Capstone Project – EDA

Hotel Booking Analysis

Aviral Singh

FLOW OF PRESENTATION:

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

OBJECTIVE:

- Hotel industry is a fast paced industry. Thus, the hotel operators need to adapt to this fast paced industry in order to operate efficiently.
- The objective of this project is to analyze the given ‘Hotel Booking Dataset’, so that we can observe, 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, cancelled bookings, customer retention, deposit type, etc.

Now, let’s take a look at the variables in our dataset.

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 a particular booking.
children : Number of children in a particular booking.
babies : Number of babies in a particular 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 the sum of all lodging transactions divided 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’.

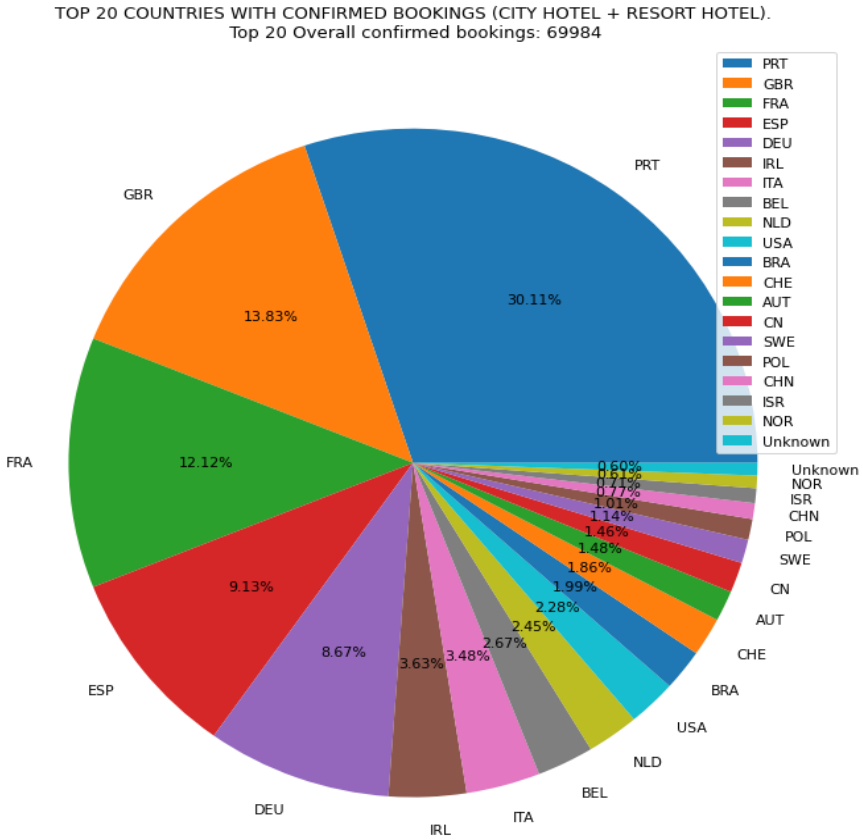
Total number of variables in our dataset : 31
Total number of records in our dataset : 119,390

EXPLORATORY DATA ANALYSIS:

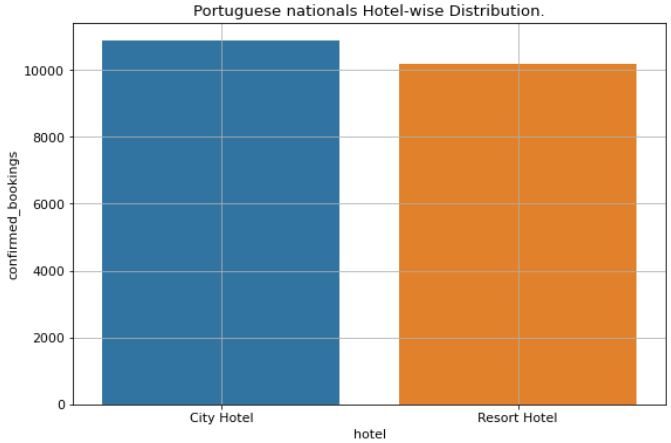
- Out of the 119,390 records:
  ‘Confirmed’ Bookings: 75,166
  ‘Cancelled’ Bookings: 44,224

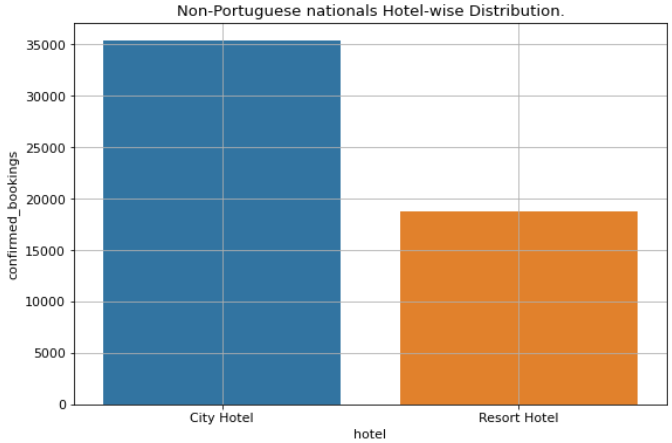
- The time duration (year-month) of our dataset : 2015-07 to 2017-08.
- Both ‘confirmed’ bookings and ‘cancelled’ bookings will be looked upon.

1. Country of Origin of Customers:



- From the pie chart we can observe that majority of the customers (with confirmed bookings) come from Portugal (PRT): 30.11% (21071).
- Portugal is followed by Britain (GBR): ~14% (9676) and France (FRA): ~12% (8481).
- There's a huge gap between Portugal and Britain. We can now definitively say that the hotels 'CITY' and 'RESORT' are located in Portugal.
- Their business is mostly being driven by the Portuguese nationals.





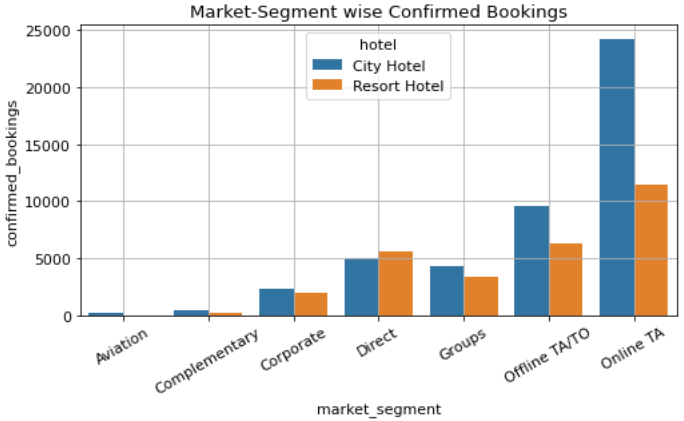
City Hotel PORTUGUESE nationals RETENTION RATE: 12.28 %
Resort Hotel PORTUGUESE nationals RETENTION RATE: 13.62 %

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

From the above shown bar plots, we can observe:

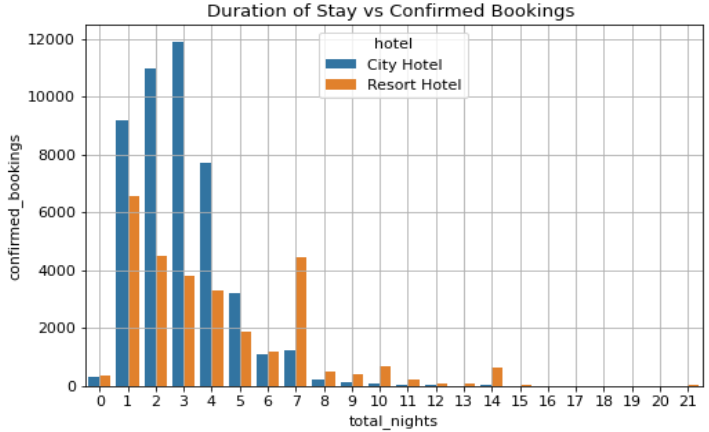
- Portuguese nationals seem to have similar preference for both 'City' and 'Resort' hotel as compared to Non-Portuguese, with ~11000 for 'City Hotel' and slightly above 10000 for 'Resort Hotel'.
- While observing only Non-Portuguese nationals, 'City Hotel' has catered to almost double the customers of what 'Resort Hotel' has.
- 'Resort Hotel' has a better retention rate among both Portuguese and Non-Portuguese customers.
- 'Resort Hotel' needs to market itself more among Non-Portuguese customers.

2. MARKET-SEGMENT wise & PAYMENT-MODE wise CONFIRMED Bookings:



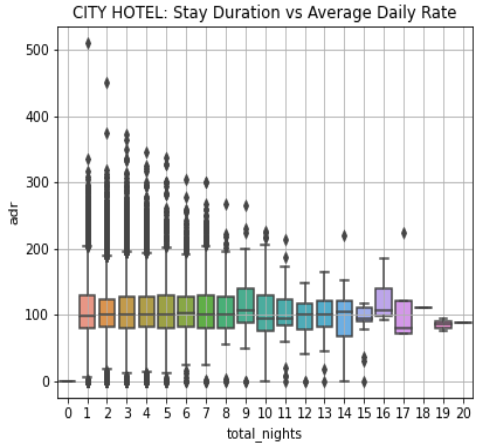
- The 'Online TA' segment brought in most of the business to both hotels.
- 'Offline TA/TO', 'Groups', 'Direct' & 'Corporate' segments also brought in business, but 'Online TA' brought in way more, which seems to be normal in the internet-age.
- 'Aviation' & 'Complementary' segments brought in very low business.
- In the terms of payment mode preferred by customers, almost all customers preferred 'No Deposit' mode of payment, i.e., they preferred to pay at the check-out time.

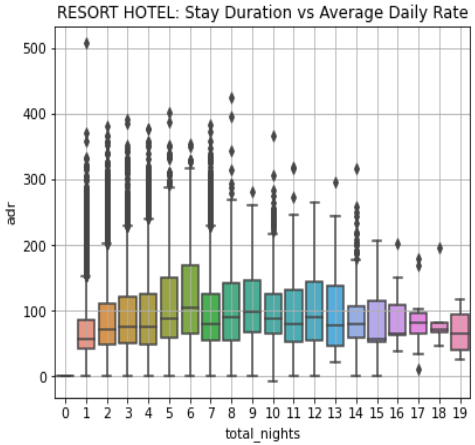
3. Duration of Stay & Average Daily Rate (ADR):



Duration of Stay vs Confirmed Bookings:

- **City Hotel:** Most customers have preferred to stay up to 3 nights. Number of customers staying beyond 3 days decreases considerably.
- **Resort Hotel:** Most customers have preferred to stay for 3-5 nights. A significant number of customers have also preferred to stay up to 7 nights. This indicates that 'Resort Hotel' customers stayed for a longer trip.

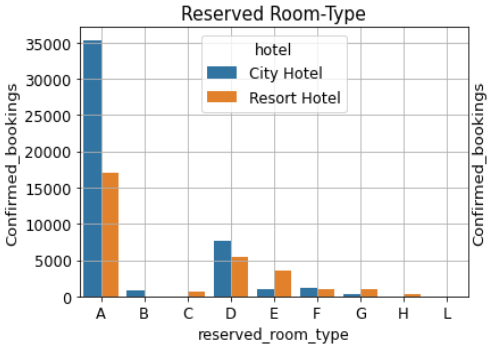


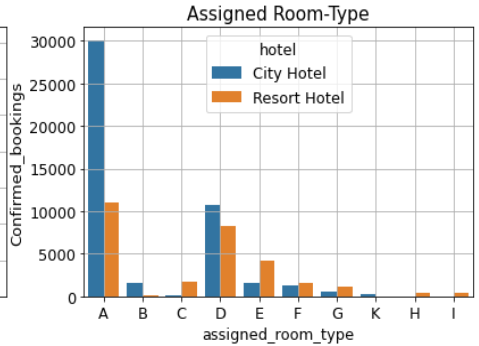


Duration of Stay vs ADR:

- **City Hotel:** Variability in ADR is higher for shorter stay duration & it's lower for longer stay duration. Median ADR stayed around 100, irrespective of the length of stay.
- **Resort Hotel:** Variability is similar to that of 'City Hotel', but the median ADR increased with the stay duration till 6th night. It oscillated afterwards, but didn't cross 100. Median ADR generally stayed below 100. Longer trips also turned out to be economical.

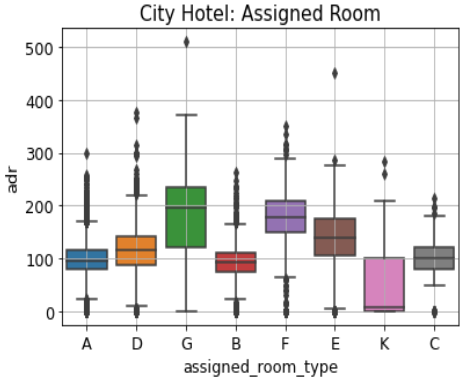
4. Room-Type & ADR:

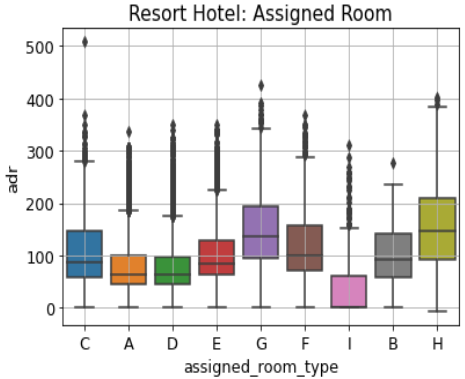




Room-Type Bar Plots:

- Room-type A had the highest demand, ~35,000 for 'City Hotel' and ~17,000 for 'Resort Hotel'.
- Everyone couldn't be assigned room-type A, as seen on comparing Assigned Room-Type bar plot with Reserved Room-Type bar plot.
- Those who couldn't be assigned room-type A, most of them were assigned room-type D.

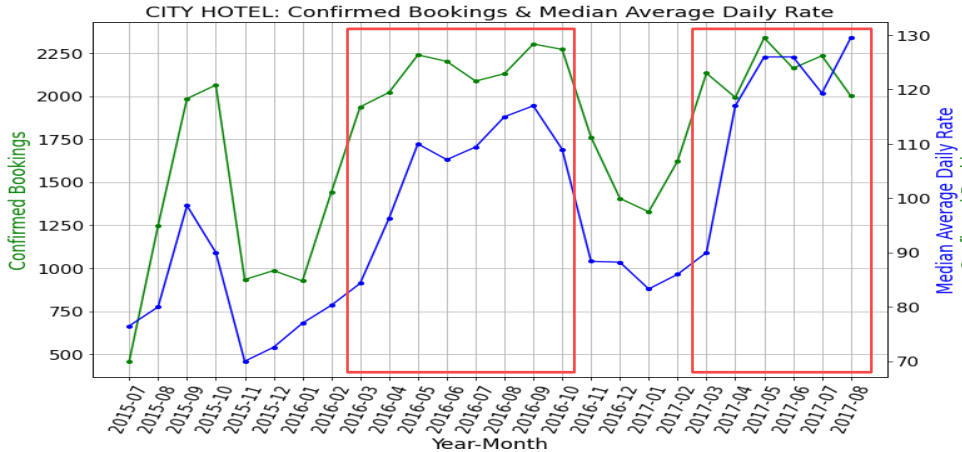


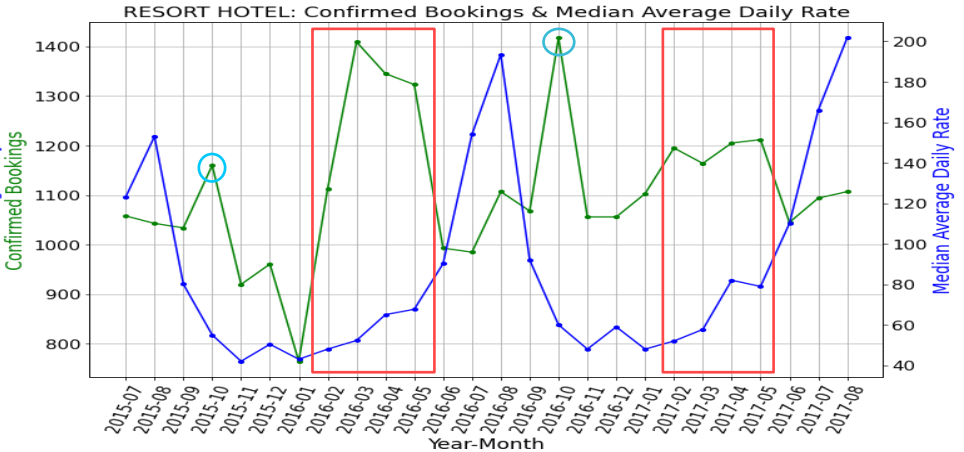


Assigned Room-Type vs ADR:

- **City Hotel:** Rooms A & D were assigned the most. The median ADR of room-type D is slightly higher than room-type A. Room-types G, F, E have considerably higher median ADR, indicating of those being high-end rooms.
- **Resort Hotel:** Room-types A & D were assigned the most. The median ADR of room-types A & D were similar. Room-types G & H had considerably higher median ADR, indicating of those being high-end rooms.

5. Year-Month wise DEMAND & Median ADR:





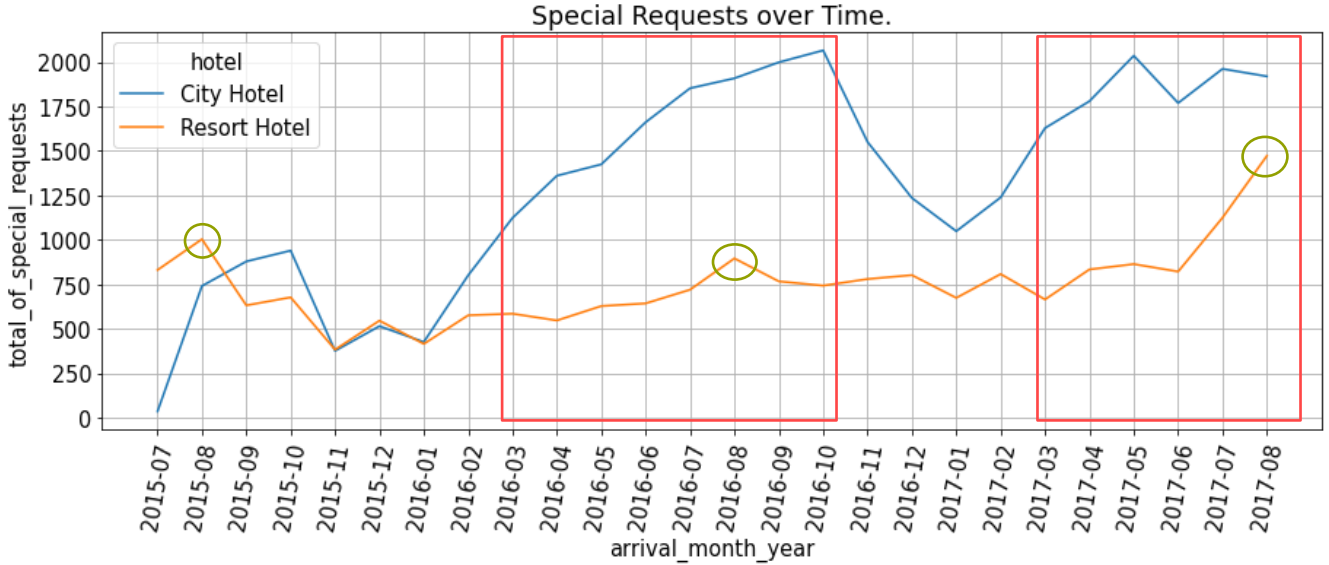
City Hotel Observations:

- The peak season is from March to October. Number of customers decrease during the winter season, i.e., from November to January. It starts going back up from February.
- The trend of ADR over months is similar to that of the demand, i.e., as the number of customers increases, the ADR also increases, and vice versa.
- From a customer's perspective, trip during the winter season would be cheaper.

Resort Hotel Observations:

- The peak season is from February to May, and there is another peak in October. Number of customers decreases during June to September & November to January.
- The trend of ADR (in blue) over months is opposite to that of the demand, i.e., as the number of customers increases, the ADR decreases, and vice versa.
- From a customer's perspective, trip during the peak season of the hotel would be economical.

6. Year-Month wise Special Requests:



- **City Hotel:** The trend of special requests is similar to the trend of its demand that we saw in previous slide, i.e., increasing from March to October, and then decreasing from November.

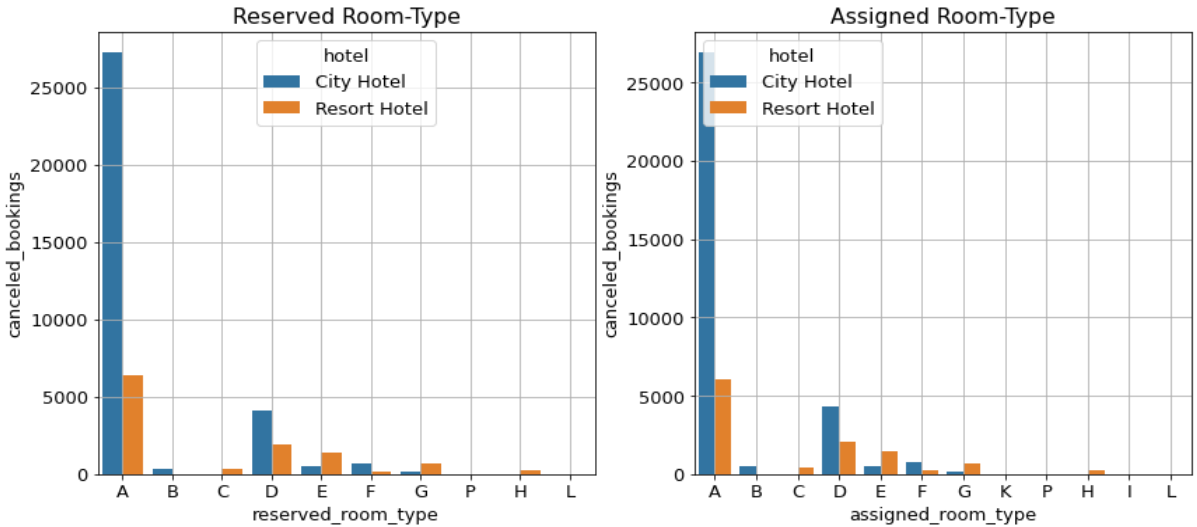
- **Resort Hotel:** The trend is fairly constant, with dips in winter season, but there are spikes in the month of August every year from 2015-2017.

CANCELLED BOOKINGS

7. Room-Type wise Booking Cancellations:

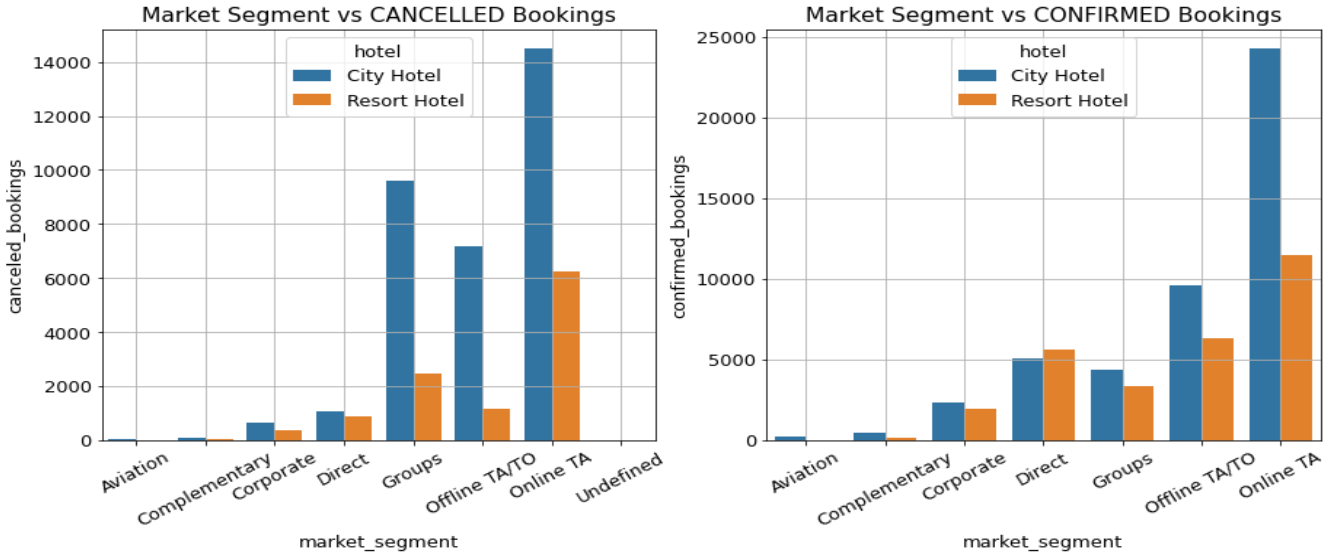
Hypothesis:
Not getting the desired/reserved room-type is a factor in booking cancellations.

Visual Inspection:



- There is almost no change between the two plots.
- ‘Not getting the desired/reserved room-type' doesn't seem to have much effect on booking cancellations.

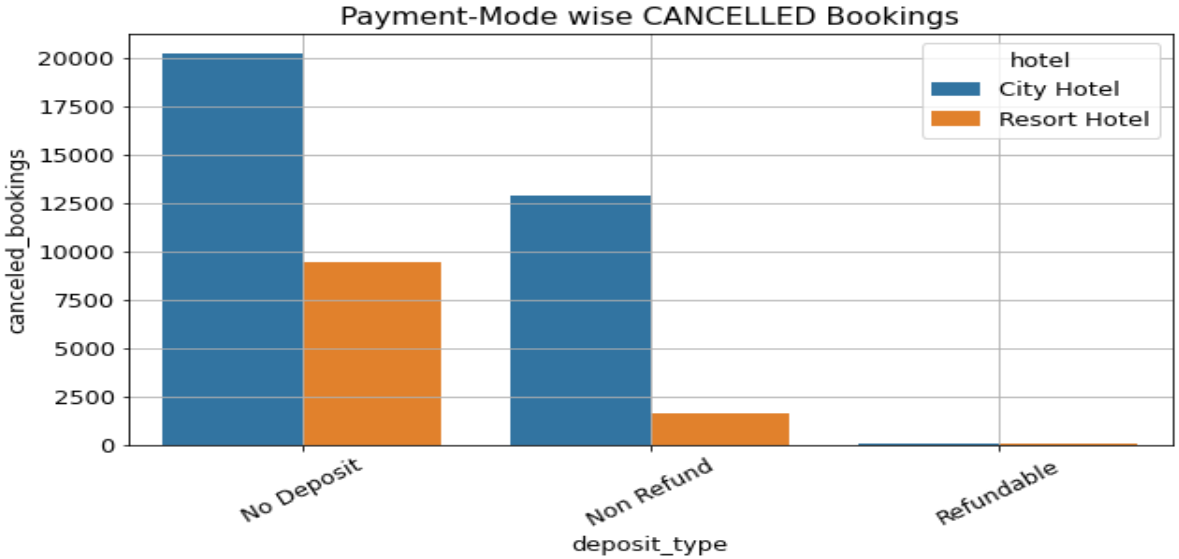
8. Market-Segment wise Booking Cancellations:



- 'Online' & 'Offline TA/TO' market-segments brought in maximum customers. So, it is understandable that maximum 'cancelled' bookings would also be from these segments.

- Totally opposite to previous observation, in the 'Group' market segment, for 'City Hotel', we see that more people have ‘cancelled’ their bookings (~ 10000) than 'confirmed' (~ 4000), which seems out of norms.

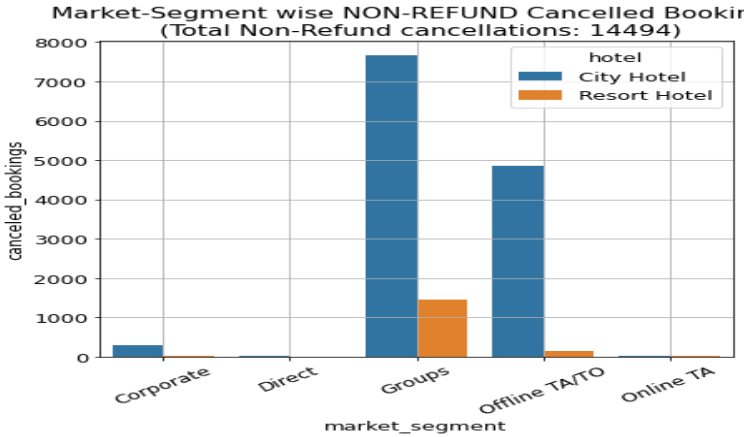
9. Payment-Mode wise Booking Cancellations:

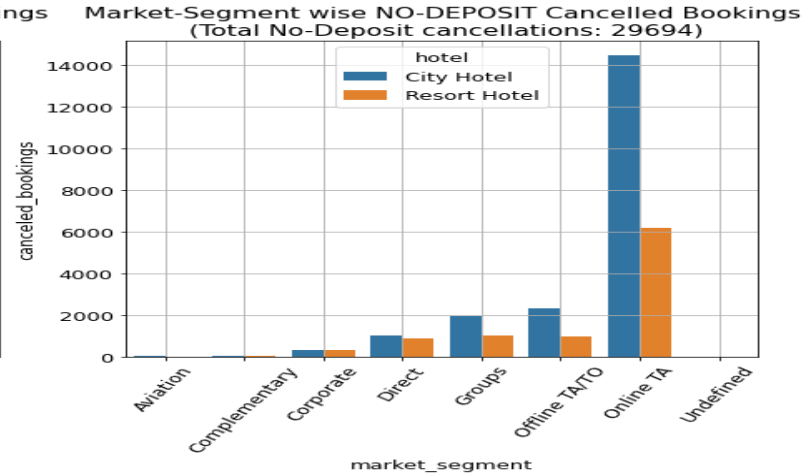


Payment-Mode wise Cancelled Bookings Observations:

- **City Hotel:** ~20000 cancelled bookings under 'No Deposit', which makes sense because customers didn't lose any money after cancellation. Surprisingly, ~12500 'Non Refund' bookings were cancelled. So many customers choosing to lose money post cancellation seems surprising.
- **Resort Hotel:** ~9000 cancelled bookings under 'No Deposit' and ~1500 cancelled bookings under 'Non Refund'. Not as high as 'City Hotel'.

10. Market-Segment wise Non-Refund & No-Deposit Cancelled Bookings:

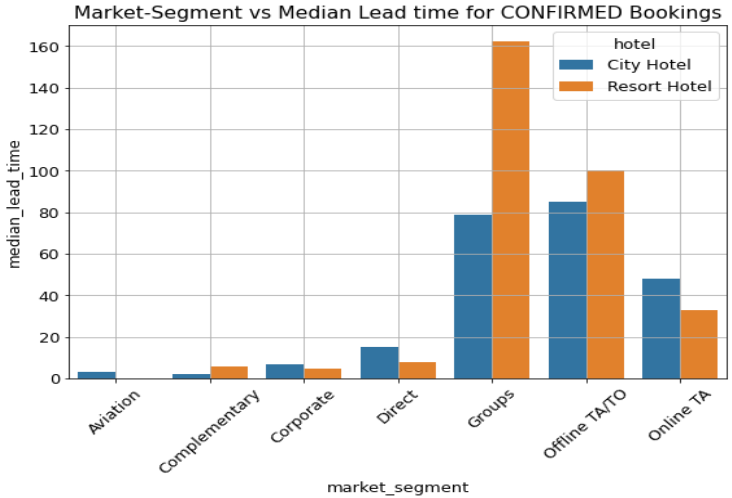


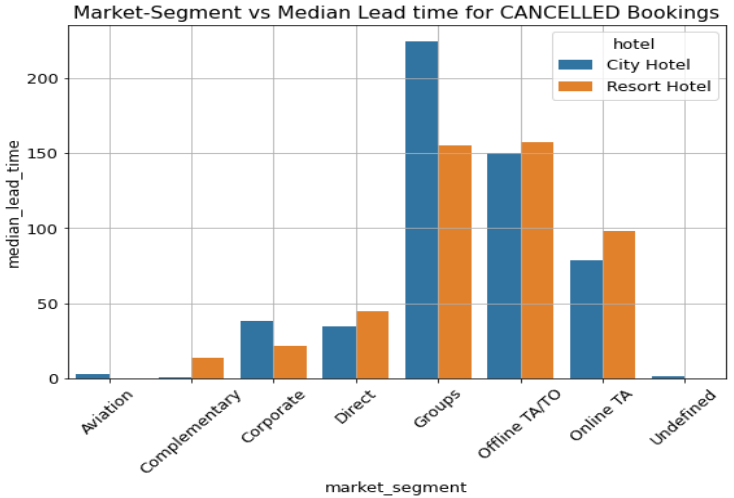


Earlier we had observed lots of cancellations from ‘Group’ segment, and the same in ‘Non-Refund’ payment mode.

- In the ‘Non-Refund’ cancellations, it can be observed that, for ‘City’ Hotel, majority of the cancellations are happening from ‘Groups’ segment and a little bit from the ‘Offline TA/TO’ segment.
- It's also clear that ‘No-Deposit’ cancellations are mostly from ‘Online TA’ segment for both hotels.
- Hotels need to worry less about the ‘Non-Refund’ cancellations, as they are getting paid in advance and they don’t have to repay the customer.
- Hotels need to worry about the ‘No-Deposit’ cancellations, that are mostly coming from ‘Online TA’ market-segment for both hotels, as they aren’t getting paid and then they have to fill those large vacant bookings with new customers.

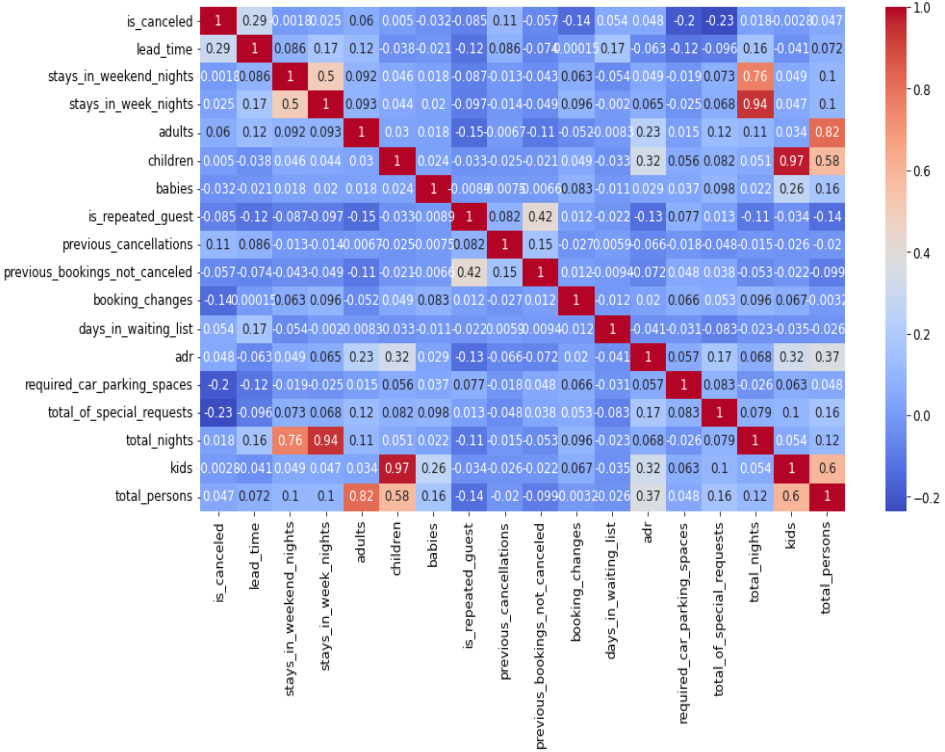
11. Market-Segment vs Median Lead Time for Confirmed & Cancelled Bookings:





- From visual observation, across the market-segments, generally the 'cancelled' median Lead Time is higher than 'confirmed' median Lead Time.
- If we pay attention to Online TA segment (where most No-Deposit cancellations are coming from), among 'confirmed' bookings, the median Lead Time for 'City Hotel' is ~50 days (under 2 months) & for 'Resort Hotel' it is ~30 days (1 month).
- Among 'cancelled' bookings, the median Lead Time for 'City hotel' is ~80 days (> 2 months) & for 'Resort Hotel' is ~100 days (> 3 months).

12. Correlation Heatmap:



- ‘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 for higher persons per booking.
- ‘kids’ is highly & positively correlated with ‘children’ than ‘babies, indicating that families with infants travel lesser as compared to families with grown kids.

CONCLUSION:

- Majority of the customers are Portuguese nationals (28% overall). 'City' & 'Resort' hotels are located in Portugal. 'City Hotel' is busier than 'Resort Hotel'.
- 'Resort Hotel' has better retention rate among both Portuguese & Non-Portuguese nationals, but it should consider marketing to attract more Non-Portuguese customers.
- 'Online TA' market-segment brought most of the business to the hotels, followed by 'Offline TA/TO'.
- Most customers prefer paying at the check-out time.
- In 'City Hotel', most of the customers stay 2-3 nights. In 'Resort Hotel', customers stay 3-5 nights, indicating customers come there for a longer trip.
- The median ADR for 'City Hotel' is around 100 irrespective of stay duration, whereas for 'Resort Hotel' it increases till 6th night of stay, but stays below 100. So, 'Resort Hotel' is economical for longer stay duration.
- The most desired room-type is 'A', followed by 'D'. Majority of those who didn’t get room-type 'A' were assigned room-type 'D'.
- The peak season for 'City Hotel' is from March to October and the trend of 'demand' & 'adr' is similar. The peak season for 'Resort Hotel' is from February to May & October, and the trend of 'demand' & 'adr' is opposite in nature.
- 'Not getting the desired room type' doesn’t have much effect on booking cancellations.
- Hotels need to focus more on cancellations via ‘Online TA’ market-segment, as that is where majority of the ‘No-Deposit' cancellations come from & 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 adults.
- Families with babies travel less than families with grown children.

THANK YOU!