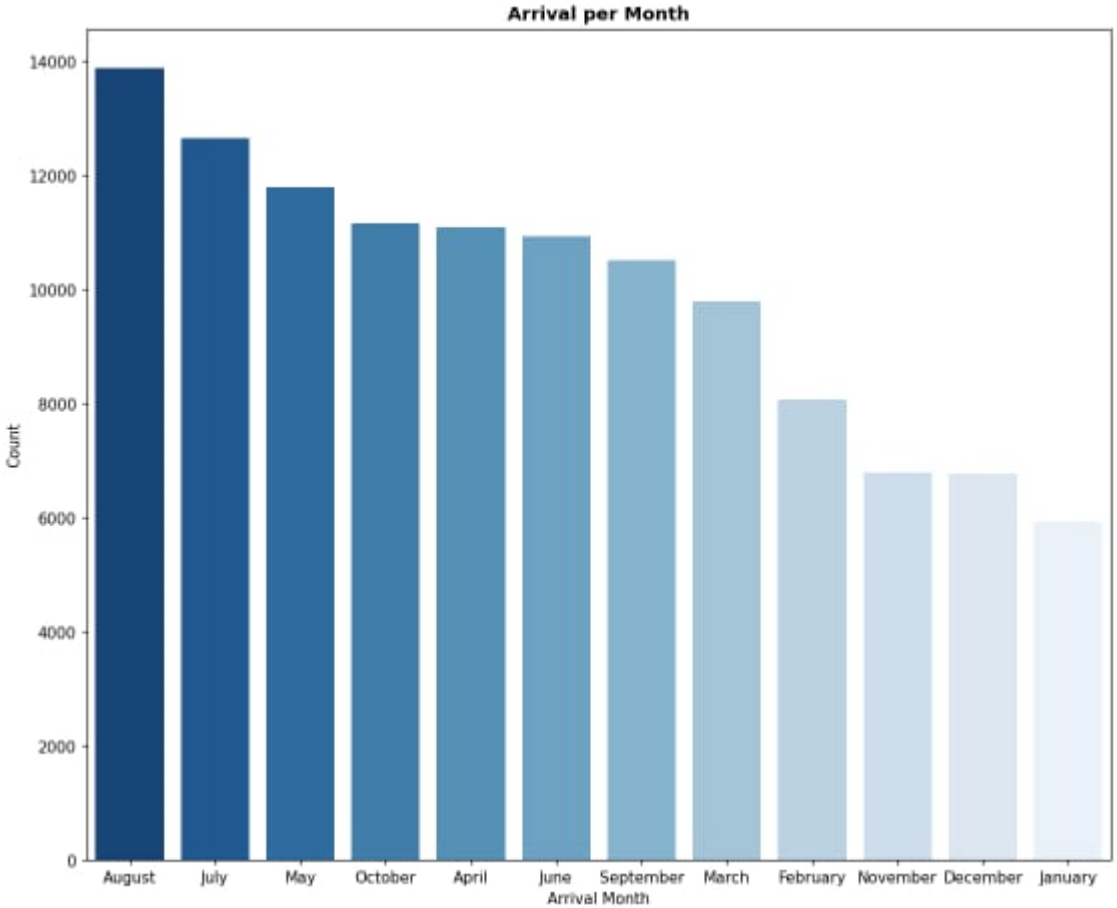
City hotels have large amount of bookings for 2-3 night stay duration and then 1 night stay and 4 night stay customers are there for City hotels.

In Resort hotels, 1 night stay customers are more and then 7 night stay customers comes in focus. Very few customers likely to stay for 8 night or more than it for both type of hotels.



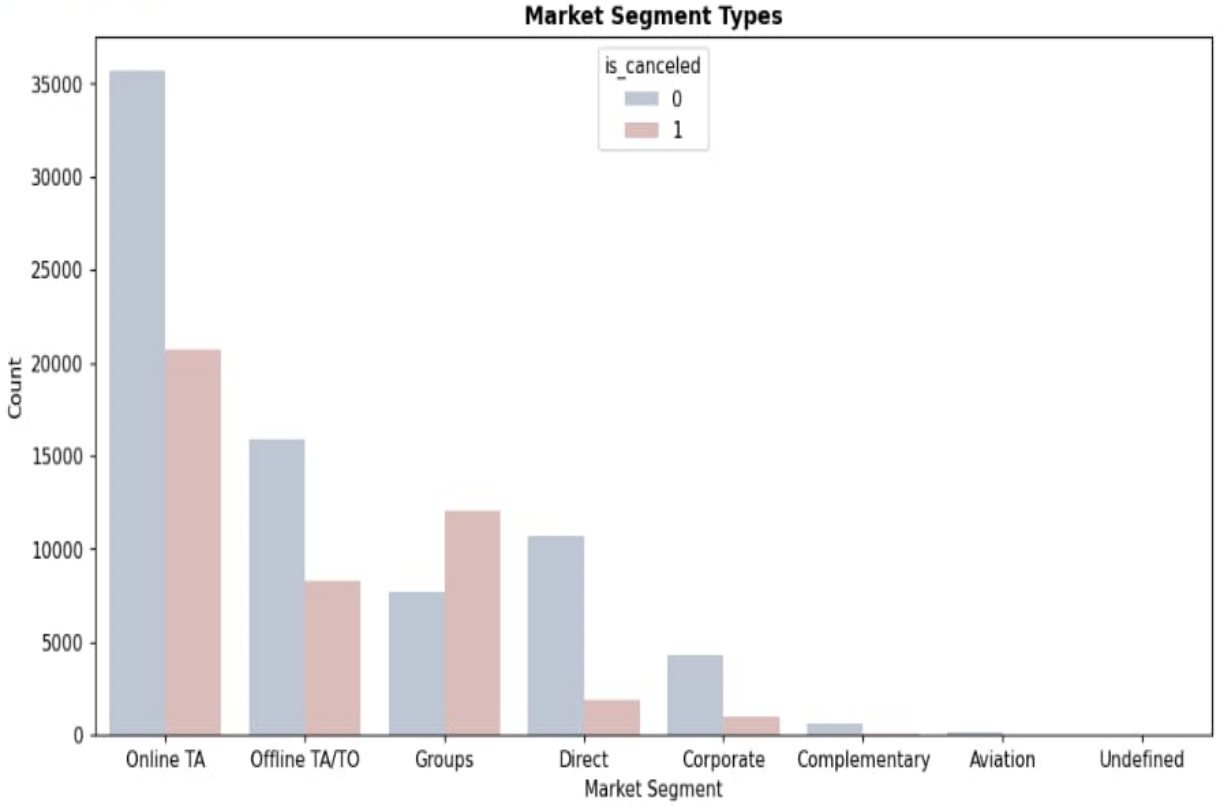
➢ Booking Trend Throught The Year

If we go through booking data along with different months, we found out that August has the highest number of bookings throughout the year then July is at second place where January has the lowest number of bookings i.e., we can assume that January will be the best month for booking to get the best rate on daily basis where booking in month of August will not be economical since it has high demand of room bookings obvious that the cost will also be high.



Total Number Of Bookings Across Various Market Segment

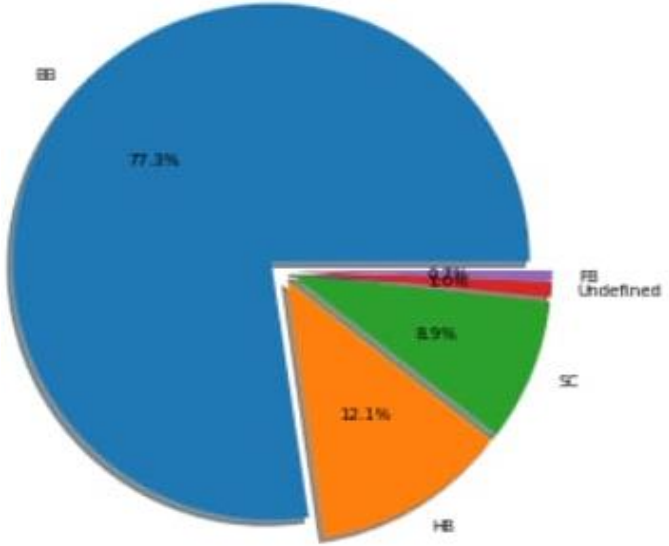
7. Online TA (Travel Agency) segment gives high amount of customers and then Offline TA/TO, Groups, Direct etc. respectively. Complementary, Aviation and Undefined has the lowest amount of customers.
8. So , from this we conclude that We can target our marketing area to be focus on these travel agencies website and work with them since majority of the visitors tend to reach out to them.



➢ Meal Category vs Count Of Booking

- Undefined/SC — no meal package
- BB — Bed & Breakfast
- HB — Half board (breakfast and one other meal — usually dinner)
- FB — Full board (breakfast, lunch and dinner)

- Maximum of the bookings are made with bed and breakfast .So, BB type of meal category is the most preferable in all type of customers, where negligible bookings are made with FB type of meal.



➢ Booking vs Customer Type

Contract — when the booking has an allotment or other type of contract associated to it

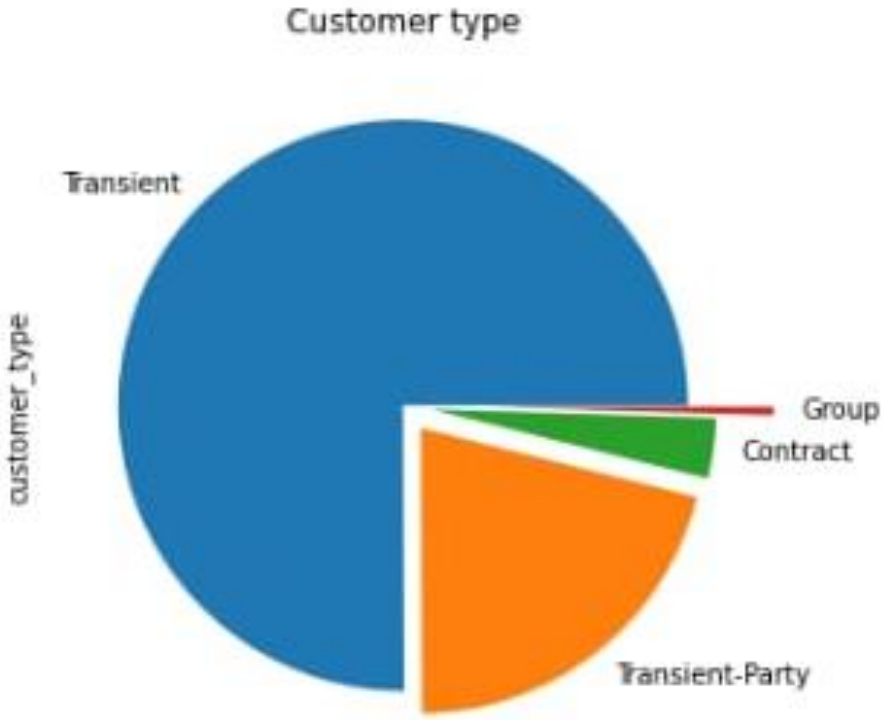
Group — when the booking is associated to a group

Transient — when the booking is not part of a group or contract, and is not associated to other transient booking

Transient-party — when the booking is transient, but is associated to at least other transient booking

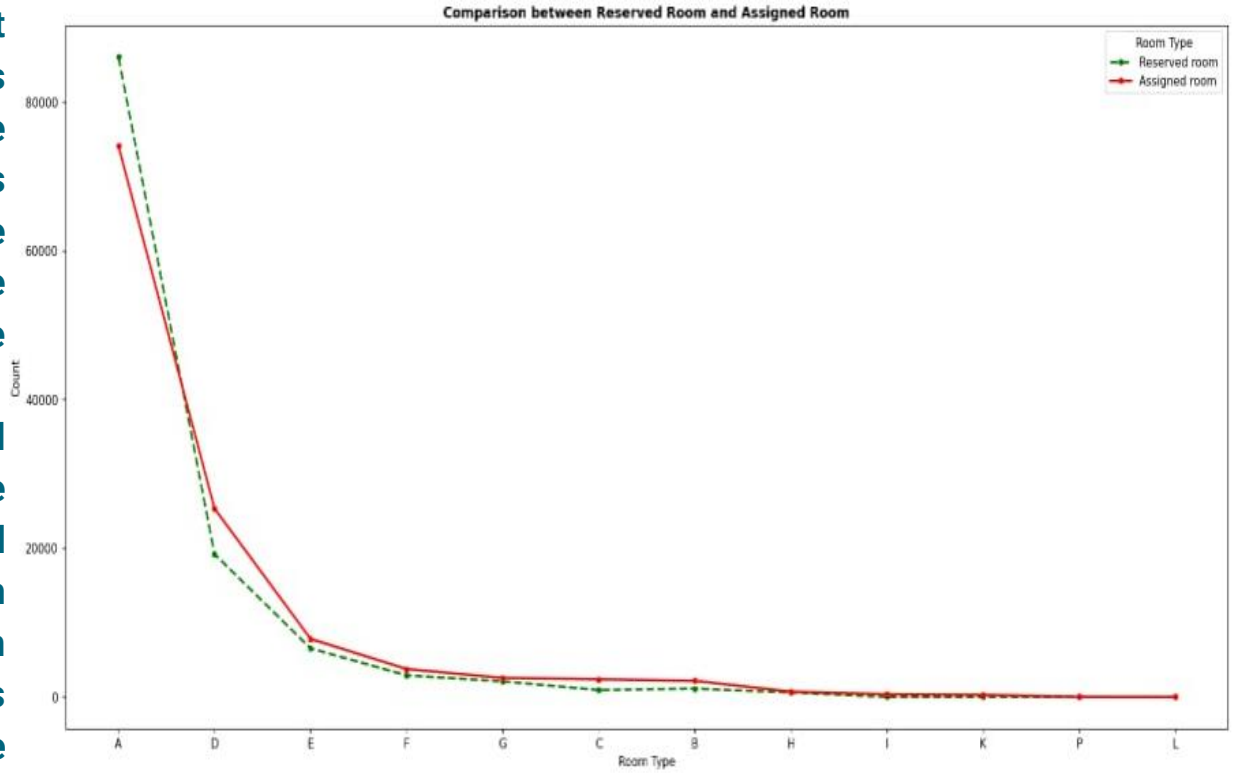
This means that the booking is not part of a group or contract. With the ease of booking directly from the website, most people tend to skip the middleman to ensure quick response from their booking.

Transient type of customer is the main source of booking because 75% of booking coming from this side after that Transient-Party, Contract and Group are coming in the focus.



➢ Booking Trend With Respect To Room Type

7. A - type of room is the most favorite in all types of customers covering all the market about more than 85% ,the D – type of room is at second place in que while negligible customers are there which are ready to stay in L - type and P – type of room.
8. So we need to upgrade L- type and P – type of room to attract more customers so that no one should be in waiting list and do not search any other hotel which results in increasing the profit of hotels as more customers will book the rooms in the hotels.



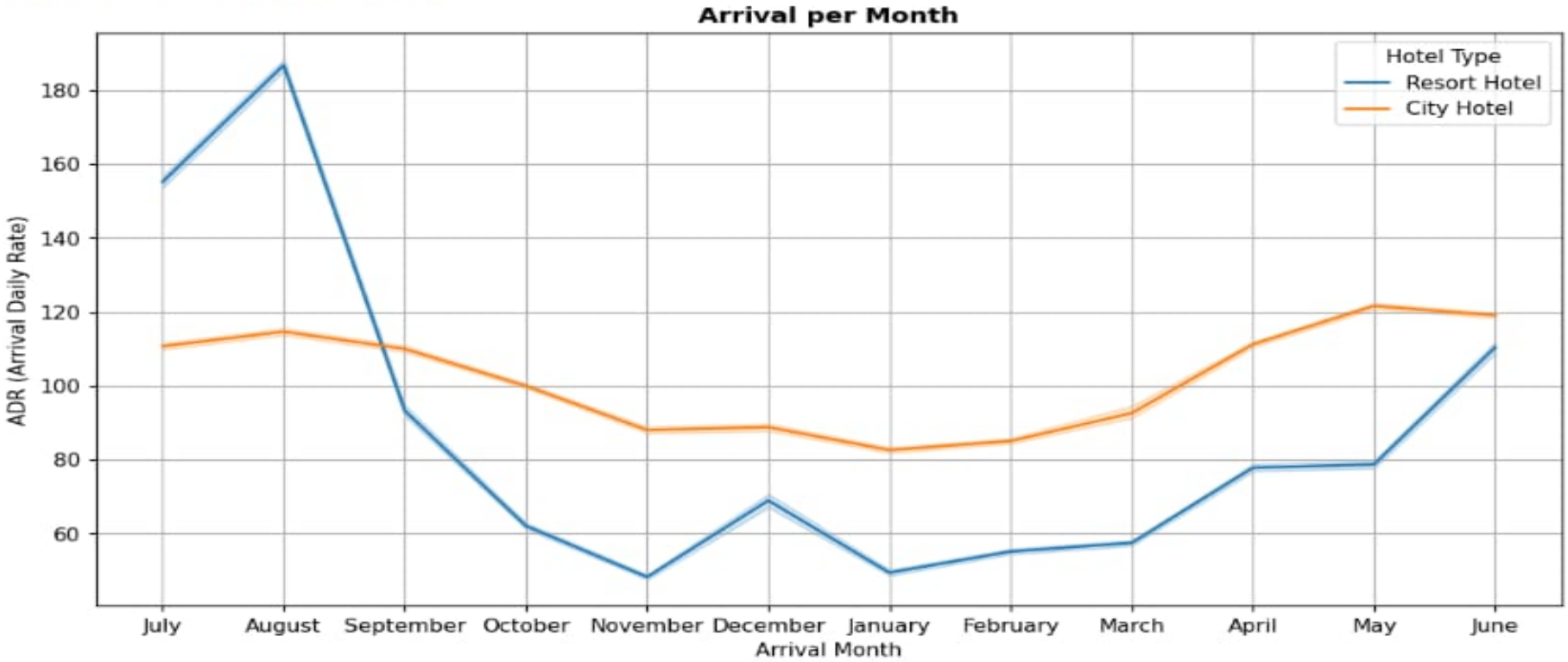
➢ Variation In ADR Across Different Months

```
# Grouping the arrival accoridng to the month and finding the mean of ADR
df.groupby(['arrival_date_month','hotel'])['adr'].mean().unstack()
```

arrival_date_month	City Hotel	Resort Hotel
April	111.251838	77.849496
August	114.680455	186.790574
December	88.826307	68.984230
February	85.088278	55.171930
January	82.628986	49.461883
July	110.734292	155.181299
June	119.074341	110.444749
March	92.643116	57.520147
May	121.638560	78.758134
November	88.069601	48.273993
October	99.974498	62.097617
September	110.004661	93.252030

- For resort hotels, the Average Daily Rate (ADR) is more expensive during August, July, June and September where it is lower for January and November.
- For city hotels, the Average Daily Rate (ADR) is more expensive during August, July, May and June where it is lower also for January and November.

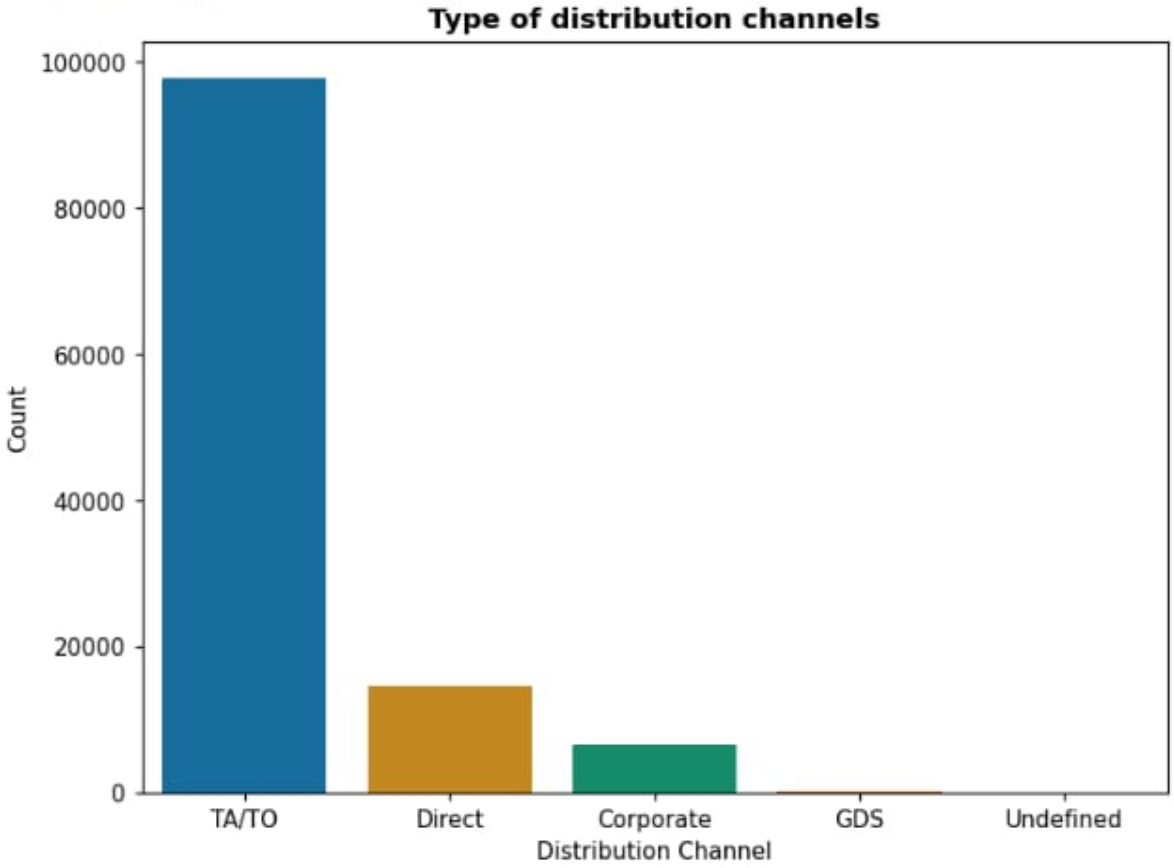
➢ Variation In ADR Across Different Months



So overall Average Daily Rate of both city hotels and resort hotels are more expensive between May and September.

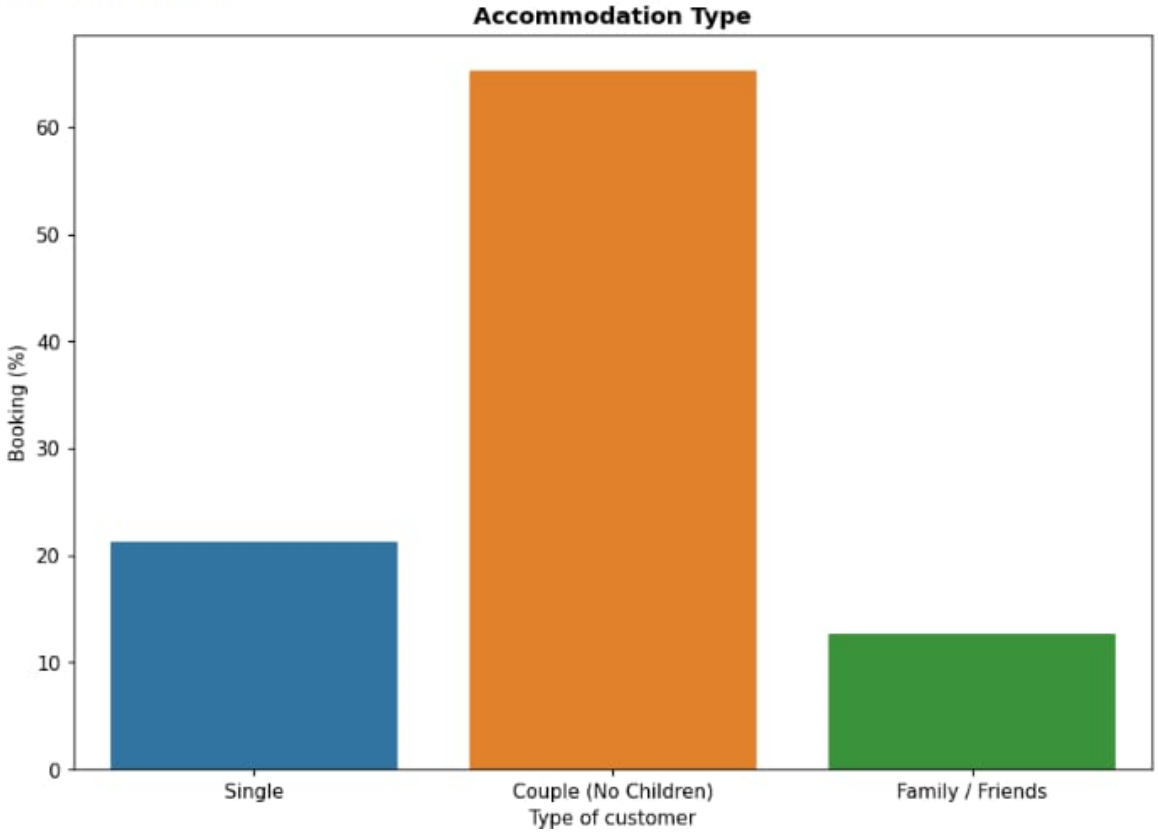
➢ Distribution Channel

- From the side graph we observed that most of the customer preferred to book the hotels through TA/TO (Travel agent / Tour operators).

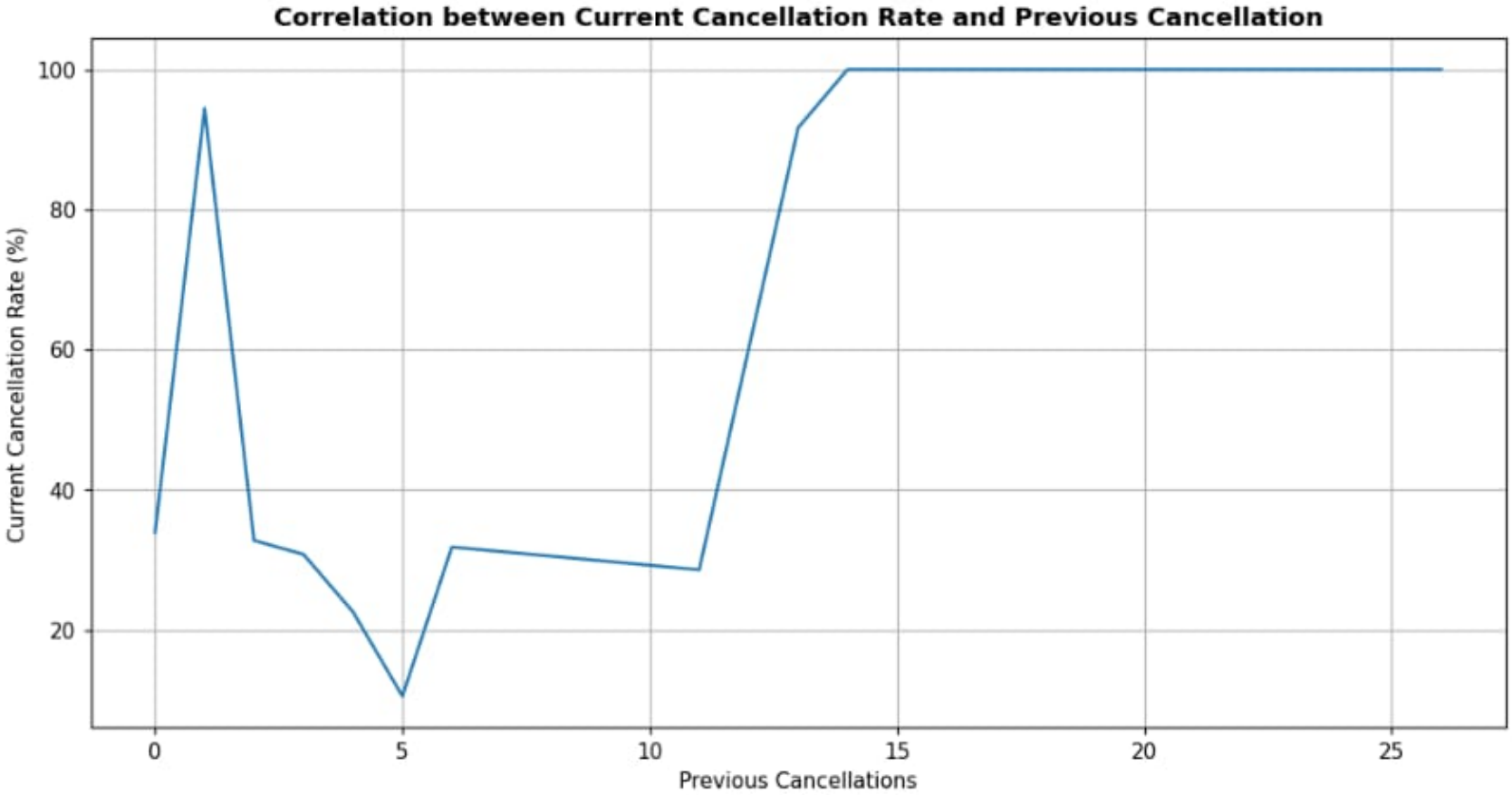


➢ Accommodation Type – Single, Couple & Family

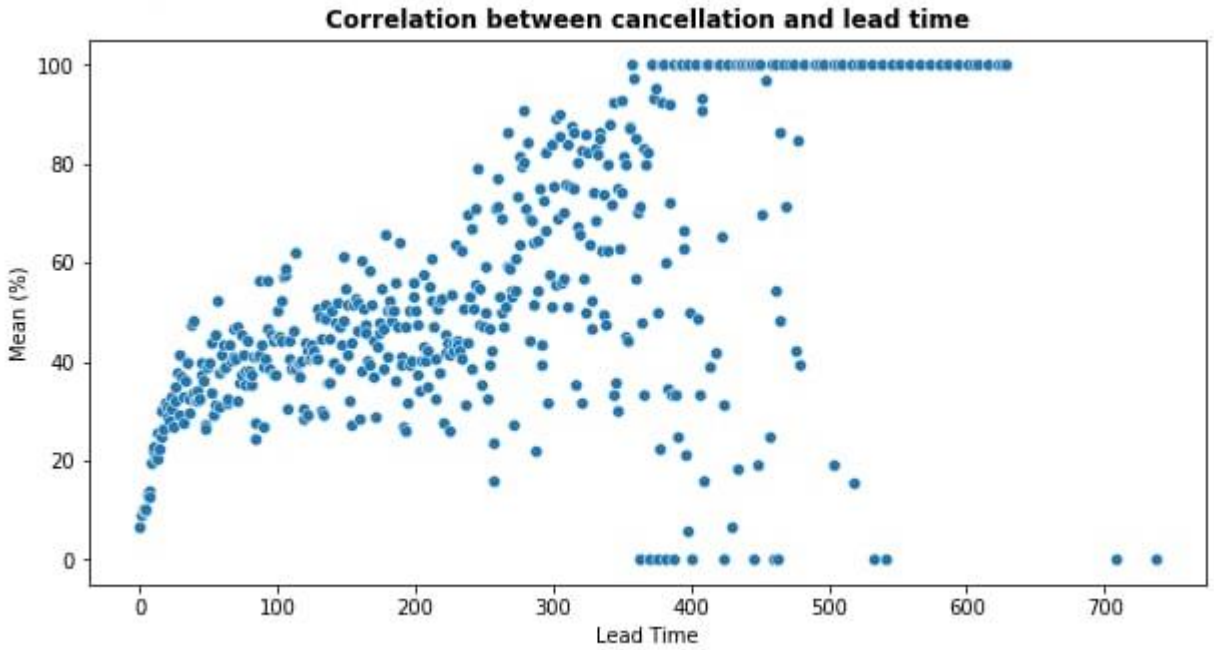
- From graph, it is seen that most number of customer are containing 2 adults customer.
- And those who have more than 2 either containing adults, children & babies have the lowest number of customer.



➢ Correlation between previous and current cancellation

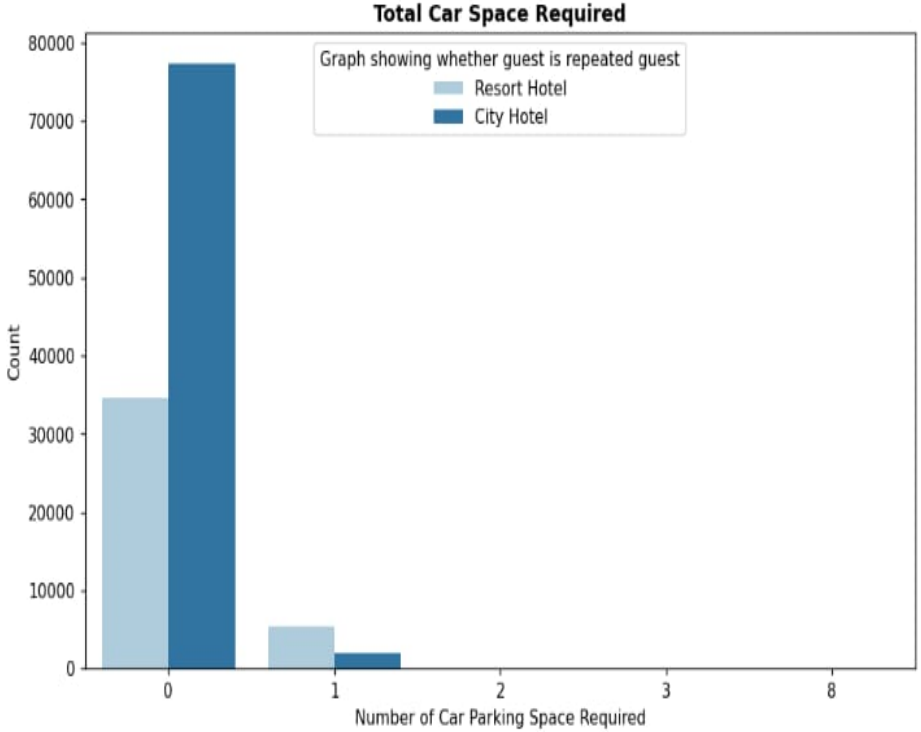


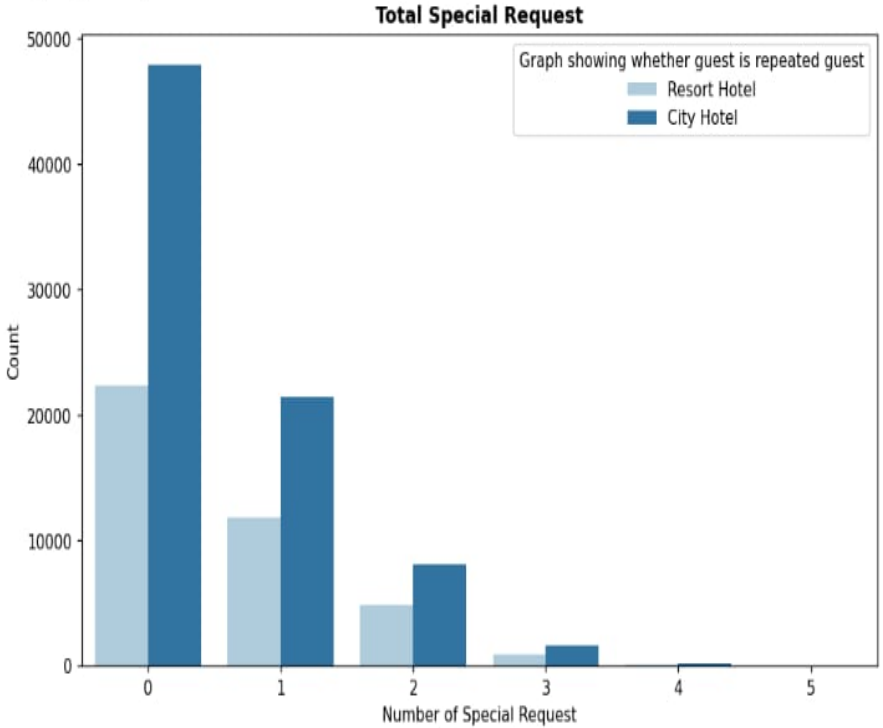
➢ Correlation between cancellation and lead time



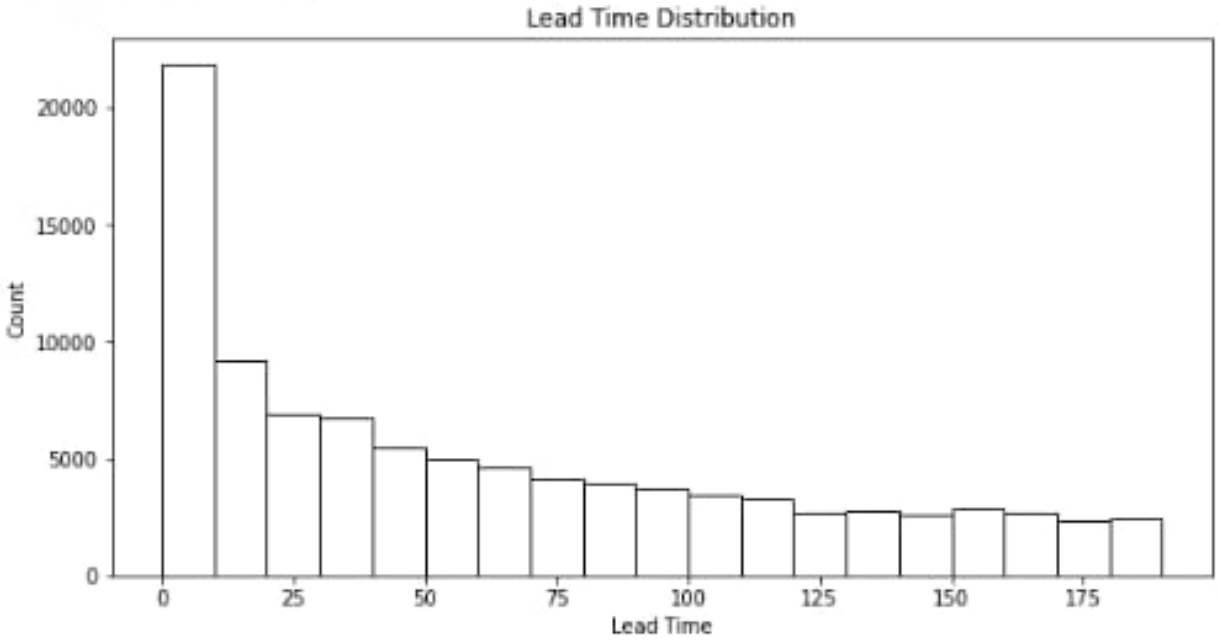
- When lead time increases lead time also increases.
- Positive Correlation between cancellation and lead time.

Number of Parking Spaces and Special Requests made by customer





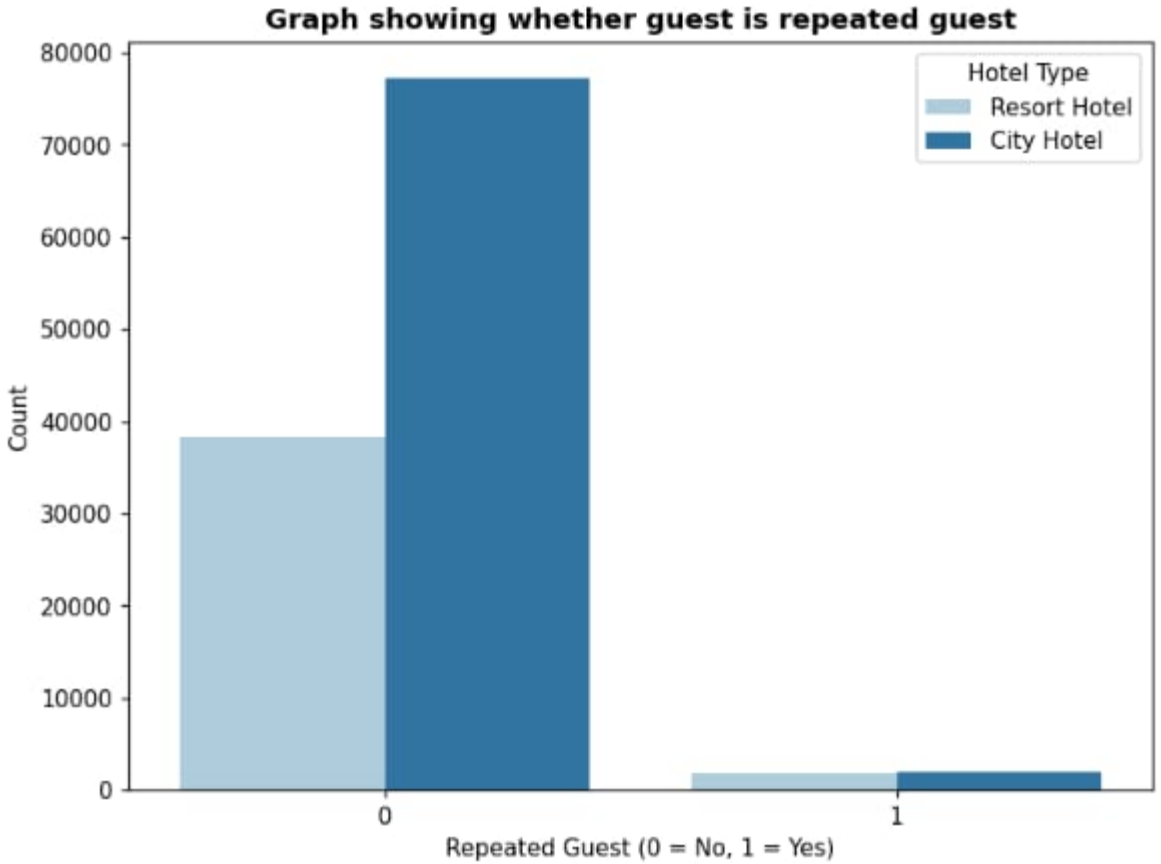
Lead Time Distribution



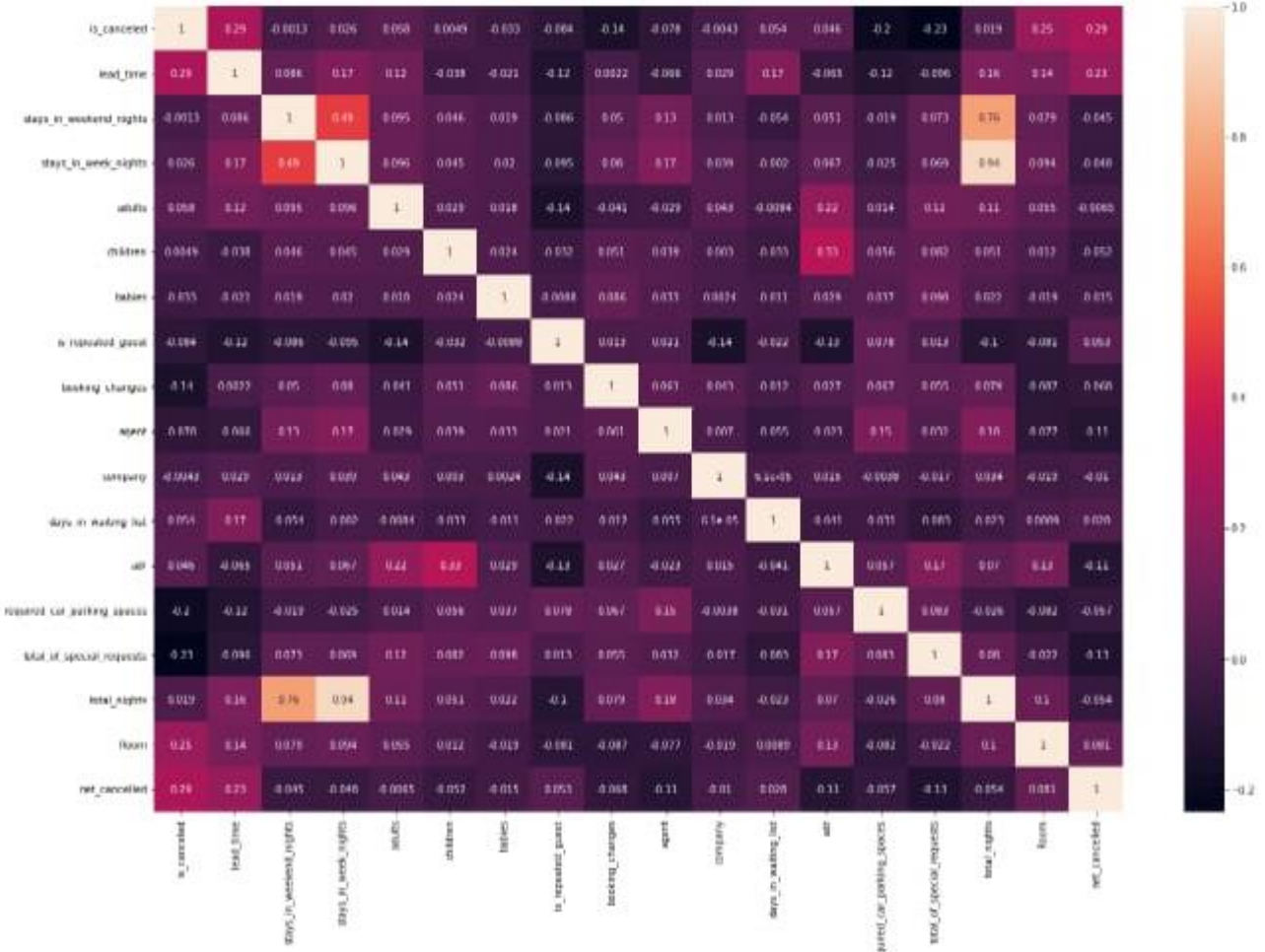
- Bins of this map is from (0,200,10).

➢ Regular Customer base

- This graph shows that only around 3.2% (3806) of customers are regulars.
- And around 96.8% (115096) of customers are new.



Numerical Column Correlation



➢ Unstacking of the Data

```
# We can find the mean of ADR across market segment per day
df.groupby(['arrival_date_day_of_month','market_segment'])['adr'].mean().unstack()
```

market_segment	Aviation	Complementary	Corporate	Direct	Groups	Offline TA/TO	Online TA	Undefined
arrival_date_day_of_month								
1	68.333333	5.068182	62.210629	121.390914	76.229038	85.508725	117.380205	NaN
2	96.000000	0.310345	67.496760	113.017468	74.612910	83.859547	113.175685	NaN
3	112.100000	1.400000	60.360000	107.317268	74.298783	81.890913	114.612456	12.0
4	95.000000	4.767391	68.575123	110.443632	79.997399	80.602023	112.516400	NaN
5	110.250000	0.193548	68.137262	111.406863	71.673768	81.829302	115.707591	18.0
6	95.000000	4.529412	72.581047	107.001595	73.359088	90.965917	115.733331	NaN
7	105.893636	3.391000	66.616568	114.710347	76.641439	86.855759	123.268789	NaN
8	97.769231	0.000000	66.286391	113.967049	70.817007	96.125248	119.144823	NaN
9	105.294118	3.590909	71.280929	106.041872	78.935172	88.048589	117.021528	NaN
10	99.090909	1.568182	64.529281	118.919791	69.414852	80.626094	117.954137	NaN

➢ Conclusion

- Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
- We also realize that the high rate of cancellations can be due high no deposit policies.
- We should also target months between May to Aug because these are peak months.
- Majority of the guests are from Western Europe. We should spend a significant amount of our budget on those area
- Given that we do not have more repeated guests, we should target our advertisement on guests to increase returning guests.

➢ Challenges

1). Huge chunk of data was to be handled by keeping in mind not to miss anything which is even of little relevance.

2). Handling with too many null values and replacing it.

References

1) https://pandas.pydata.org/

2) https://matplotlib.org/

3) https://seaborn.pydata.org/

4) Geeks for Geeks

5) Analytics Vidhya

6) Stack overflow.

7) Kaggle.

Thank You!