Capstone Project - EDA

Hotel Booking Analysis

-Abhinav Singh

FLOW OF THE PRESENTATION:

- OBJECTIVE
- DATASET DESCRIPTION
- EXPLORATORY DATA ANALYSIS (EDA)
- CONCLUSION

OBJECTIVE:

- Hotel industry is a fast moving industry and if someone is operating in the hotel business, they also need themselves to be adaptable with this fast moving industry.
- The objective of this project is to analyze the given ‘Hotel Booking Dataset’, in order to see, understand and gain insights from the factors that govern the bookings in the hotel industry.
- The dataset contains data for two hotels, named: ‘City’ hotel and ‘Resort’ hotel, having data variables like date, duration of stay, total guests, market segment, is the booking cancelled, is the customer a repeat customer, deposit type, etc.

So let’s take a look at these variables in dataset description.

DATASET DESCRIPTION:

hotel :Resort Hotel or City Hotel
is_canceled : ‘1’ if booking is cancelled and ‘0’ if it is not cancelled.
lead_time: Days between confirmed booking status and the day customer is scheduled to arrive at the hotel.
arrival_date_year : Year of arrival
arrival_date_month : Month of arrival
arrival_date_week_number : Week number of year
arrival_date_day_of_month : Day of arrival.

stays_in_weekend_nights : Nights stayed in weekends.
stays_in_week_nights : Nights stayed in week days.
adults : Number of adults in the booking
children : Number of children in the booking
babies : Number of babies in the booking
meal : Type of meal booked by the customer.
country : Country of origin.
market_segment : Market segment the customer belongs to.
distribution_channel : Distribution channel through which the booking came in.
is_repeated_guest : ‘1’ if the customer is a repeated guest, ‘0’ if not.
previous_cancellations : Number of previous bookings cancelled by the customer before the current booking.
previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer before the current booking.
reserved_room_type : Type of room reserved at the time of booking.
assigned_room_type : Actual room allotted by the hotel.
booking_changes : Number of changes made to the booking before check-in.
deposit_type : No Deposit, Non Refund , Refundable.
agent : ID of the travel agent that made the booking
company : ID of the travel company that made the booking
days_in_waiting_list : No. of days the booking was in the waiting list before confirmation.
customer_type : Type of customer: Contract,Group,Transient,Transient party.

adr : Average Daily Rate is defined as dividing the sum of all lodging transactions by the total duration of stay.
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.
reservation_status : Reservation last status- ‘Check-out’, ‘Canceled’, ‘No-Show’.

These were all 31 variables in our dataset and the total number of records in our dataset are 1,19,390.

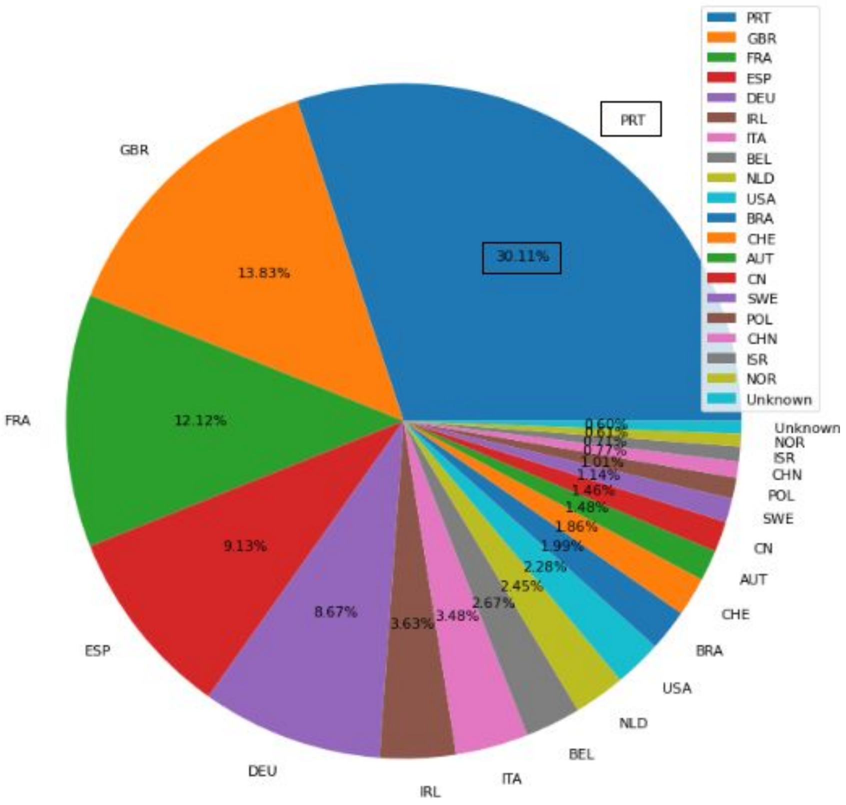
EDA:

Out of the 1,19,390 records,
# ‘Confirmed’ Bookings: 75,166
# ‘Cancelled’ Bookings: 44,224

The time duration (year-month) of our dataset is from: 2015-07 to 2017-08.

Both ‘confirmed’ bookings and ‘cancelled’ bookings will be looked upon.

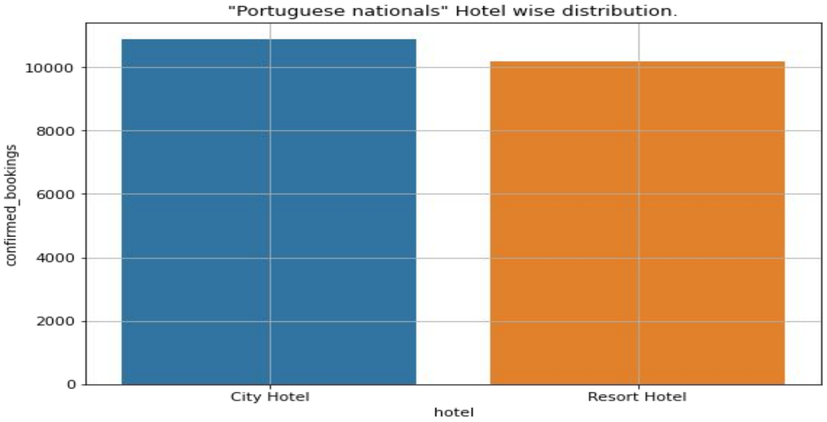
1) Where are the customers coming from?

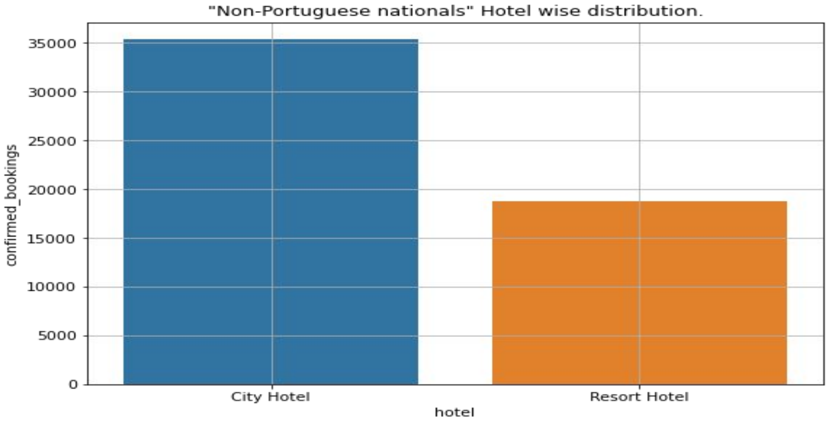

TOP 20 COUNTRIES WITH CONFIRMED BOOKINGS (CITY HOTEL + RESORT HOTEL).
Top 20 Overall confirmed bookings: 69984

- From the pie chart of the ‘TOP 20 COUNTRIES WITH CONFIRMED BOOKINGS’ for both ‘CITY’ & ‘RESORT’ hotel, we see that majority of the of the customers belong to Portugal (PRT): 30.11% (21071).

- The next closest to Portugal are Britain (GBR): ~14% (9676) and France (FRA): ~12% (8481).

- Since the gap is huge between Portugal and Britain, it **confirms** that the hotels ‘CITY’ and ‘RESORT’ are located in Portugal & it is the Portuguese nationals who are driving the business majorly.





City Hotel PORTUGUESE nationals RETENTION RATE is: 12.28 %
Resort Hotel PORTUGUESE nationals RETENTION RATE is: 13.62 %

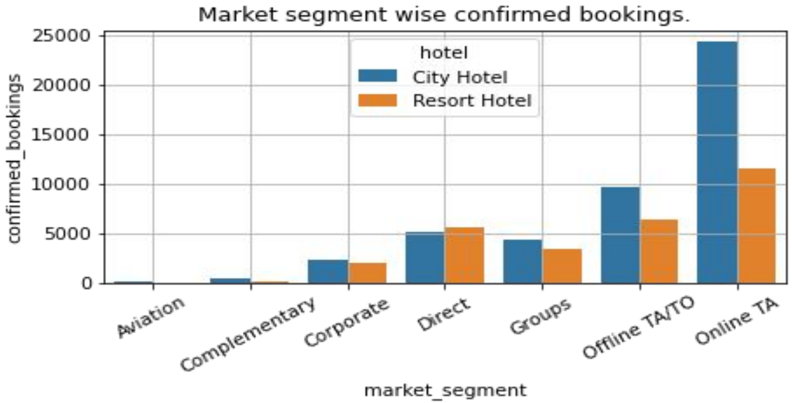
City Hotel non-PORTUGUESE nationals RETENTION RATE is: 0.72 %
Resort Hotel non-PORTUGUESE nationals RETENTION RATE is: 1.49 %

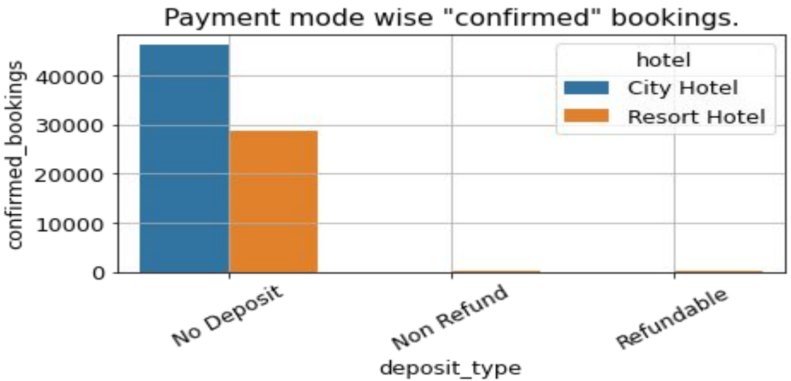
—--

<u>From ' "Portuguese nationals" Hotel wise distribution ' & ' "Non-Portuguese nationals" Hotel wise distribution ' bar plot, we observe</u>:

- The <u>preference</u> of <u>Portuguese nationals</u> is <u>similar</u> for both 'City' and 'Resort' hotel as compared to Non-Portuguese, with ~11000 for 'City' and slightly above 10000 for 'Resort'.
- But in the case of <u>Non-Portuguese nationals</u>, <u>'City Hotel' has catered</u> to almost <u>double</u> the people of what 'Resort Hotel' has (~35000 & ~19000 respectively) & overall also we can conclude that ‘City Hotel’ is the busier of the two.
- Looking at the retention rate in both Portuguese & non-Portuguese cases, <u>'Resort Hotel' has a better retention rate</u>.
- 'Resort hotel' could market itself more among the non-Portuguese customers.

2) Market segment wise 'confirmed' bookings & Payment mode.

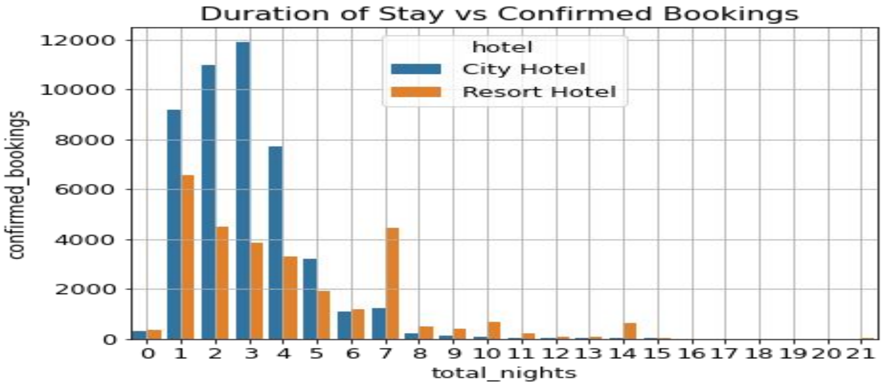


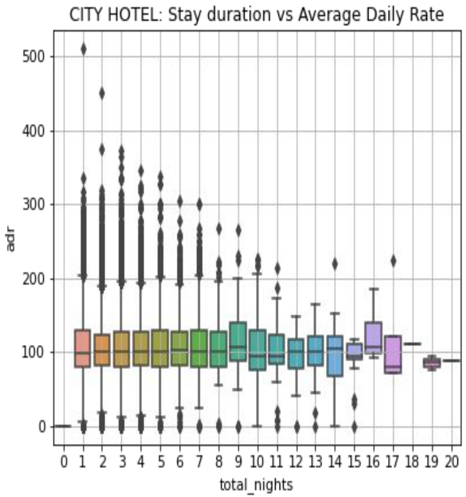


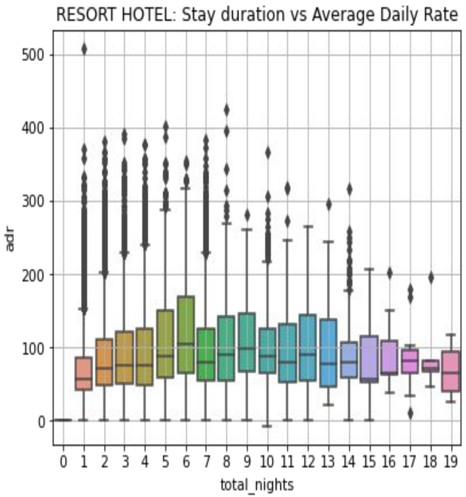
- The 'Online Travel Agency(TA)' market segment is bringing most of the business to the hotels.
- 'Offline TA/TO', 'Groups', 'Direct' & 'Corporate' market segments are also bringing business, but 'Online TA' is way ahead, which is normal in the age of internet.
- 'Aviation' & 'Complementary' are on the lower side in terms of bringing the business.

- In the terms of payment mode preferred by customers, almost all customers prefer 'No Deposit' mode of payment, ie they prefer to pay at the time of checking out from the hotel.

3) Duration of Stay & ADR.


Duration of Stay vs Confirmed Bookings


CITY HOTEL: Stay duration vs Average Daily Rate


RESORT HOTEL: Stay duration vs Average Daily Rate

From Duration of stay vs Confirmed Bookings bar plot, we see:

- For 'City Hotel', customers have mostly preferred to stay upto 3 nights, after that it decreases considerably.

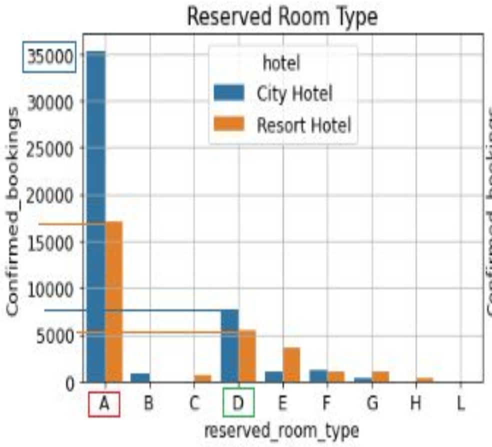
- Whereas for 'Resort Hotel', seeing the confirmed bookings, customers have preferred to stay for 3-5 nights & there are customers who prefer to stay till 7 nights also. This is an indicator that people coming to 'Resort Hotel' are coming for a longer trip.

From Duration of stay vs ADR box-plot we see:

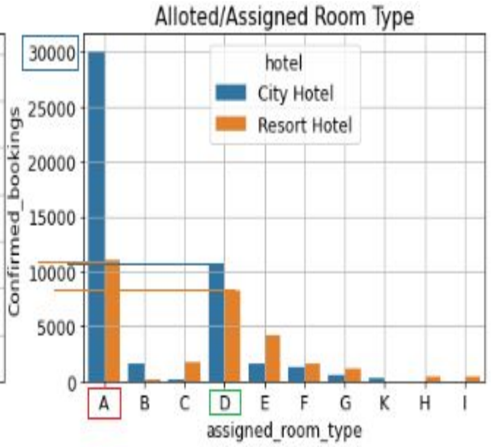
- For 'City Hotel', the variability in ADR is higher when stay duration is less & it decreases when duration increases, but the median ADR is generally around 100, irrespective of the length of stay.

- For 'Resort Hotel', variability is similar to that of 'City Hotel', but the median ADR increases with the duration of stay, till 6th night (100). After that it oscillates but doesn't cross 100. And even till 6th day, median is below 100, hence customers can plan to stay for longer duration for lower ADR.

4) Room Type & ADR.

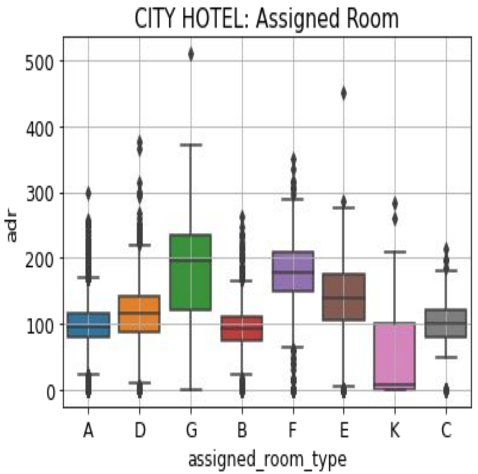


Reserved Room Type

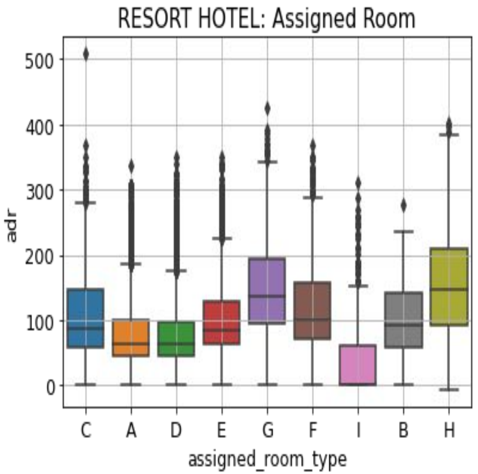


Alloted/Assigned Room Type

- From the 'Reserved' room type bar plot, we can see that there is a very high demand of room type 'A' for both 'City' and 'Resort Hotel', ~35000 and ~17000 respectively.
- But not everyone is getting room type 'A', which we can see from 'Alloted/ Assigned' room type bar plot where, room type 'A' confirmed bookings are ~30000 for 'City Hotel' and ~12000 for 'Resort Hotel'.
- We see for 'City Hotel' assigned room 'D' increases to ~12000 from ~7000 (in assigned room bar plot) and for 'Resort Hotel' room type 'D' bookings increases slightly to ~8000 from ~5000. So those who are not getting room type ‘A’ are majorly getting room type ’D’.



CITY HOTEL: Assigned Room

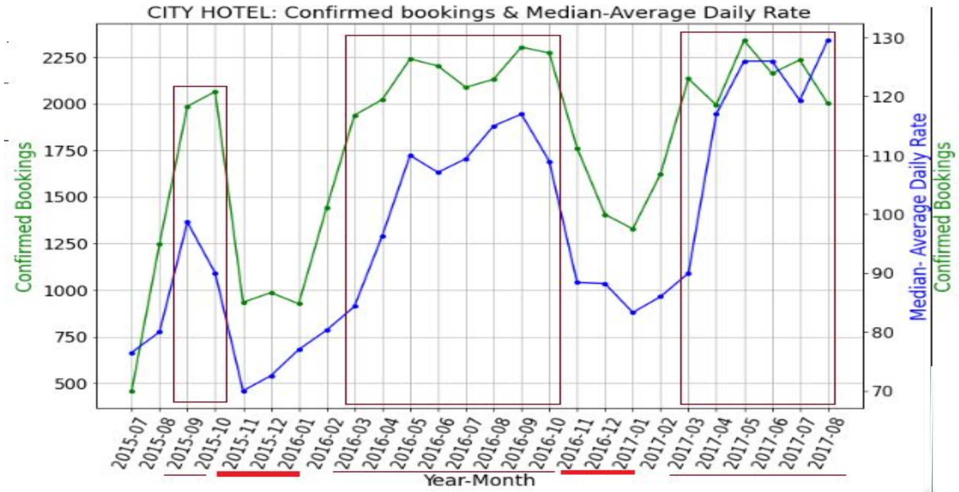


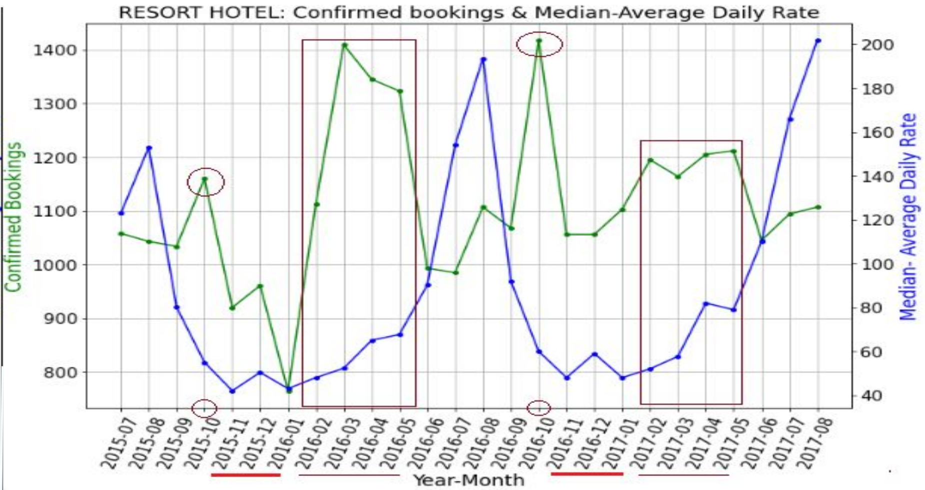
RESORT HOTEL: Assigned Room

Observations from 'assigned_room_type' vs 'adr' box-plot:

- For 'CITY HOTEL', from box plot we see, the median ‘ADR’ of room 'D' is slightly higher than room 'A'. Room 'G','F','E' have considerably higher median ‘ADR’, which is an indicator of those being high end rooms.
- For 'RESORT HOTEL' , the median ADR of room 'D' is similar to that of room 'A'. Room 'G' & 'H' have considerably higher median ADR, which is an indicator of those being high end rooms. Room 'C','E','F','B' lie somewhat between ['A','D'] and ['G','H'].

5) 'Year-Month' wise "Demand" & "Median ADR":





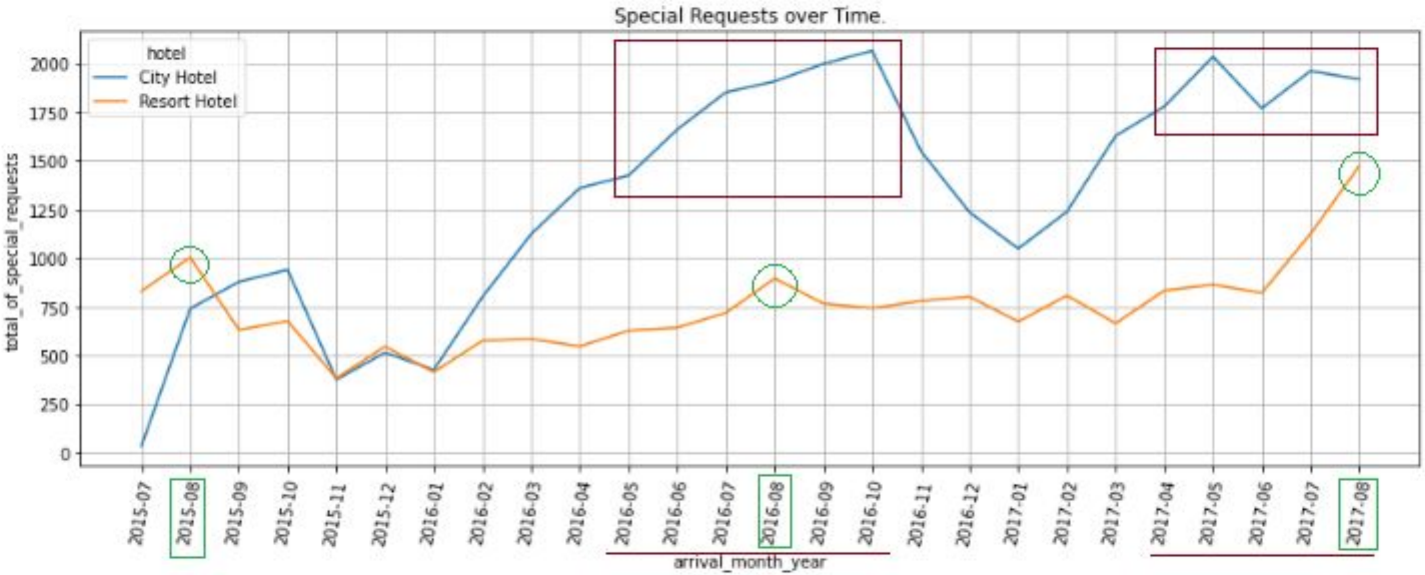
For 'City Hotel' we observe that:

- The peak season is from the month of March to October (brown rectangle) and no of customers decrease during winter season, ie from November to January. Then again it starts going up starting from February (Customers/demand is in green in the line plot)
- The trend of ADR (in blue) over months is similar to that of the demand, ie as the no. of customers starts increasing, the ADR also starts increasing and vice versa (PS- demand and ADR are on different scale).
- From a customer perspective, trip during winter season would be cheaper.

For 'Resort Hotel' we observe that:

- The peak season is from the month of February to May (brown rectangle) & there is a peak in October. No of customers decrease during June to September & November to January.
- The trend of ADR (in blue) over months is opposite to that of the demand, ie as the no. of customers starts increasing, the ADR starts decreasing and vice versa.
- From a customer perspective, trip during the peak season of the hotel would be economical.

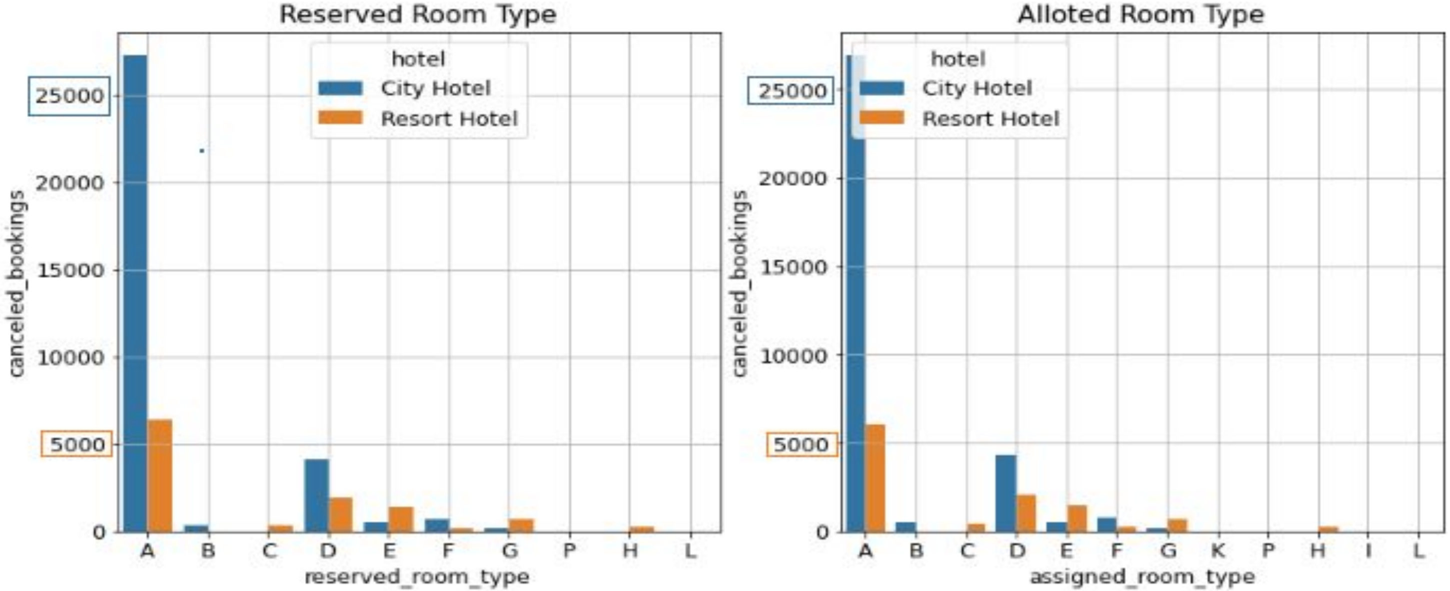
6) 'Year-Month' wise Special Requests:



- For 'City Hotel' we see that the trend of special requests is similar to the trend of its demand, which we saw on previous slide, ie increasing in its peak period (brown rectangle), May to October and then decreasing from the start of winter season from November.
- For 'Resort Hotel' we see that the trend is fairly constant, with dips in winter season, but there are spikes in the month of August in every year from 2015-2017 (green circle).

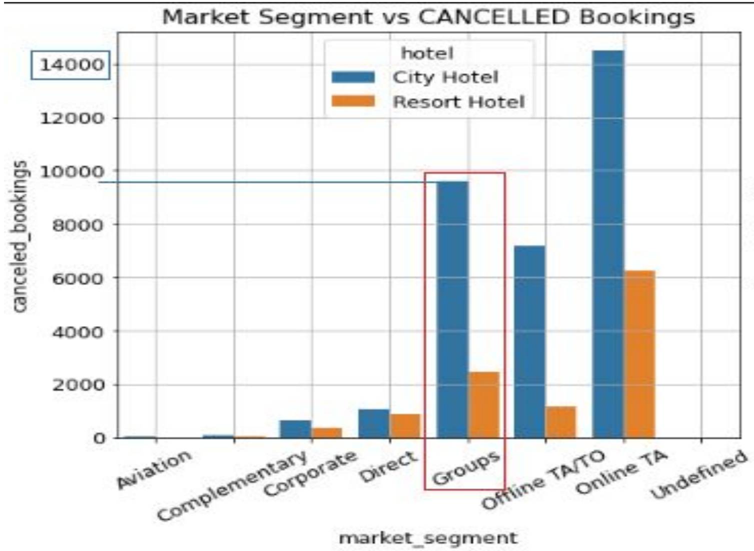
7) Cancelled Bookings.

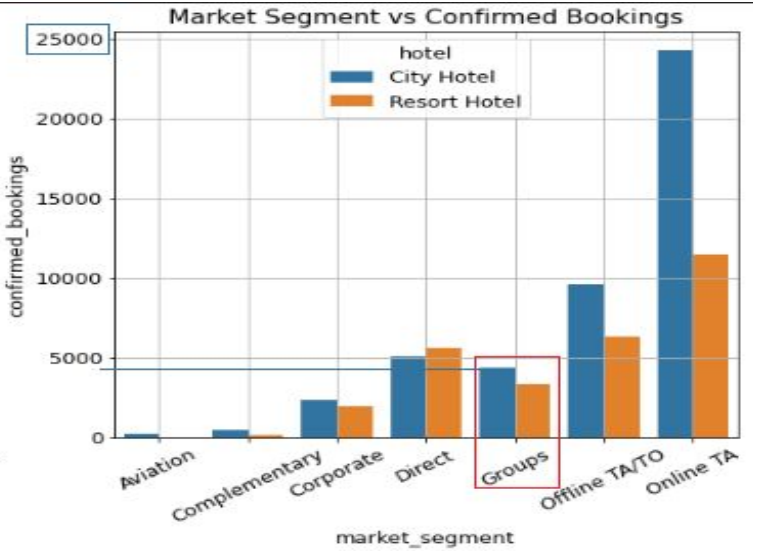
Hypothesis: Not getting the desired/reserved room type is a factor for booking cancellations.
Visual Inspection:



From the bar plot for 'Reserved' and 'Assigned' room type we see that there is almost no change between two plots, hence 'not getting desired room type' does not affect booking cancellations.

So after this, 'market segment' wise booking cancellations were checked.

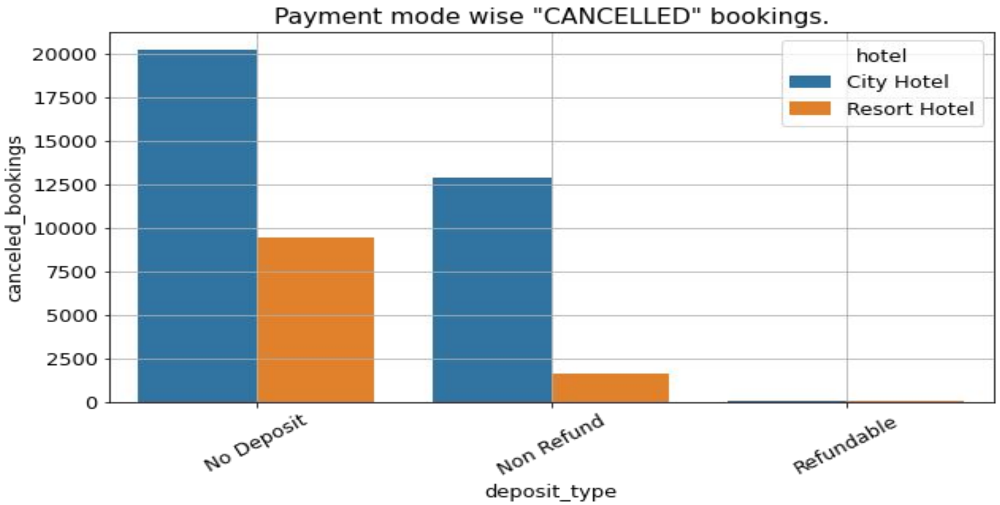




- From above ‘confirmed’ bookings barplot (on the right), we see that 'Online' & 'Offline TA/TO' market segment brings maximum customers so it is understandable that more 'CANCELLED' bookings will also be from these segments, but less than booked ones (bar plot on the left).

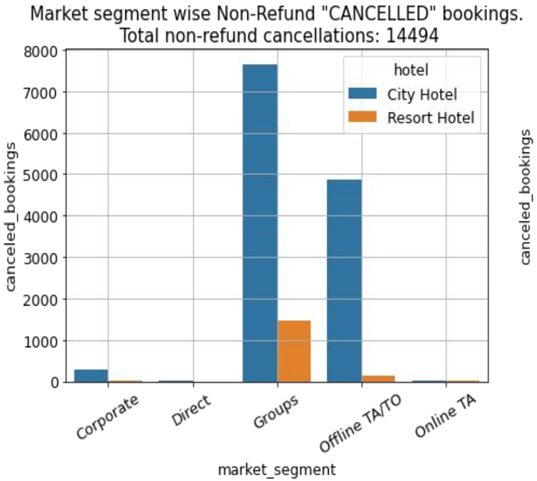
- On a total contrast, from the CANCELLED bookings bar plot (on the left),in the 'Group' market segment, for 'City Hotel', we see that more people have CANCELLED their bookings (~10000) than 'confirmed' (~4000, barplot on the right), which is surprising.

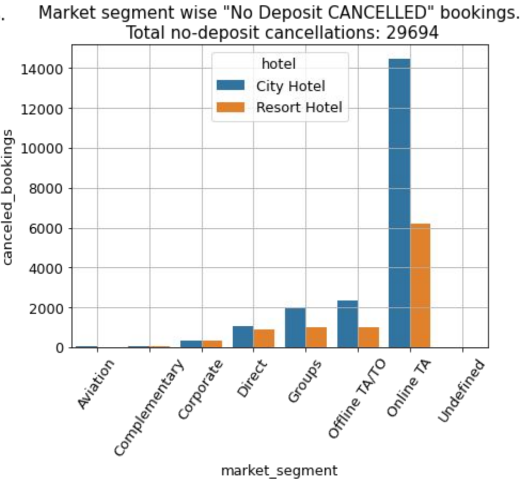
After this, ‘Payment mode’ wise cancelled bookings were checked.



In the ‘Payment mode’ wise ‘cancelled’ bookings we see that,

- For ‘City Hotel’, there are ~20000 cancelled bookings in ‘No Deposit’, which makes sense because customer isn’t losing money, but surprisingly there are ~12500 customers who had paid under ‘Non refund’ mode and then cancelled the booking, which is surprising, because customer is losing a lot of money!
- For ‘Resort Hotel’, there are ~9000 ‘cancelled’ bookings in ‘No Deposit’ and ~1500 cancelled bookings in ‘Non Refund’, which is not as high as ‘City Hotel’.





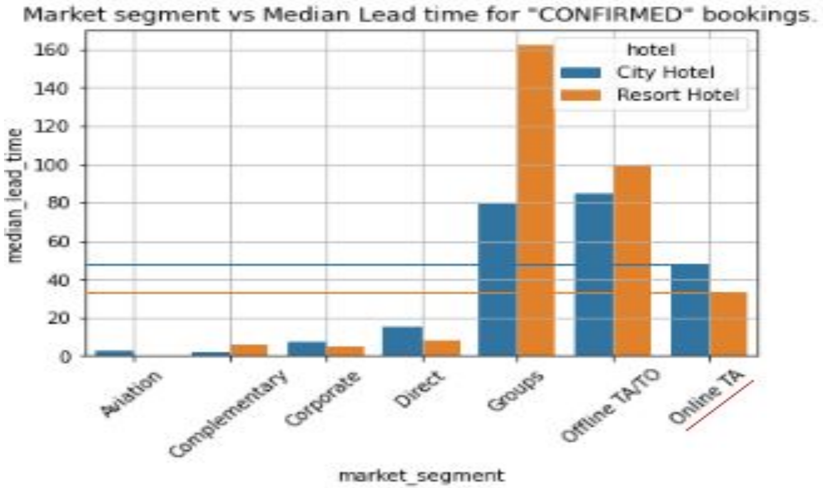
Earlier we saw lots of cancellations in ‘Groups’ market segment and now in ‘Non refund’ payment mode. Is there a relation? Yes!

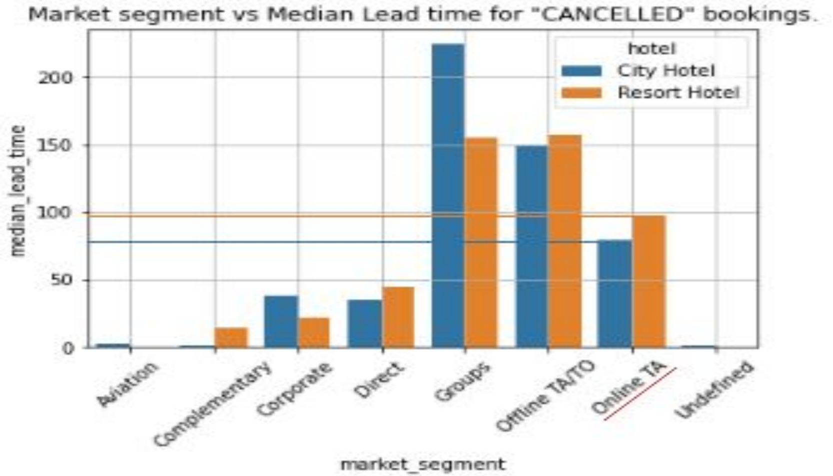
- In the ‘NON REFUND’ cancellations (bottom left bar plot on the left), it is observed that, for ‘City’ Hotel, majority of the cancellations are happening from ‘Groups’ and as it turn out, the ‘Offline TA/TO’ market segment also, but for ‘Resort’ hotel it is majorly from ‘Groups’.

- So it is now also clear that ‘NO DEPOSIT’ cancellations are majorly from ‘Online TA’ market segment for both hotels (bottom right bar plot on the left).

- On the previous slide we saw lots of cancellations in ‘Non Refund’ payment mode (Groups & Offline), but hotels need to worry less about it, since they are getting paid in advance and they don’t have repay the cancelling customer.
- What they need to worry about are the ‘No Deposit’ cancellations, which are majorly coming from ‘Online TA’ market segment for both hotels, because they are not getting paid any money and there comes another challenge to fill those large vacant bookings via new customers.

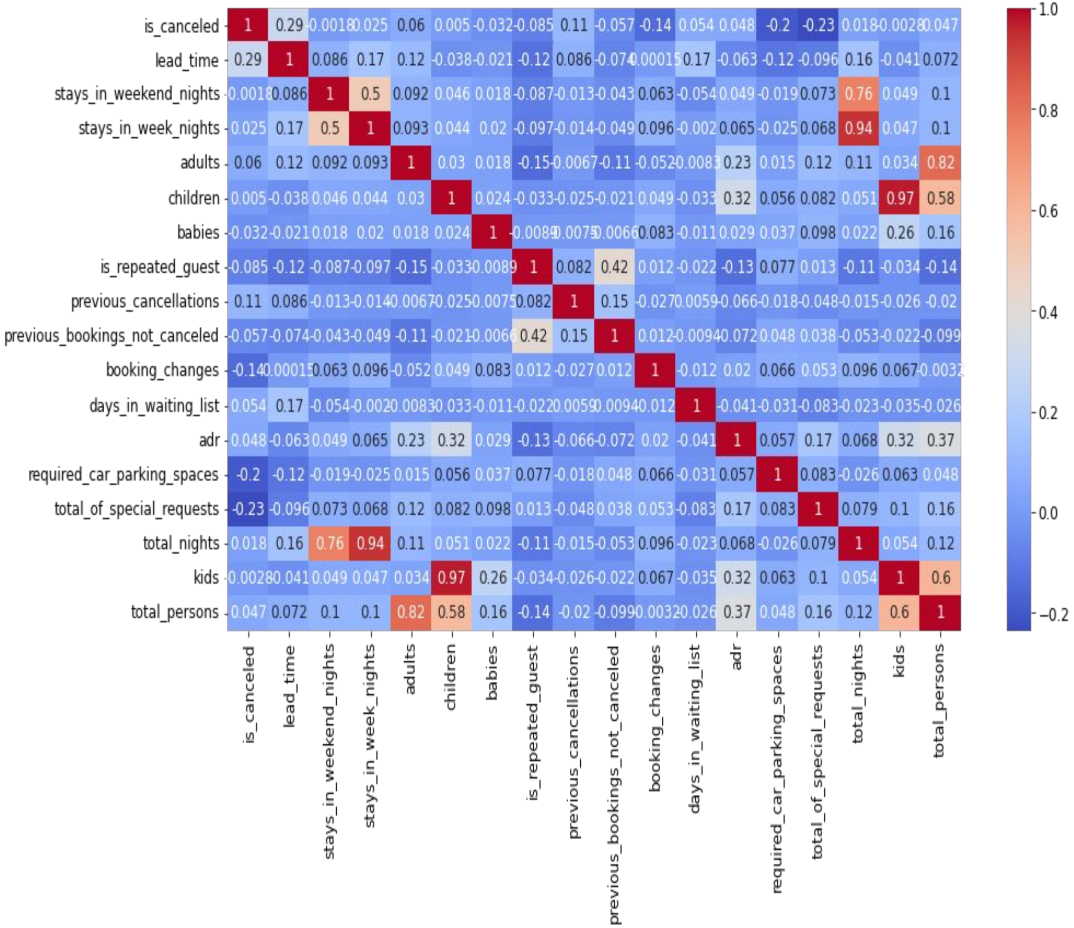
Here we are looking at ‘Market segment’ wise median Lead Time for both “Confirmed” (left) & “Cancelled” (right) bookings:


Market segment vs Median Lead time for "CONFIRMED" bookings.


Market segment vs Median Lead time for "CANCELLED" bookings.

- Visually we can see that, generally for each market segment, the “cancelled” median Lead Time is higher than “confirmed”.
- If we see for ‘Online TA’ market segment (from where most “No Deposit” cancellations are coming), in “confirmed” bookings, the median Lead Time for ‘City Hotel’ is ~50 days (under 2 months) & for ‘Resort Hotel’ it is ~30 days (1 month). In “cancelled” bookings, the median Lead Time for ‘City hotel’ is ~80 days (> 2 months) & for ‘Resort Hotel’ is ~100 days (> 3 months).

8) Correlation Heatmap:



	is_canceled	lead_time	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_nights	kids	total_persons
is_canceled	1	0.29	-0.0018	0.025	0.06	0.005	-0.032	-0.085	0.11	-0.057	-0.14	0.054	0.048	-0.2	-0.23	0.018	-0.0028	0.047
lead_time	0.29	1	0.086	0.17	0.12	-0.038	-0.021	-0.12	0.086	-0.074	0.00015	0.17	-0.063	-0.12	-0.096	0.16	-0.041	0.072
stays_in_weekend_nights	-0.0018	0.086	1	0.5	0.092	0.046	0.018	-0.087	-0.013	-0.043	0.063	-0.054	0.049	-0.019	0.073	0.76	0.049	0.1
stays_in_week_nights	0.025	0.17	0.5	1	0.093	0.044	0.02	-0.097	-0.014	-0.049	0.096	-0.002	0.065	-0.025	0.068	0.94	0.047	0.1
adults	0.06	0.12	0.092	0.093	1	0.03	0.018	-0.15	-0.0067	-0.11	-0.052	-0.0083	0.23	0.015	0.12	0.11	0.034	0.82
children	0.005	-0.038	0.046	0.044	0.03	1	0.024	-0.033	-0.025	-0.021	0.049	-0.033	0.32	0.056	0.082	0.051	0.97	0.58
babies	-0.032	-0.021	0.018	0.02	0.018	0.024	1	-0.0089	-0.0075	-0.0066	0.083	-0.011	0.029	0.037	0.098	0.022	0.26	0.16
is_repeated_guest	-0.085	-0.12	-0.087	-0.097	-0.15	-0.033	-0.0089	1	0.082	0.42	0.012	-0.022	-0.13	0.077	0.013	-0.11	-0.034	-0.14
previous_cancellations	0.11	0.086	-0.013	-0.014	-0.0067	-0.025	-0.0075	0.082	1	0.15	-0.027	0.0059	-0.066	-0.018	-0.048	-0.015	-0.026	-0.02
previous_bookings_not_canceled	-0.057	-0.074	-0.043	-0.049	-0.11	-0.021	-0.0066	0.42	0.15	1	0.012	-0.0094	-0.072	0.048	0.038	-0.053	-0.022	-0.099
booking_changes	-0.14	0.00015	0.063	0.096	-0.052	0.049	0.083	0.012	-0.027	0.012	1	-0.012	0.02	0.066	0.053	0.096	0.067	-0.0032
days_in_waiting_list	0.054	0.17	-0.054	-0.002	-0.0083	-0.033	-0.011	-0.022	0.0059	-0.0094	-0.012	1	-0.041	-0.031	-0.083	-0.023	-0.035	-0.026
adr	0.048	-0.063	0.049	0.065	0.23	0.32	0.029	-0.13	-0.066	-0.072	0.02	-0.041	1	0.057	0.17	0.068	0.32	0.37
required_car_parking_spaces	-0.2	-0.12	-0.019	-0.025	0.015	0.056	0.037	0.077	-0.018	0.048	0.066	-0.031	0.057	1	0.083	-0.026	0.063	0.048
total_of_special_requests	-0.23	-0.096	0.073	0.068	0.12	0.082	0.098	0.013	-0.048	0.038	0.053	-0.083	0.17	0.083	1	0.079	0.1	0.16
total_nights	0.018	0.16	0.76	0.94	0.11	0.051	0.022	-0.11	-0.015	-0.053	0.096	-0.023	0.068	-0.026	0.079	1	0.054	0.12
kids	-0.0028	-0.041	0.049	0.047	0.034	0.97	0.26	-0.034	-0.026	-0.022	0.067	-0.035	0.32	0.063	0.1	0.054	1	0.6
total_persons	0.047	0.072	0.1	0.1	0.82	0.58	0.16	-0.14	-0.02	-0.099	-0.0032	-0.026	0.37	0.048	0.16	0.12	0.6	1

From correlation heatmap we see:

- ‘is_canceled’ is positively correlated with ‘lead_time’, which we saw on the previous slide.
- Stays in week & weekend nights are highly & positively correlated to total nights, but weeknights is higher, indicating stays during weeknights are more.
- ‘adults’ is more positively correlated with ‘total_persons’ than ‘kids’ indicating majority of the customers are adults.
- ‘adr’ is positively correlated with ‘total_persons’, indicating hotels can earn more if persons per booking are more.
- ‘kids’ is highly & positively correlated with ‘children’ than ‘babies, indicating that families with infants travel less as compared to families with grown kids.

CONCLUSION:

- Majority of the customers are Portuguese nationals (28% of overall), ‘City’ & ‘Resort’ Hotels are located in Portugal and ‘City’ hotel is the busier of the two hotels.
- The retention rate of ‘Resort’ hotel is better for both Portuguese & Non-Portuguese nationals, but it could market on attracting Non-Portuguese customers more.
- “Online TA” market segment is bringing most of the business to the hotels followed by “Offline TA/TO”.
- Almost all customers prefer paying at the time of checking out from hotel.
- For ‘City’ hotel, majority of the customer stay 2-3 nights where as for ‘Resort’, customer stay 3-5 nights, indicating customers there come for a longer trip.
- The median ‘ADR’ for ‘City’ hotel is around 100 irrespective of duration of stay, whereas for ‘Resort’ it increases till 6 nights of stay, but stays below 100. So for longer duration ‘Resort’ hotel is economical.
- The most desired room type is ‘A’, followed by ‘D’. Majority of those who didn’t get room type ‘A’ were allotted room type ‘D’.
- The peak season for ‘City’ hotel is from March-October and the trend of ‘demand’ & ‘adr’ is similar, whereas for ‘Resort’, the peak season is from February-May & October and the trend of ‘demand’ & ‘adr’ is opposite in nature.
- Not getting the desired room type, doesn’t affect booking cancellations.
- Hotels need to focus more on cancellations via ‘Online TA’ market segment, because that is where majority of the ‘No Deposit’ cancellations are & the median lead time for “cancelled” bookings is considerably higher than “confirmed” bookings.
- Stay during weeknights is higher than stay during weekend nights.
- The majority of the customers are the adult population
- Families with babies travel less than families with grown children.

THANK YOU!