CAPSTONE PROJECT

Exploratory Data Analysis on Hotel Booking

By: Sabbavarapu Malathy Lata

Points to Discuss

1. Background and Research Objective
2. Data Summary
3. Overview of Hotel Industry in India
4. Univariate Analysis
5. Hotel Type Analysis
6. Distribution Channel and Market Segment Analysis
7. Analysis Based on Time Period
8. Analysis Based on Cancellation of Bookings
9. Analysis Based on Special Requests
10. Correlation Heatmap
11. Optimal Stay for customers
12. Conclusion

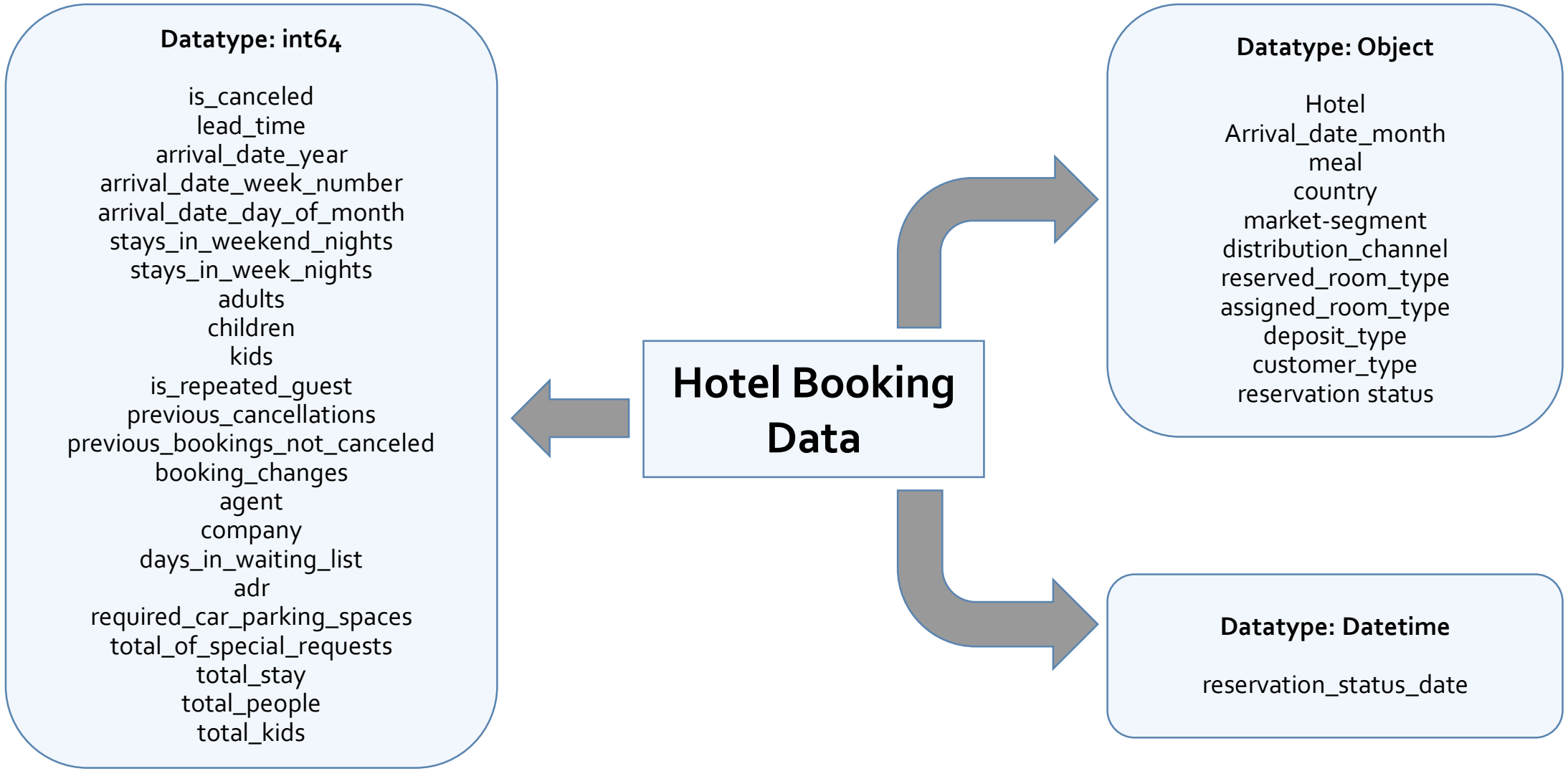
Background and Research Objective

- In this project, we will be analyzing the Hotel Booking data for the given period from 2015 to 2017. The data contains information about bookings and cancellations made for the city and resort hotels, information concerning the bookings made such as waiting period, lead time, if any agent is involved during the booking process, the country of origin of the guest, stay period, month-wise bookings, and more.

- The hotel industry is mainly boosted by tourism and the number of bookings made depends on the factors provided in the dataset.

- The main objective of this exploratory data analysis is to explore and analyze data to discover important factors that govern and aid in hotel booking and to give insights to hotel management for improving the booking percentage and retention of customers.

Data Summary



Data Summary Contd...

Table 1: Data column description for datatype **int64**

Sl.No.	Column Name	Description
1	is_canceled	Corresponds to the value indicating if the booking was canceled (1) or not canceled (0)
2	lead_time	Indicates the period between the booking date and the actual arrival date
3	arrival_date_year	Indicates the year of actual arrival date
4	arrival_date_day_of_month	Indicates the date of the month of actual arrival
5	stays_in_weekend_nights	Indicates the number of weekend nights (Saturday and Sunday), the guest stayed or booked to stay at the respective hotel
6	stays_in_week_nights	Indicates the number of weekday nights (Monday and Friday), the guest stayed or booked to stay at the respective hotels
7	adults	Indicates the total number of adults in a particular booking
8	children	Indicates the total number of children in a particular booking
9	kids	Indicates the total number of kids in a particular booking
10	is_repeated_guest	Indicates if the guest is a repeated guest (1) or not a repeated guest (0) for a respective hotel type
11	previous_cancellations	Indicates the number of cancellations made for bookings before the current booking
12	previous_bookings_not_canceled	Indicates the number of bookings that were not canceled before the current booking
13	booking_changes	Indicates the number of changes done for a booking from the day of booking till any cancellation or check-in occurs
14	agent	Indicates the identification number of the travel agency that made the booking
15	company	Indicates the identification number of the company that made the booking
16	days_in_waiting_list	Indicates the number of days a booking is wait-listed before confirmation of the booking
17	ADR	Average Daily Rate (ADR) indicates the average revenue earned for an occupied room on a given day
18	required_car_parking_spaces	Indicates the number of car parking spaces required by the guests for a particular booking
19	total_of_special_requests	Indicates the number of special requests made by the guests for a particular booking

Data Summary Contd...

Table 2: Data column description for datatype **object**

Sl.No.	Column Name	Description
1	Hotel	Indicates the two hotel types- City and Resort hotels
2	Arrival_date_month	Indicates the month of actual arrival date
3	meal	Indicates the meal type chosen by guests- BB (Breakfast and Bed), FB (full board extended: 3 meals a day), SC (Self-catering: no meals included), HB (half board: 2 meals a day)
4	country	The value represents the country of origin of the guest
5	market_segment	Represents the market segments direct, corporate, offline and online tour agencies(TA)/tour operators(TO), complementary, groups, and aviation
6	distribution_channel	Represents the distribution channels direct, corporate, offline and online tour agencies(TA)/tour operators(TO), and Global distribution systems (GDS)
7	reserved_room_type	Indicates the code of the reserved type of room at the time of booking
8	assigned_room_type	Indicates the code of the assigned type of room during the arrival of the guest
9	deposit_type	Indicates the type of deposit the hotel requires the guests to pay for confirmation of the booking. This includes refundable deposits, non-refundable deposits, and no deposit
10	customer_type	Indicates the type of customer visiting the hotel namely transient, contract, and group
11	reservation_status	Indicates the reservation status of the booking namely check-out, canceled, and no-show

Table 3: Data column description for datatype **datetime64[ns]**

Sl.No.	Column Name	Description
1	reservation_status_date	The value indicates the date of the status specified for the reservation of the booking

Overview of Hotel Industry In India

- The hotel industry is the sub-segment of the hospitality industry which is a broad category of fields within the service industry.
- The hotel industry provides services such as accommodation by hotels, resorts, and professionally run guest houses which can be booked directly, through provider's websites, travel/tourism agents, online travel agencies (OTA), and more.
- In 2020, the revenue for the Indian travel, hospitality, and tourism industry was USD 75 billion. By 2027, it is estimated to grow and reach a revenue of USD 125 billion.
- The hotel industry accompanied by tour operations and the restaurant segment in tourist areas was the main driver of the growth of the industry.

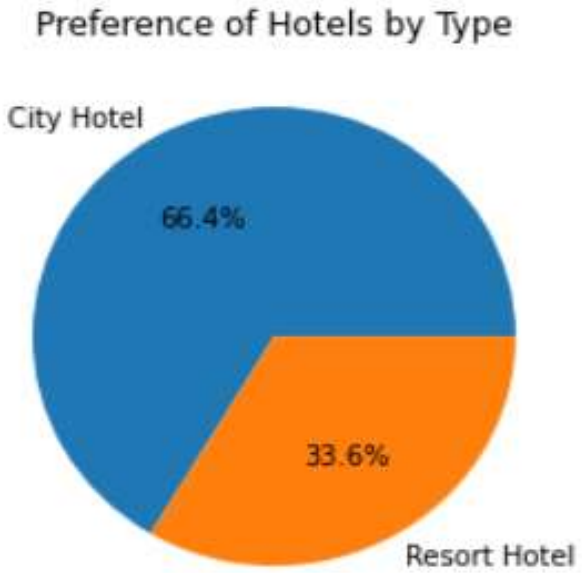
- In India, in 2020, around 8% of the total employment corresponds to the 39 million jobs under Indian tourism.
- In 2021, a total of 144 thousand hotel rooms were present in India.
- By 2028, through visitor exports (spending within a country by international tourists for leisure or business travel) Indian hospitality and tourism industry is expected to reach a revenue of USD 50.9 billion.

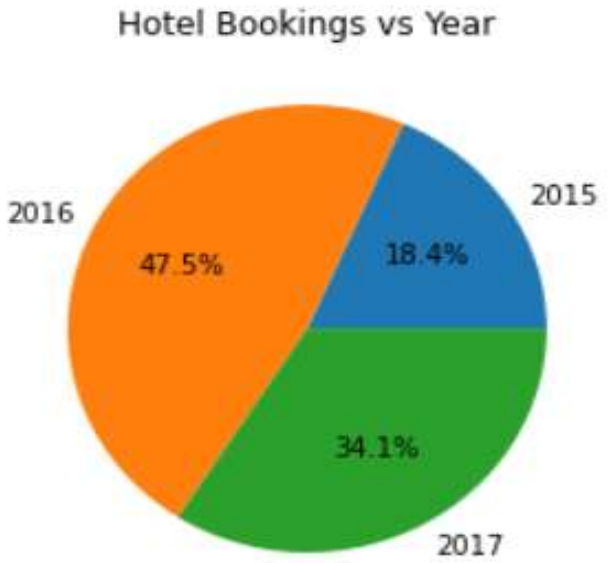
Univariate Analysis

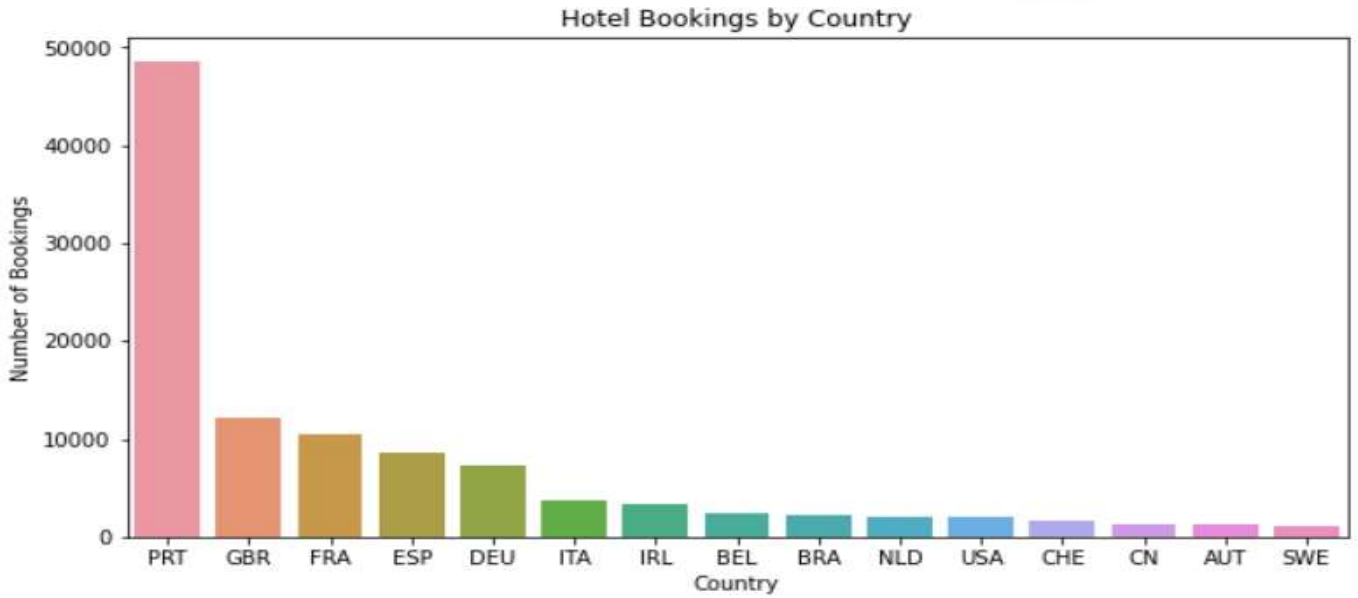
- The city hotel is the hotel type with more bookings, around 65%, across the years mentioned in the dataset. Whereas the resort hotel bookings constituted around 35%.

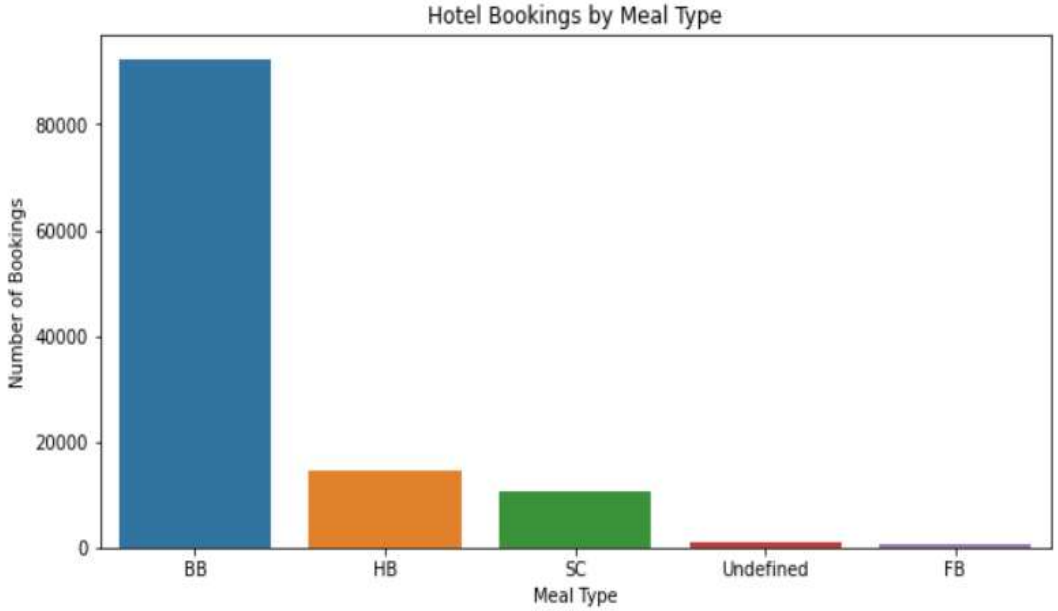
- From the pie chart, it is evident that the number of bookings in 2016 was greater than in 2015, and 2017. It consisted of around 50%. The number of bookings increased in 2016 and decreased again in 2017 to around 35%.

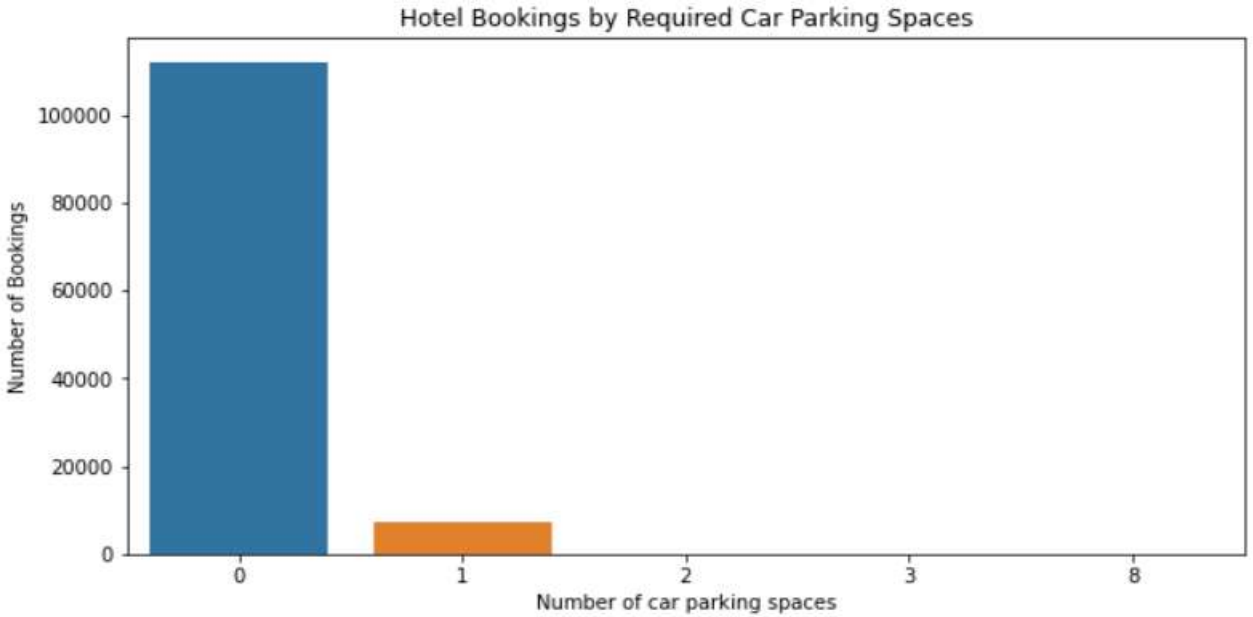
- The most bookings made were from the European countries, predominantly from Portugal followed by Great Britain, France, and Spain.







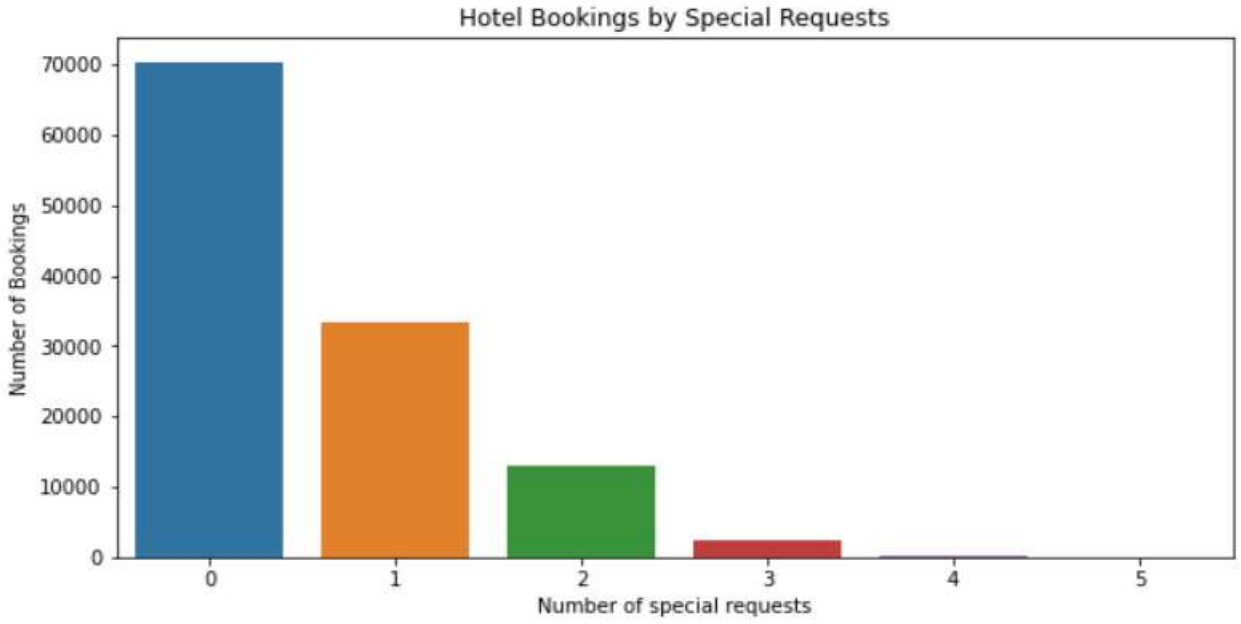


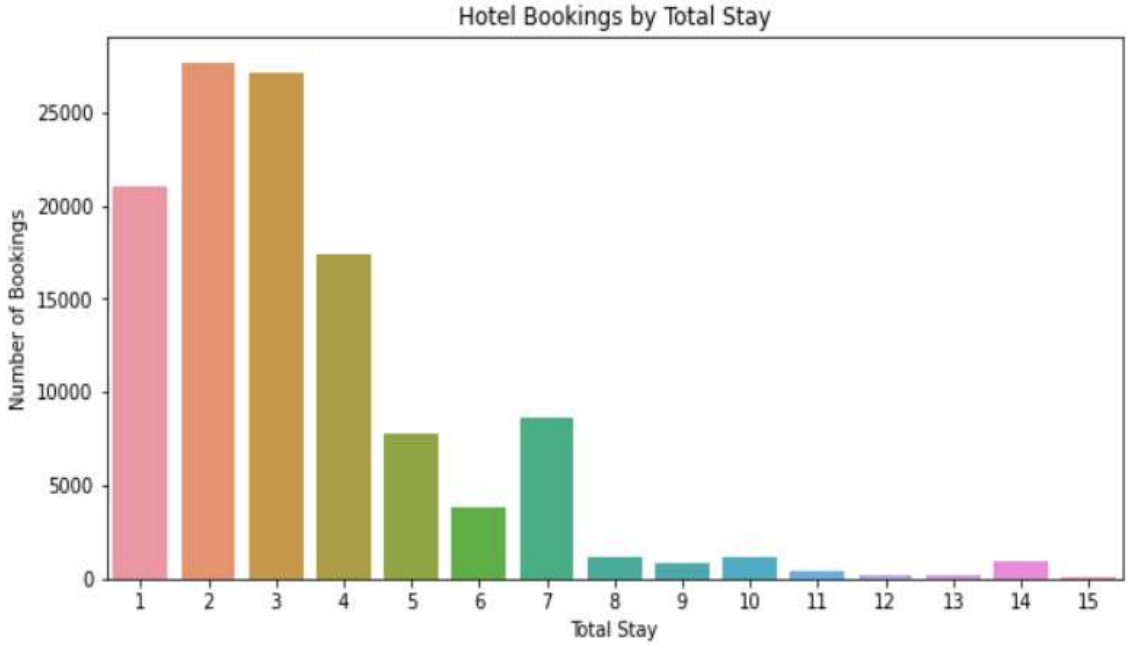


- Bed and Breakfast (BB) was the most preferred meal type for both the hotels, approximately 77%. Half board (HB) and Self catering (SC) are similarly preferred.

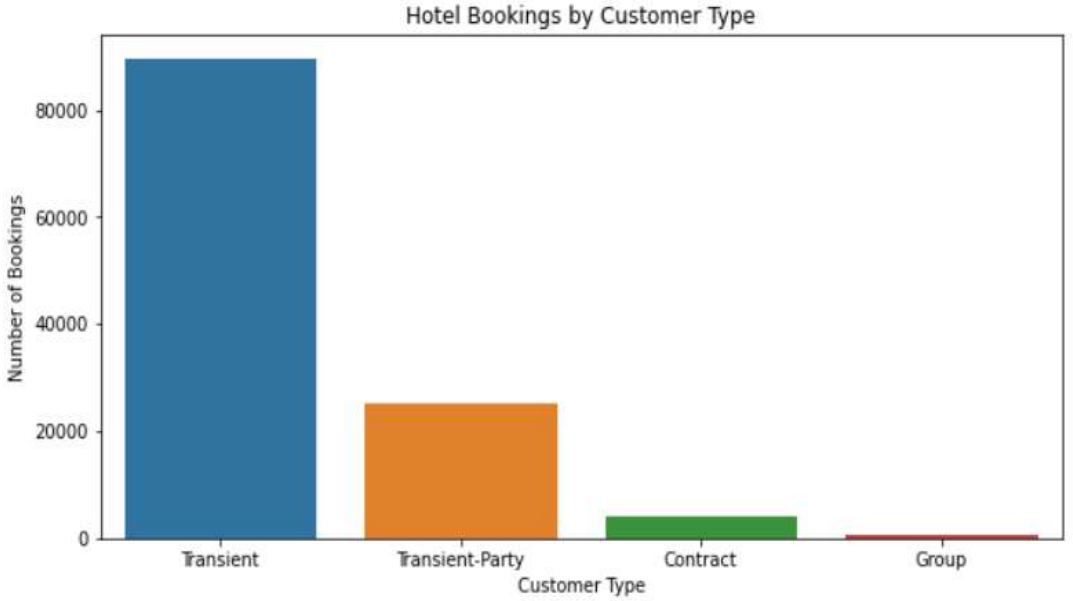
- Most of the bookings, approximately 94%, required no parking spaces, followed by 6% requiring 1 car parking space and a negligible <1% requiring 2, 3, and 8 car parking spaces.

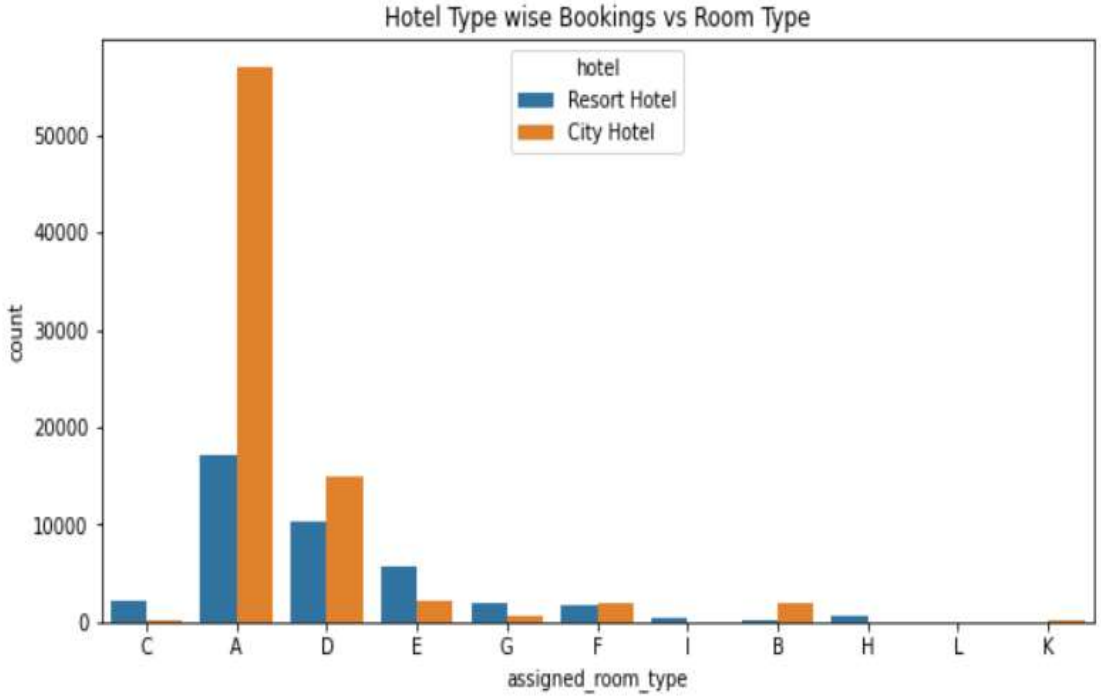
- The majority of the hotel bookings were not received with any special requests. Around 59% of bookings received no special requests, followed by around 28% receiving 1 special request, followed by around 10% receiving 2 special requests, and a less than 5% of bookings receiving 3, 4, and 5 special requests.

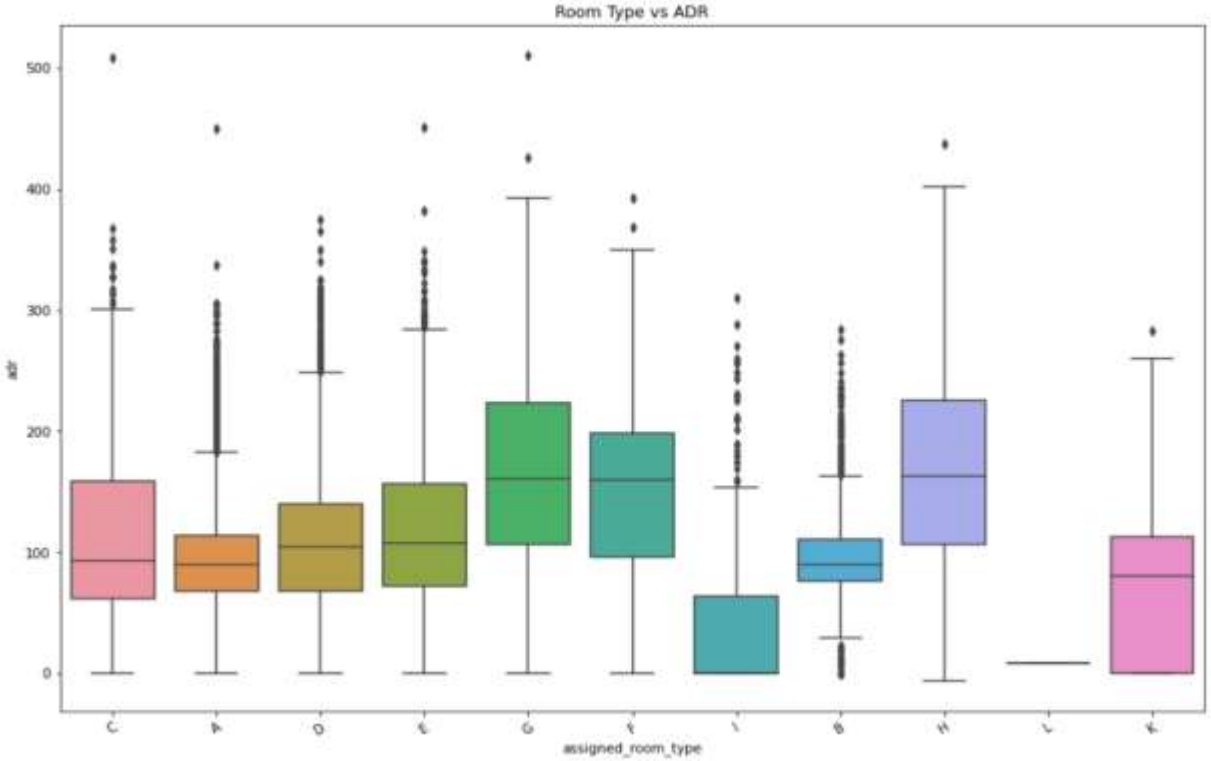




- The most preferable stay length for the guest in the hotels is of less than 7 days.
- Agent number 9 has done the most number of bookings followed by agent number 240 and 1.
- Almost 75% of the bookings made were from the transient customer. This is followed by 21% of bookings made from transient-party type, followed by less than 5% of bookings made from contract and group type of customers.

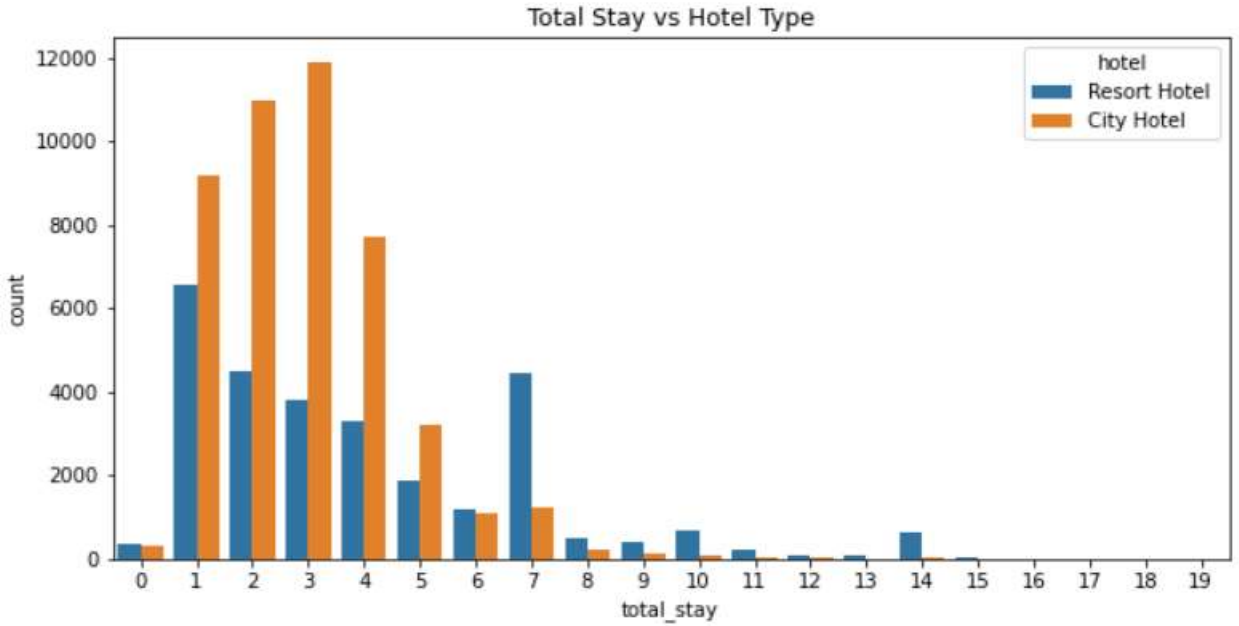


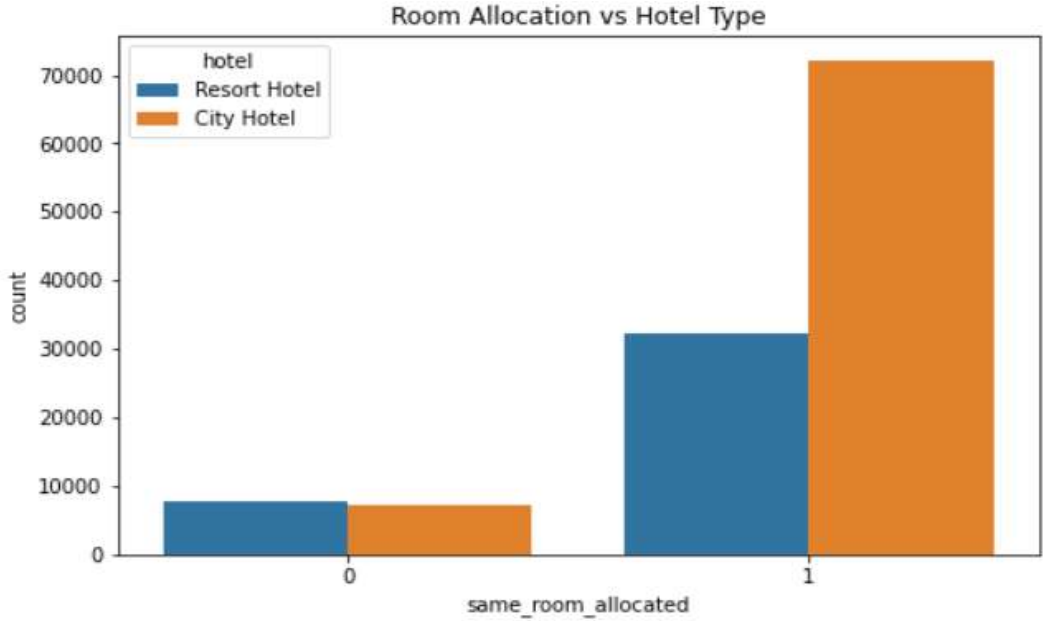




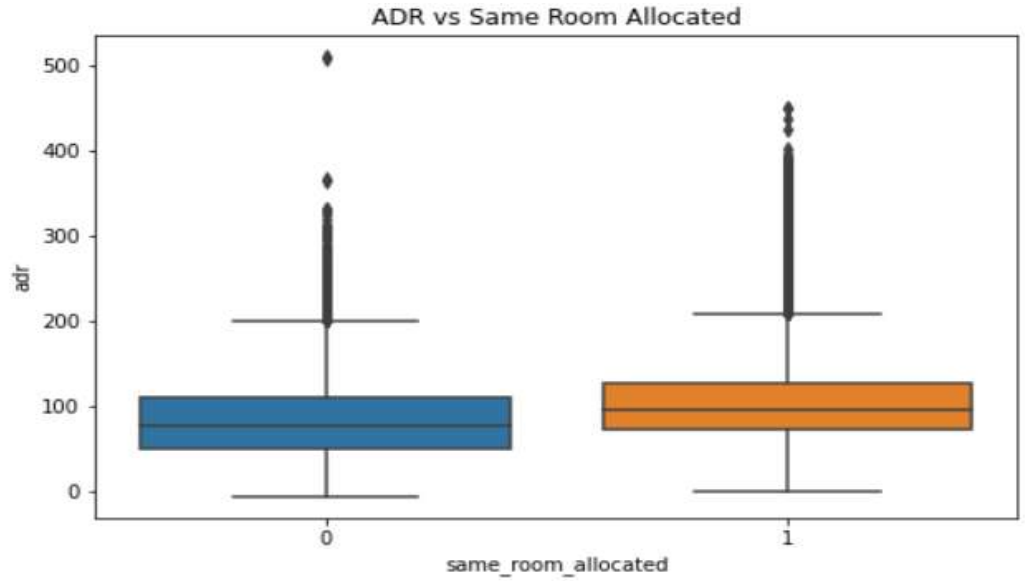
- Room type A is the most demanded room type for both city and resort-type hotels.
- Room types G, F, and H generated higher ADRs when compared to other room types. Hence, hotels can increase the focus on their sales of room types G, F, and H to increase their revenue.
- For most of the bookings, the hotels did not require any deposit with a booking percentage as high as approximately 88%. Around 12% of guests also paid a non-refundable deposit and less than 1% paid a refundable deposit.

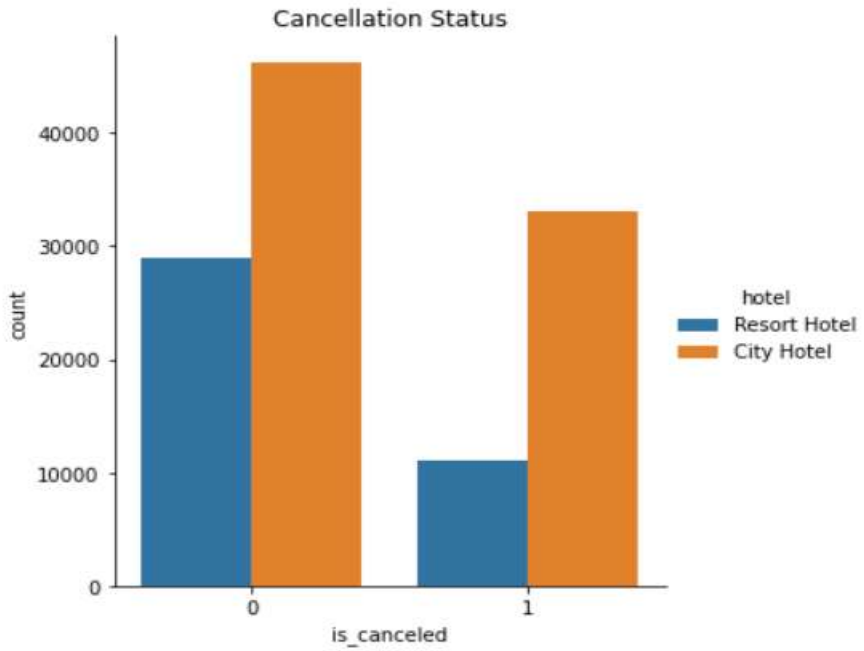
Hotel Type Analysis

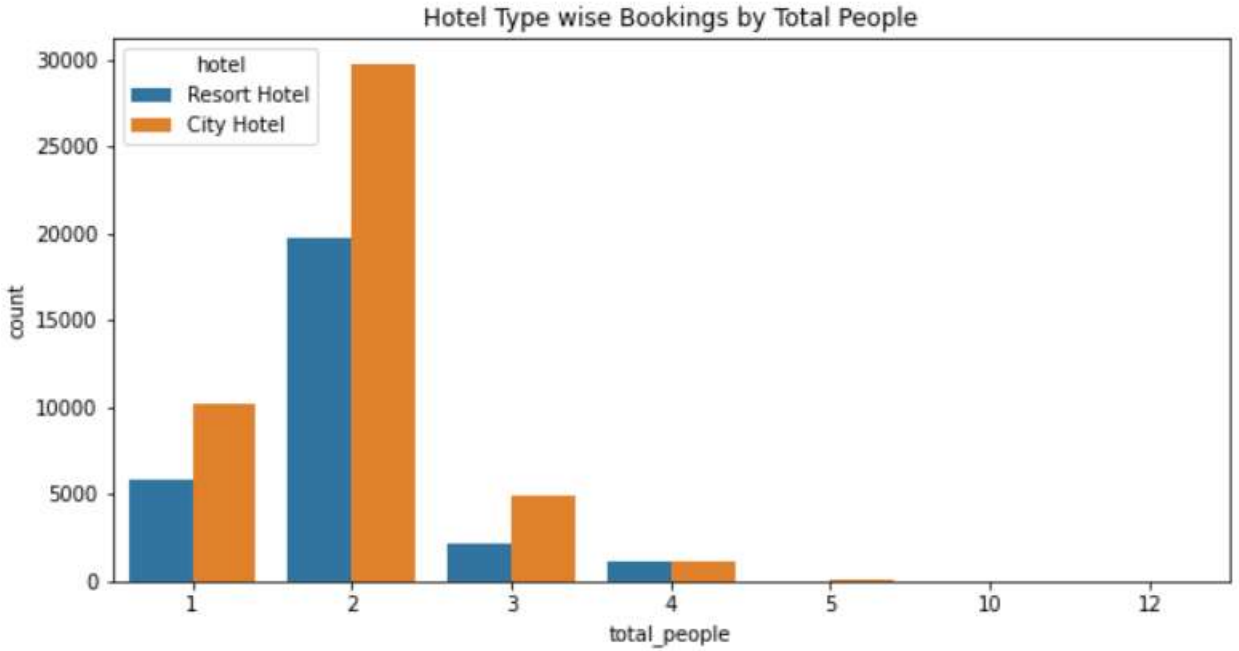




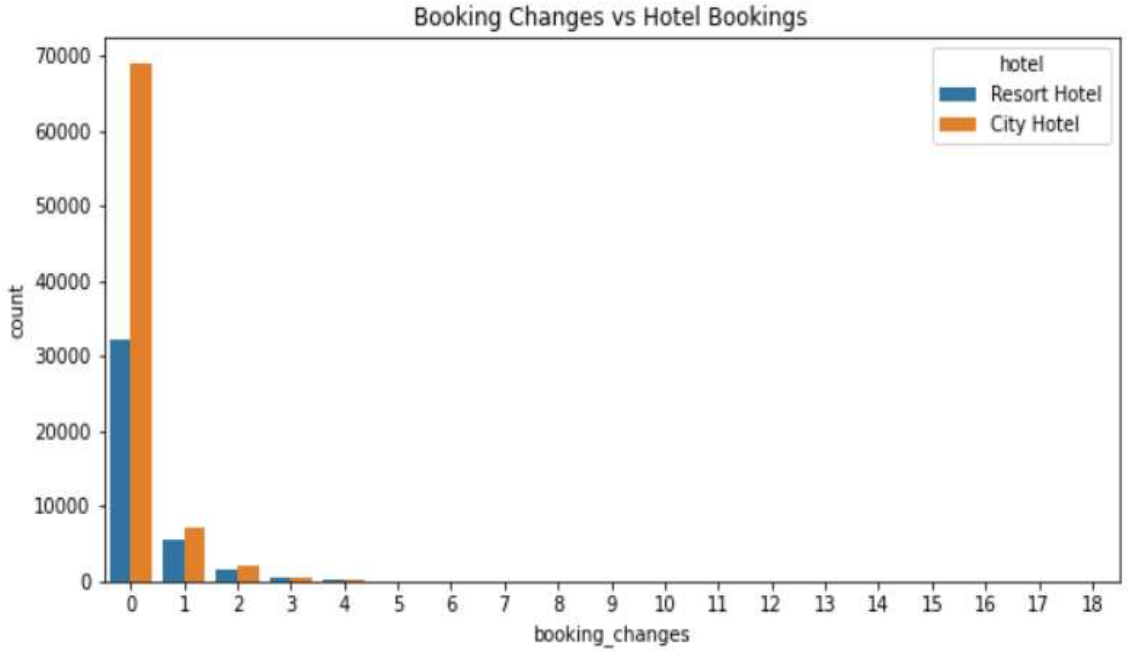
- For city hotels, the preferred stay length was <= 5 days and that for resort hotels was <= 7 days. For longer stays (greater than 5 days) resort hotels were preferred.
- For both the hotel types, the majority of the bookings were allocated with the same room type and a comparatively small proportionate were not allocated with the same room as they reserved.
- The average ADR for the same room allocated as reserved was slightly higher than the rest. Thus, allocating the same room as reserved can aid in increased revenues.

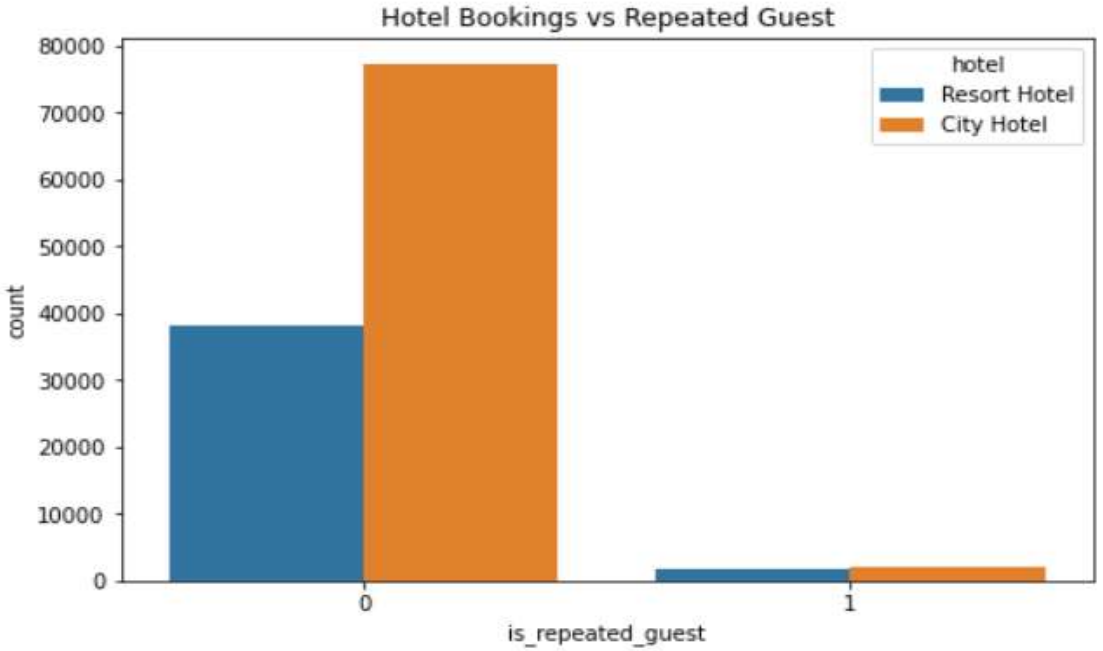


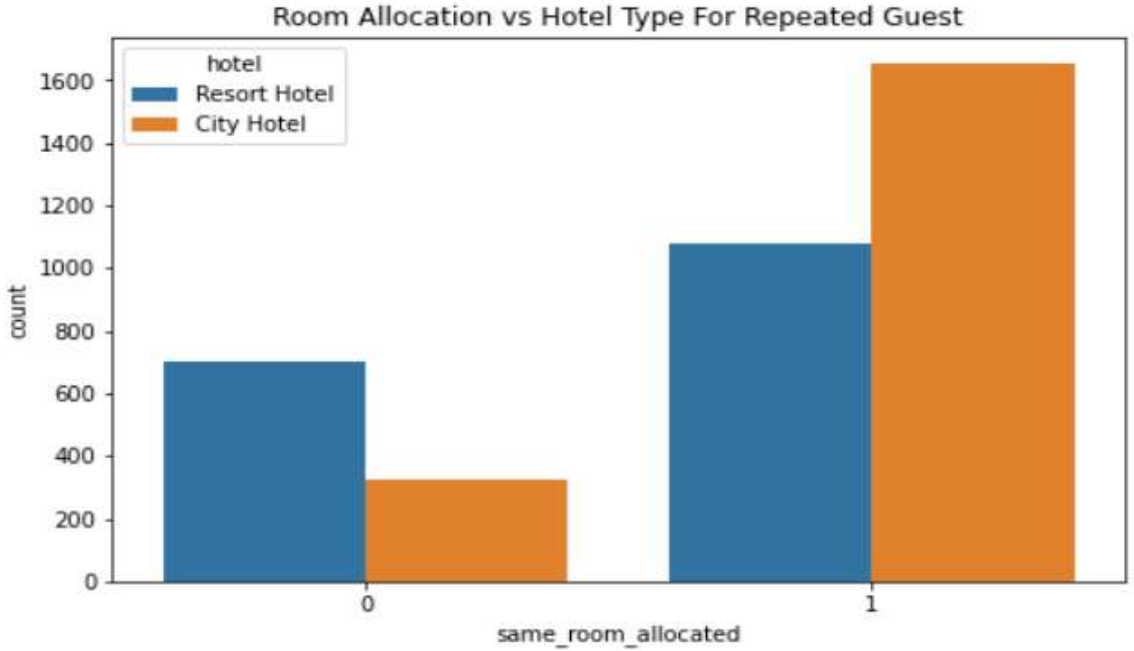


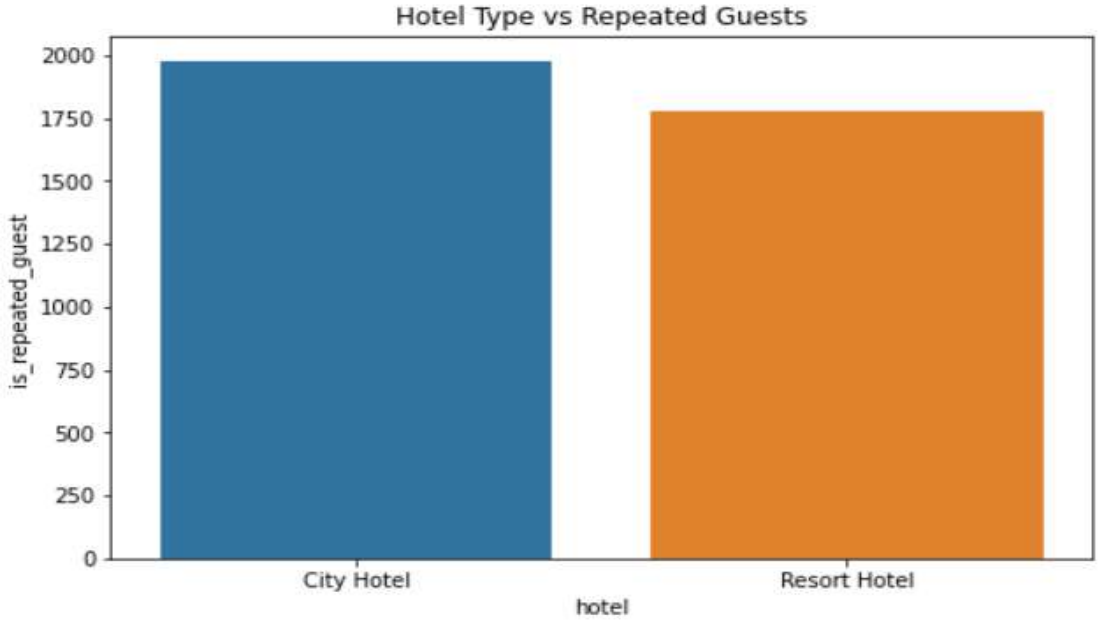


- Around 44,000 bookings were canceled from the year 2015 to 2017.
- Out of the total bookings done a larger number of cancellations were made for city hotels when compared to resort hotels constituting around 42%.
- Out of the total cancellations done a larger number of cancellations were made for city hotels when compared to resort hotels constituting around 75%.
- The most number of bookings were made by 2 people (couple) in both the type of hotels. This is followed by the number of bookings made for 1 person.
- There were no booking changes done for most of the bookings for both the city and resort hotels. Less than 10000 bookings were made with 1 change

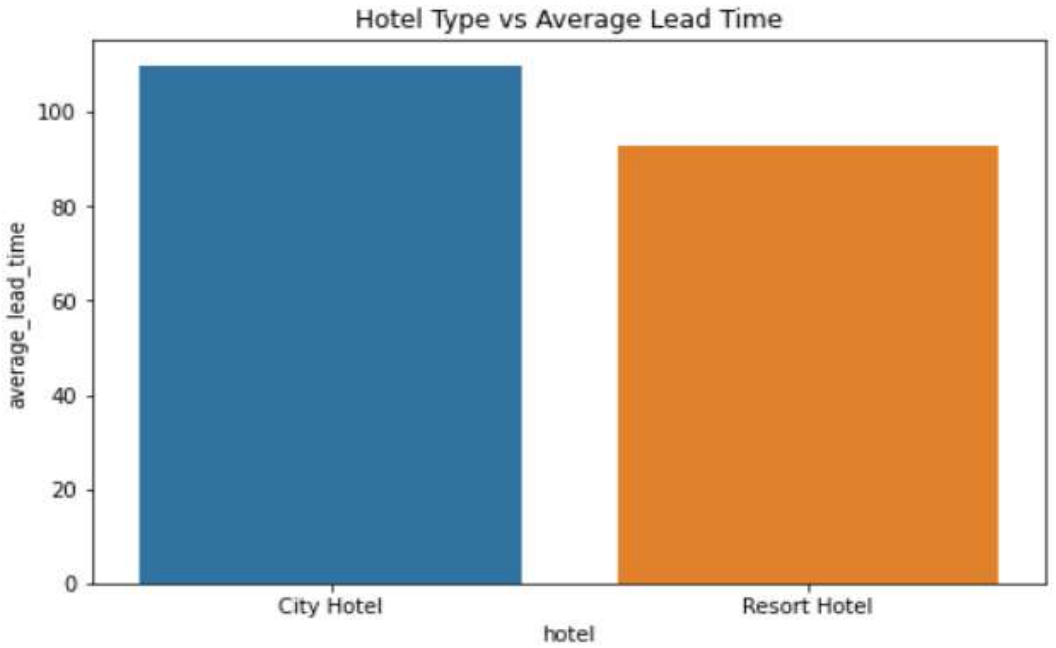


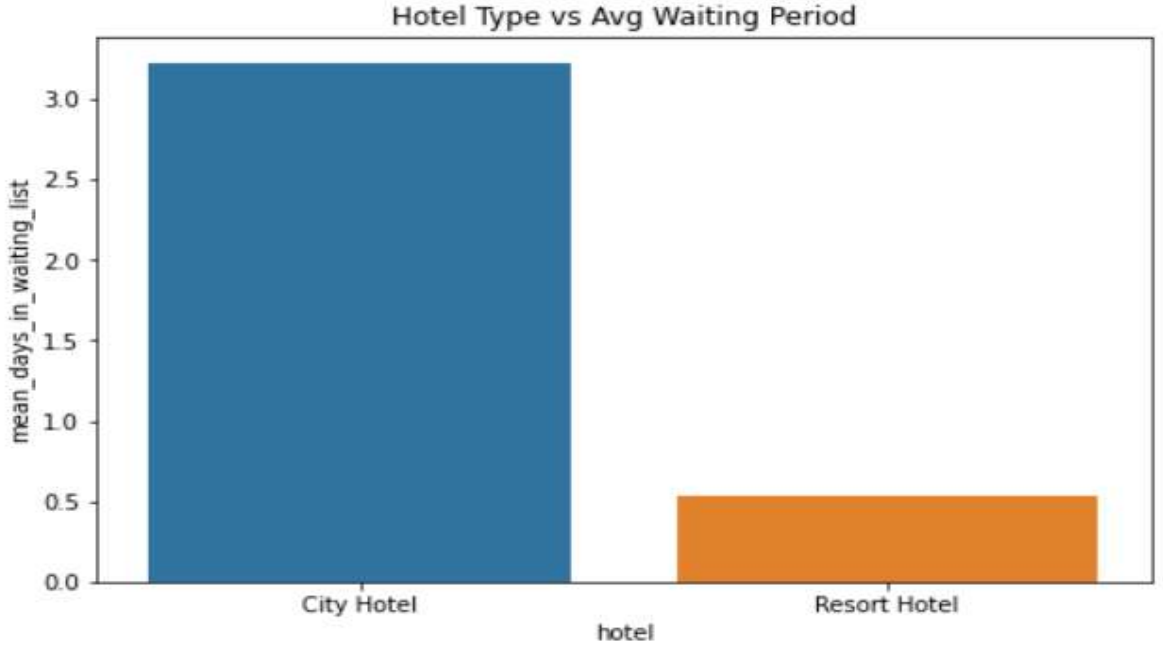




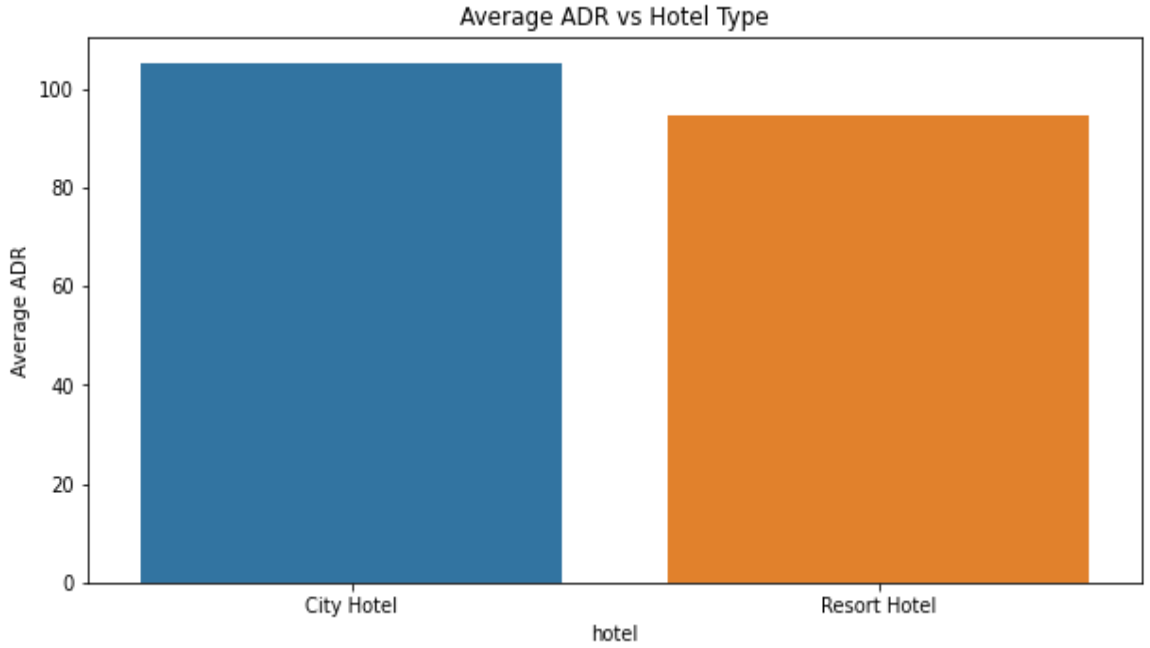


- A minimal proportionate of bookings constituted repeated guests for both city and resort hotel types.
- Among the repeated guests, the city hotel recorded a slightly higher number of bookings from previous guests when compared to resort hotels.
- For a greater proportion of repeated guests, for both the hotel types, it could be estimated that the same room is allocated as they reserved. Hence this could be a reason for customer retention.
- From the analysis, it was also observed that customer retention was not dependent on factors such as amenities namely car parking spaces, allowance of special requests, and booking changes.

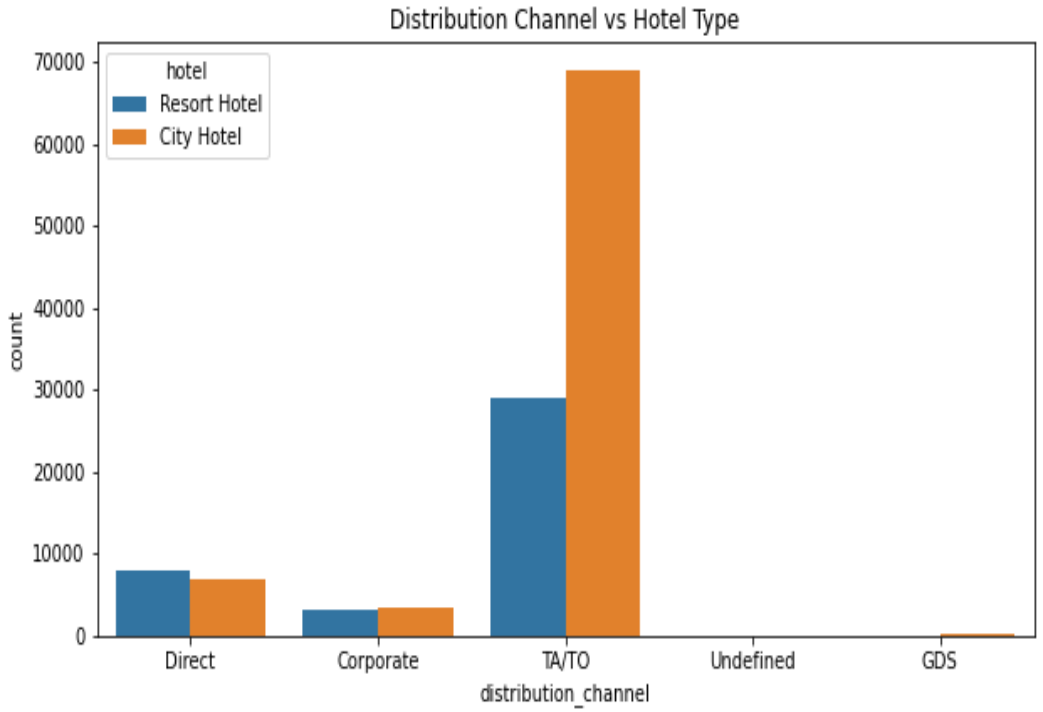


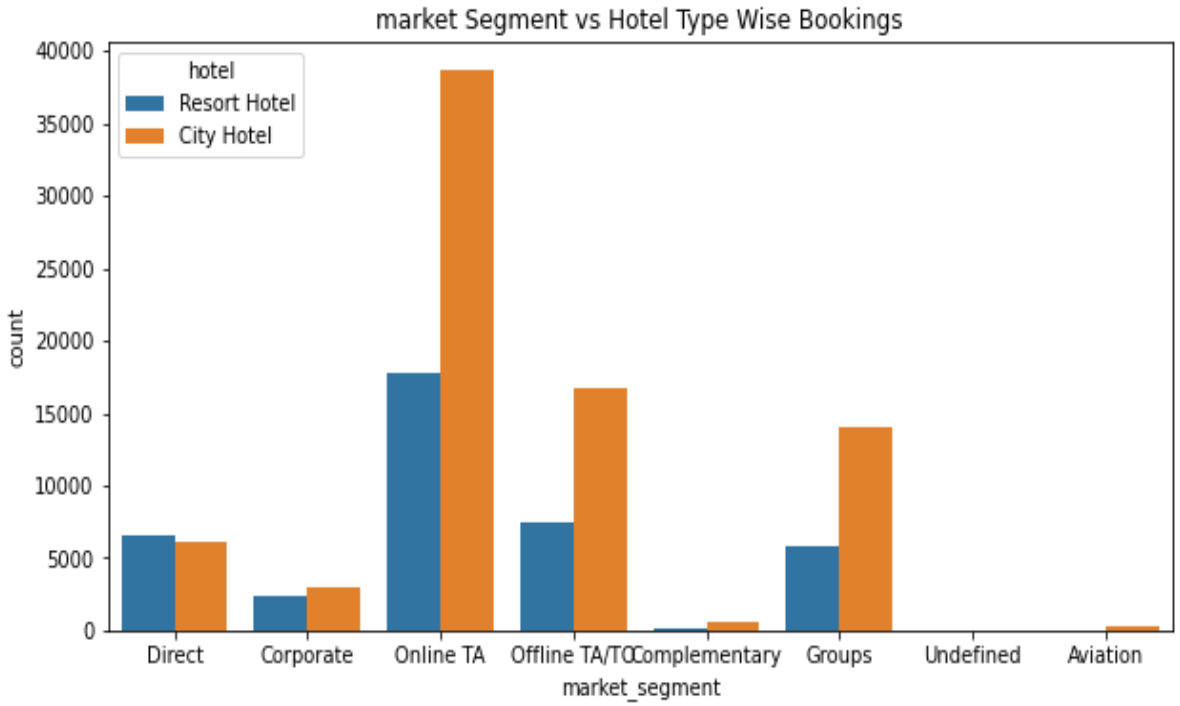


- The city hotel type has a greater average lead time when compared to the resort hotel type. Also, the median lead time is significantly higher for both hotels, which means customers generally plan their hotel visits way early.
- For city hotel type, both the max and the average waiting period were greater than resort hotels.
- The greater lead time and waiting period for city hotels represent that a city hotel has greater demand and is busier than a resort hotel.
- The city hotel has a higher average ADR than a resort hotel.



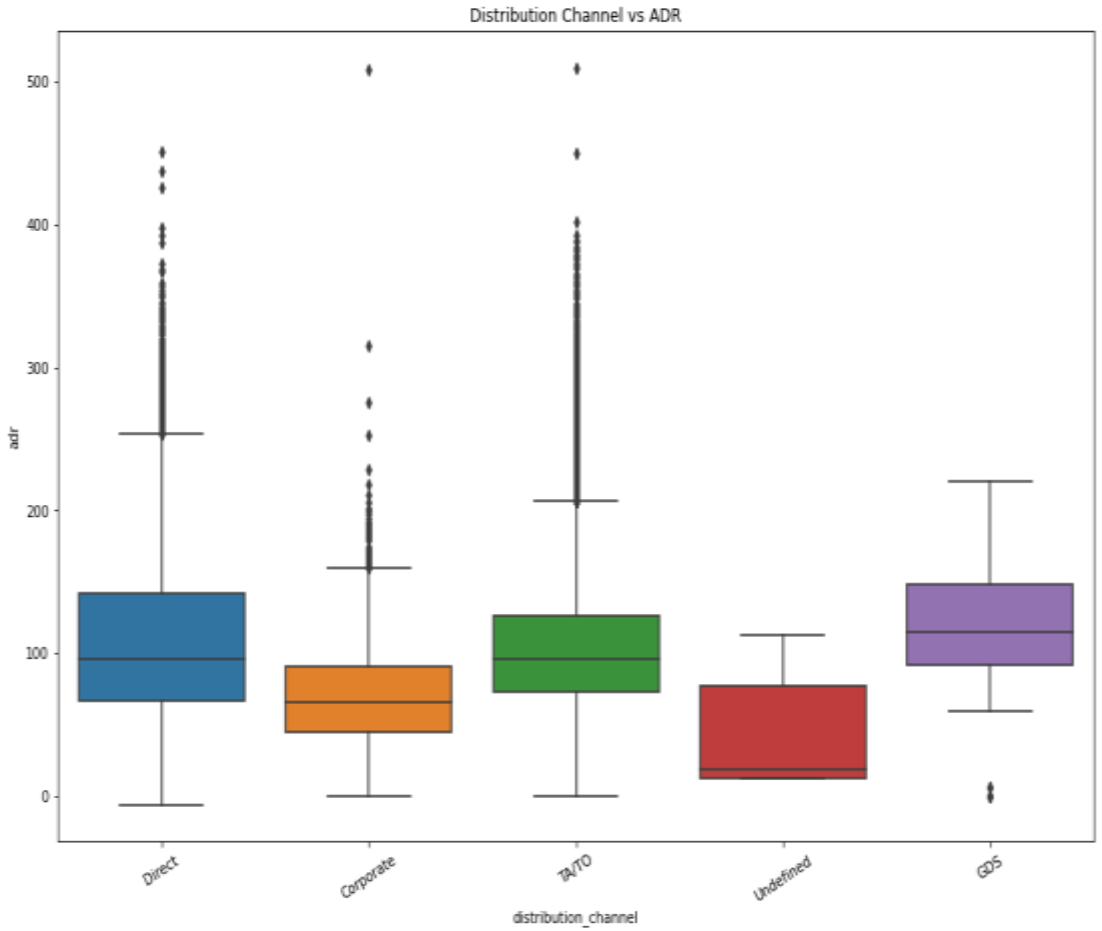
Distribution channel and Market Segment Analysis

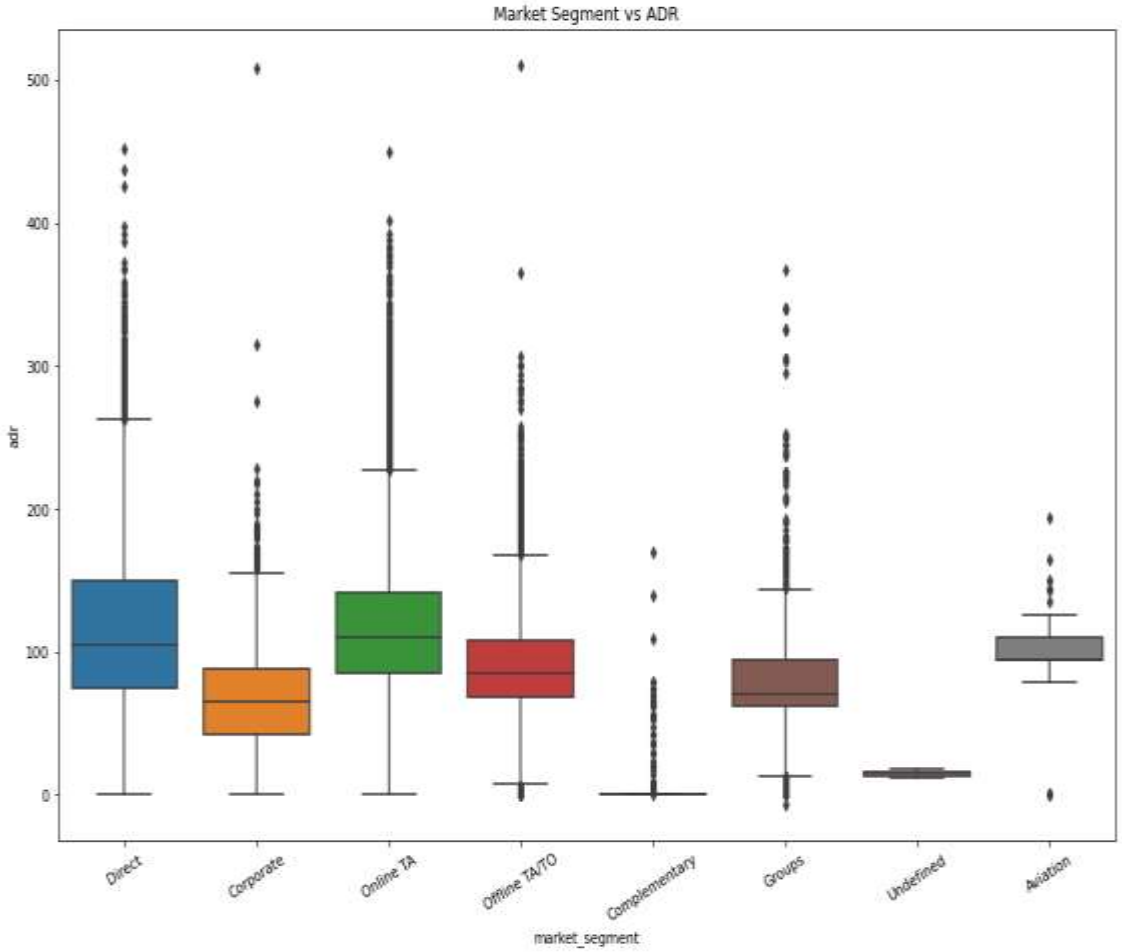




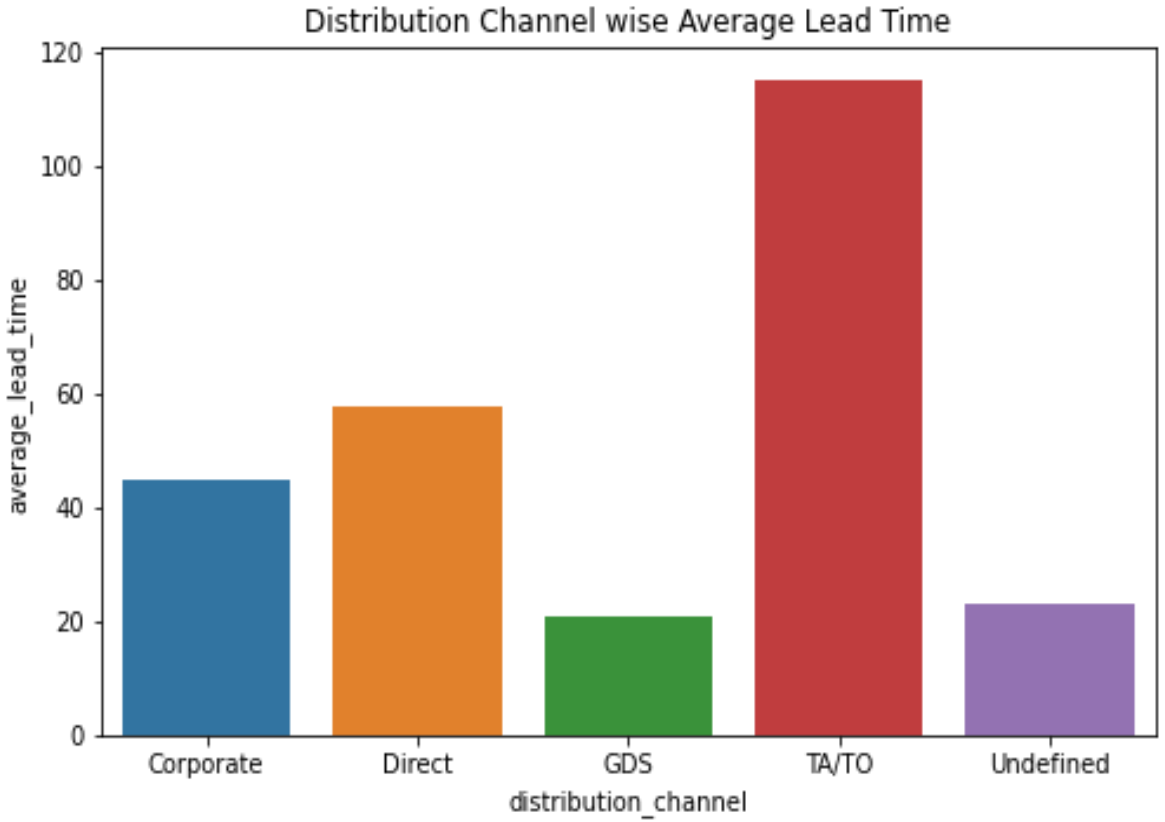
- The most bookings made were through Travel Agents (TA) and Tour Operators (TO). The second most demanded distribution channel was direct booking, followed by corporate booking.

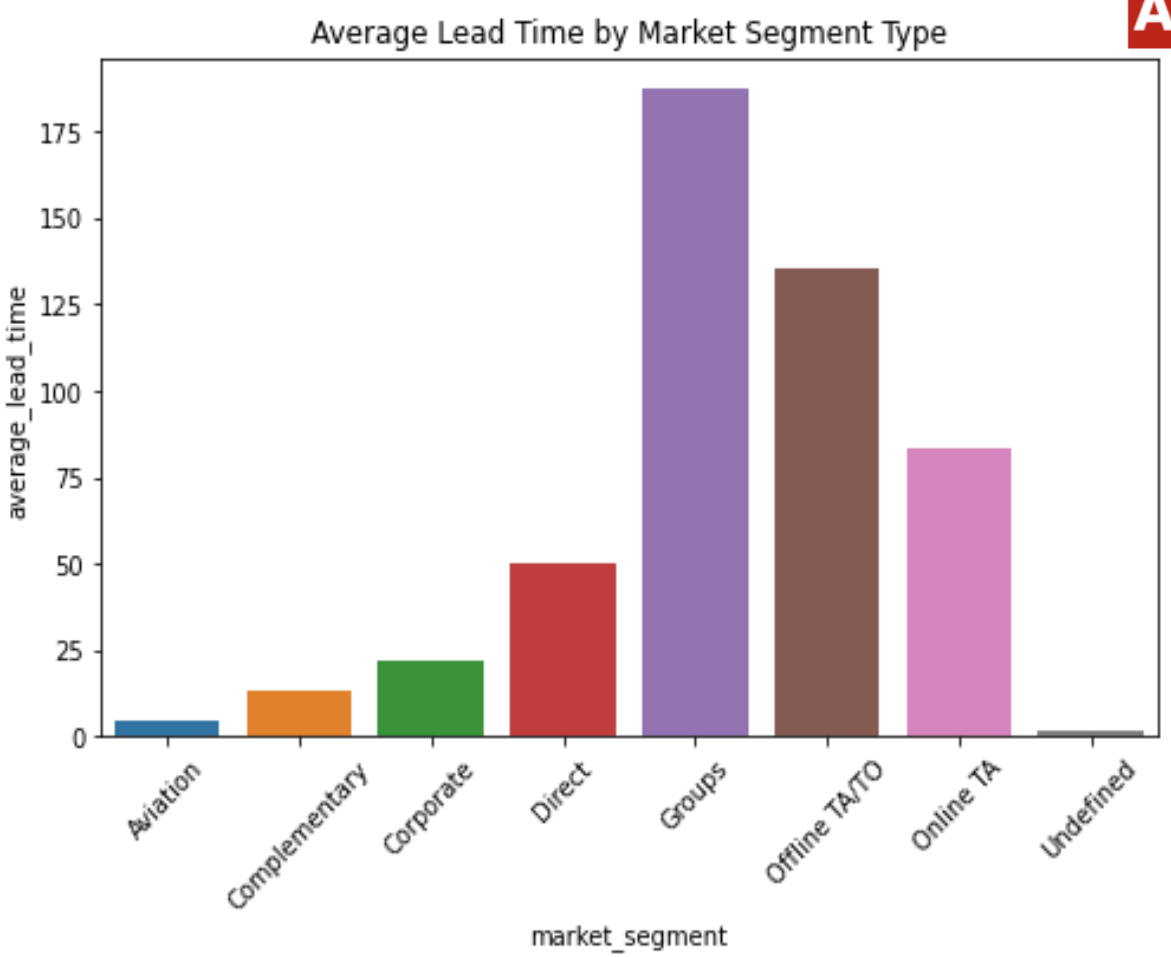
- Online Travel Agents is the market segment with the most number of bookings from the dataset for both types of hotels. This is followed by Offline Travel Agents(TA)/Tour Operators (TO), groups, and direct customers.





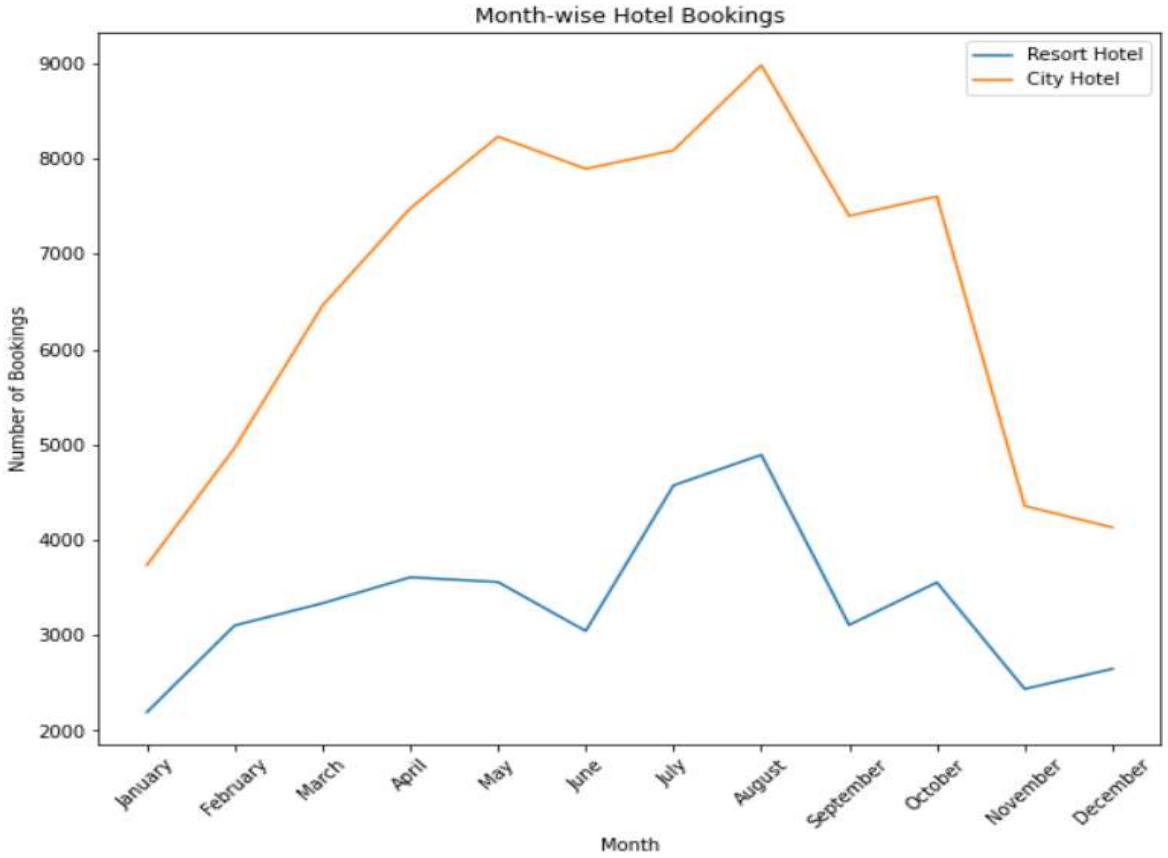
- The distribution channels generating higher ADRs were direct, GDS (Global Distribution Systems), and Travel Agents (TA)/Tour Operators (TO) channels of distribution.
- Hotels can focus on the GDS channel to generate more revenue.
- Online Travel Agents and Direct segments are the market segments that were generating higher ADR

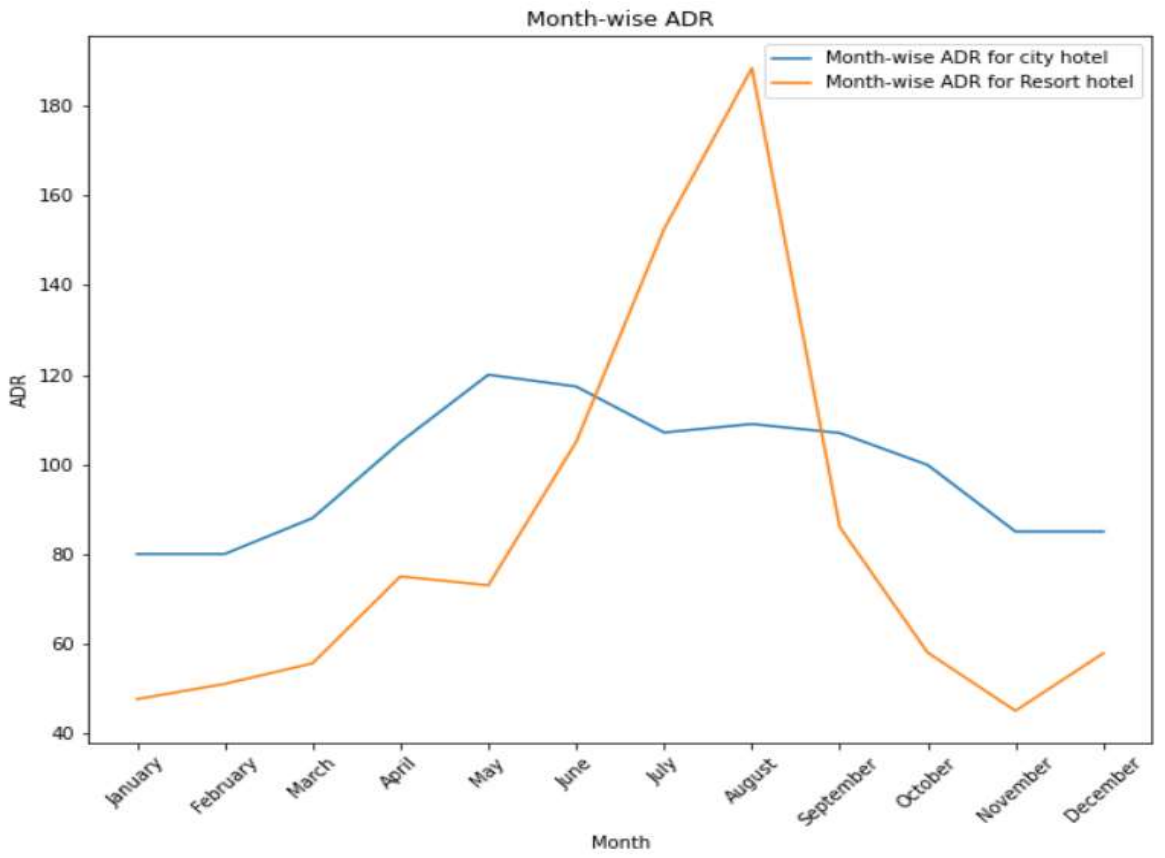




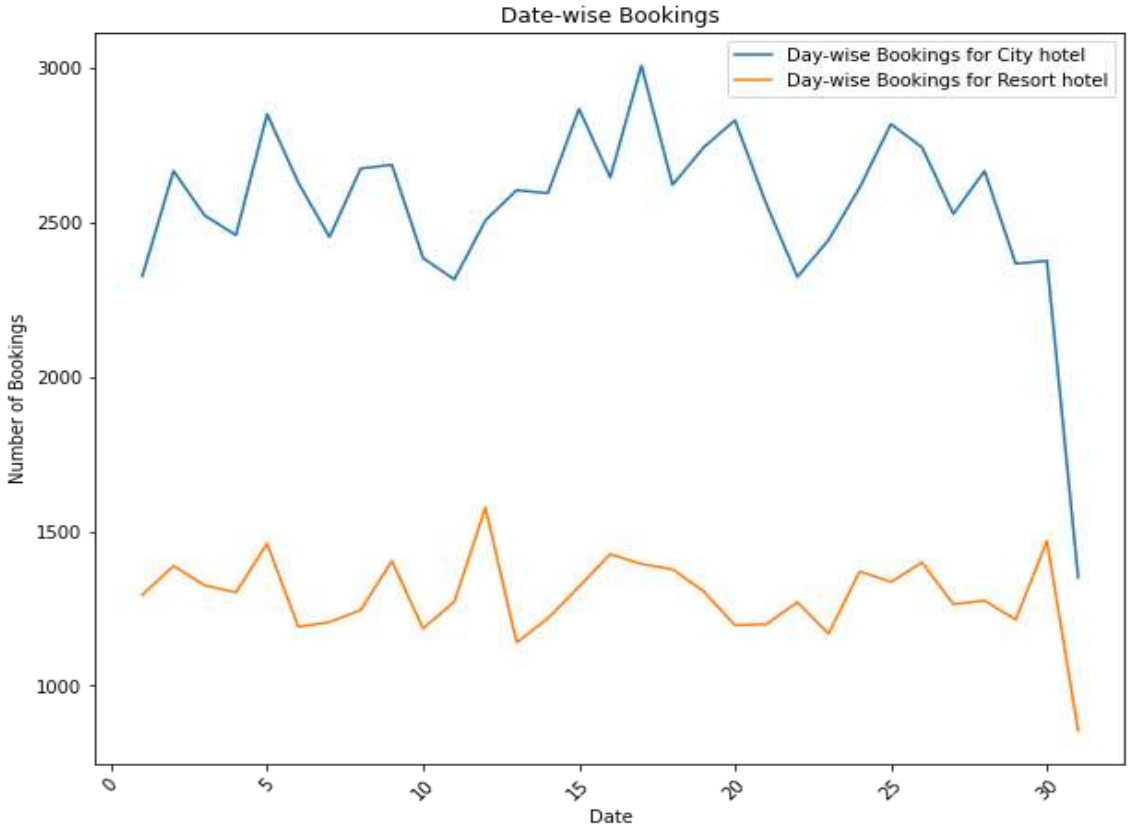
- The average lead time is higher for the bookings made via TA/TO distribution channel.
- The market segments with higher average lead time are groups, offline TA/TO, and online TA.
- Thus the bookings made via TA/TO channel are booked way ahead of the actual arrival date.

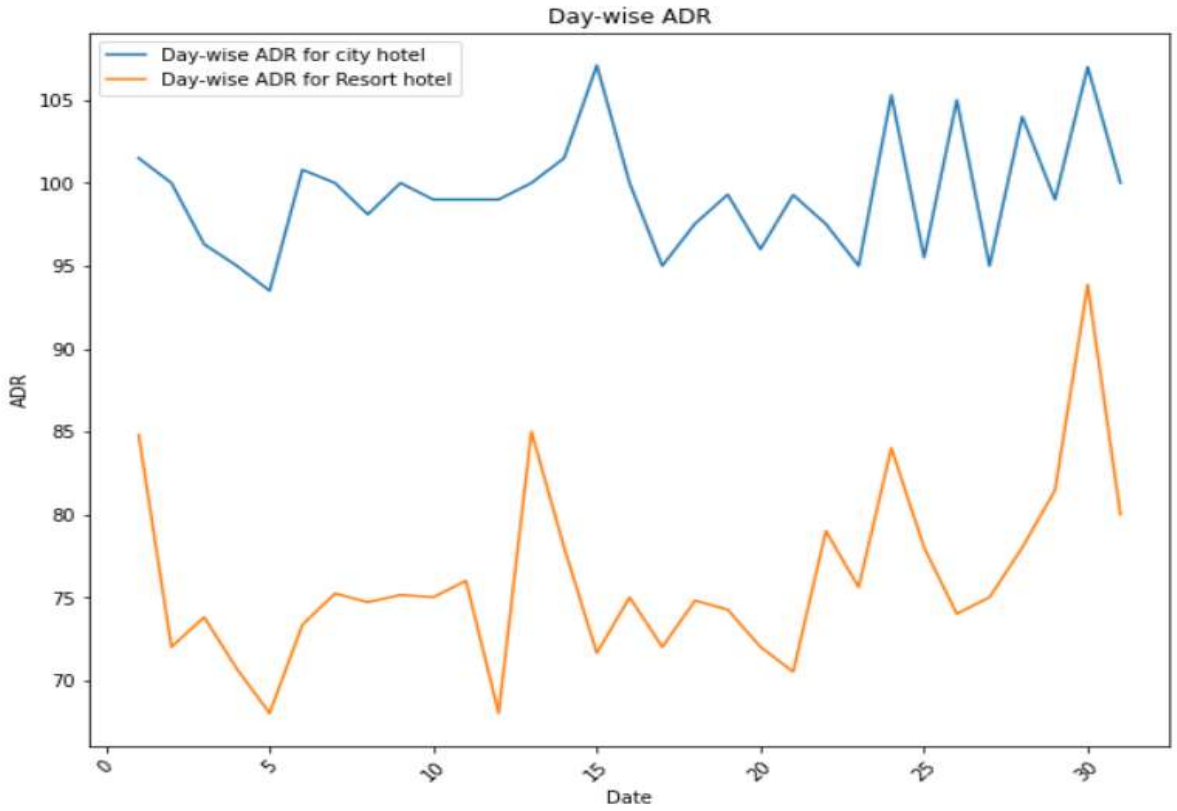
Analysis Based on Time Period





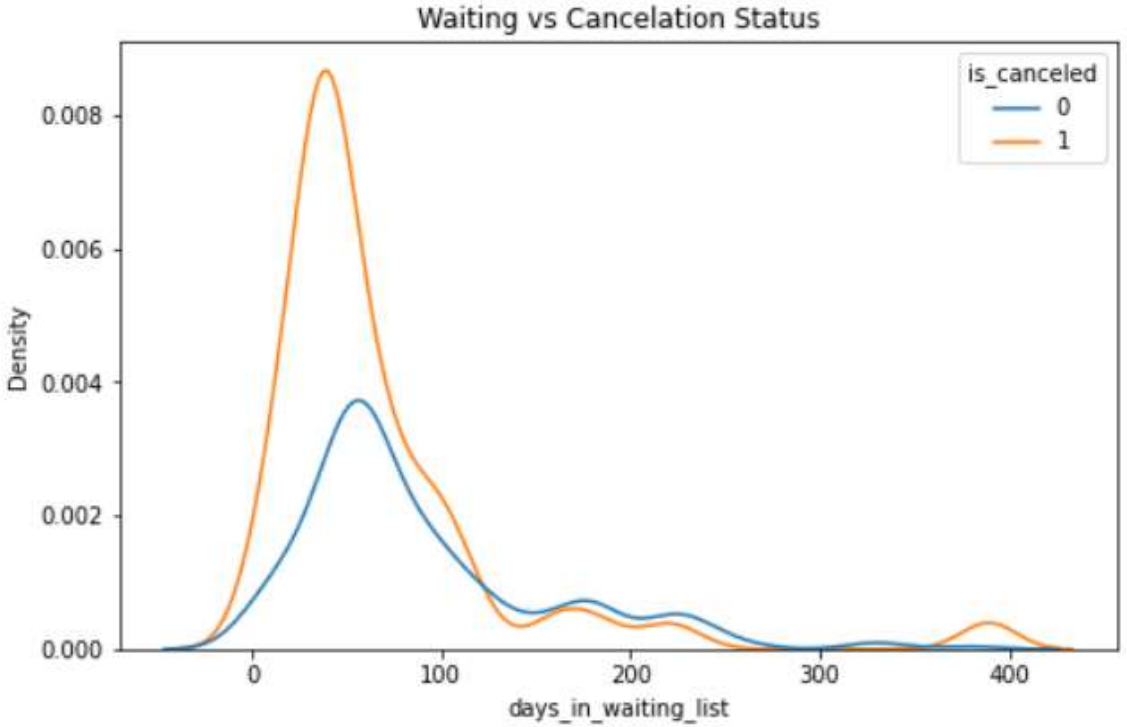
- According to the dataset, for both city and resort hotels, August month witnessed the highest number of hotel bookings. The hotel bookings gradually increased from January to reach the peak in August and then gradually declined till December.
- According to the dataset, for resort hotels, the month of August witnessed the highest Average Daily Rate (ADR). The ADR witnessed a sudden spike from May to August and then declined till November.
- According to the dataset, for city hotels, the month of May witnessed the highest Average Daily Rate (ADR). The ADR gradually increased from January to May and then declined till December.

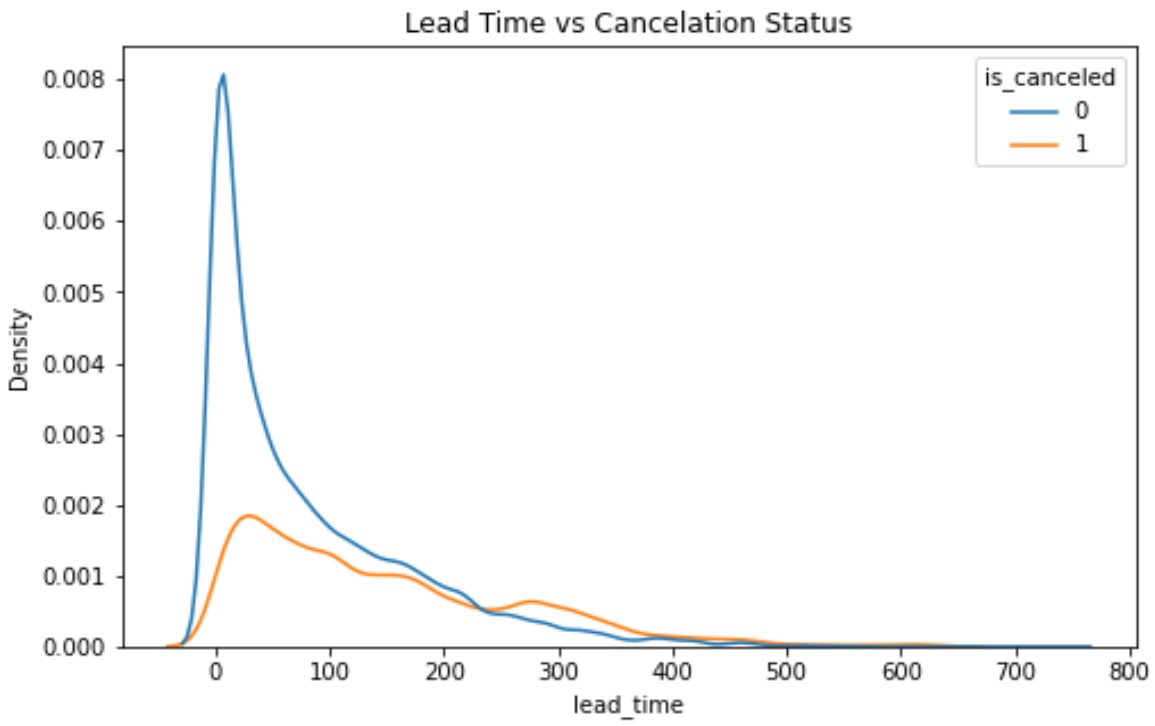




- According to the dataset, for both city and resort hotels, small peaks at regular intervals of days can be seen which can be due to an increase in arrival during weekends.
- For resort hotels, the maximum ADR is observed at the end of the month, which represents that most of the hotel bookings were made for the end of the month.
- For city hotels, the maximum ADR was observed both in the mid and at the end of the month, representing the demand at both times of the month.

Analysis Based on Cancellation of Bookings

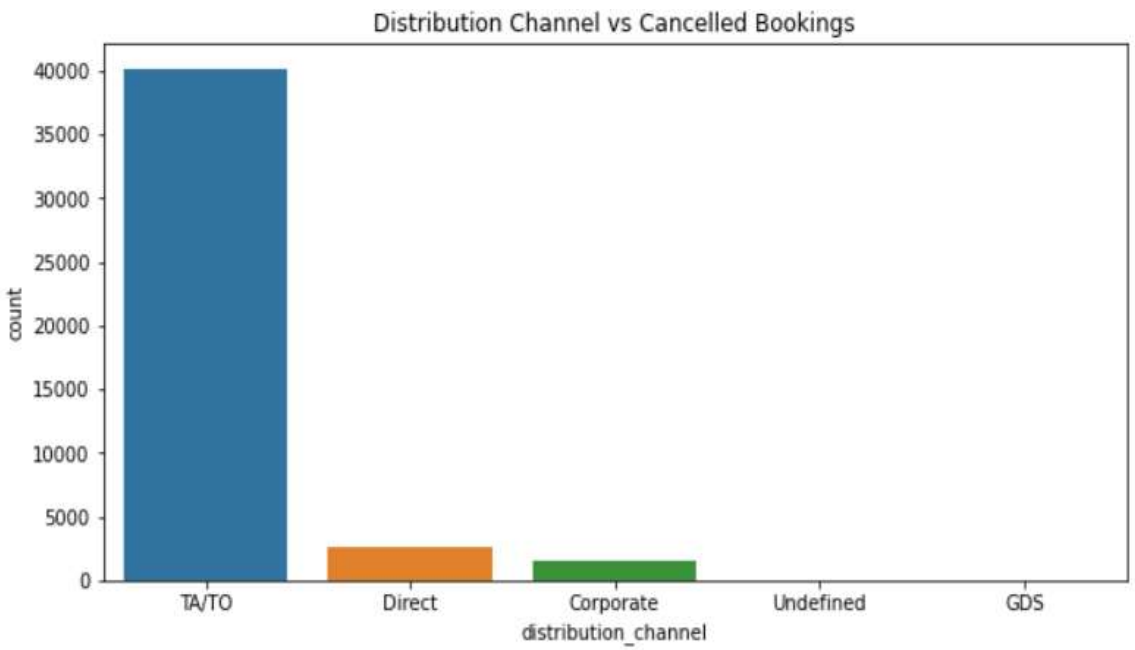


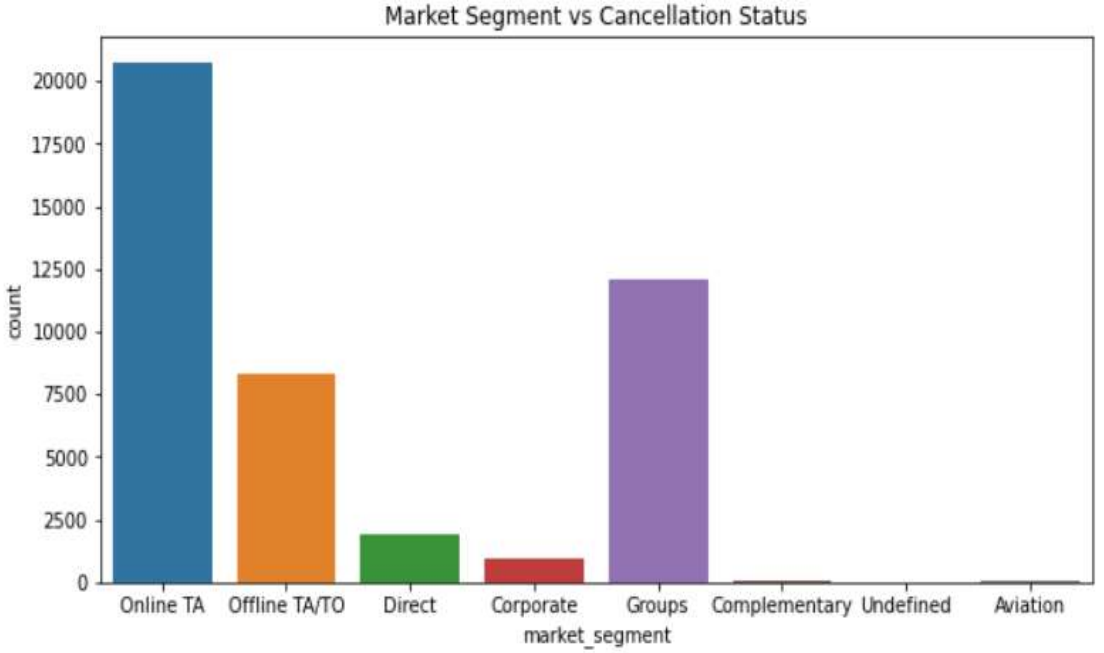


- As the waiting period is increasing, both the number of canceled and not canceled bookings distribution curves are similar. Hence, it is estimated that the waiting period has no role to play in the cancellation of hotel bookings for both hotel types.

- As the lead time is increasing, both the number of canceled and not canceled bookings are decreasing. Most of the bookings that got canceled have a lead time of fewer than 300 days and most of the bookings that are not canceled also have a lead time of fewer than 300 days.

- Hence, lead time has no role to play in the cancellation of hotel bookings for both hotel types.



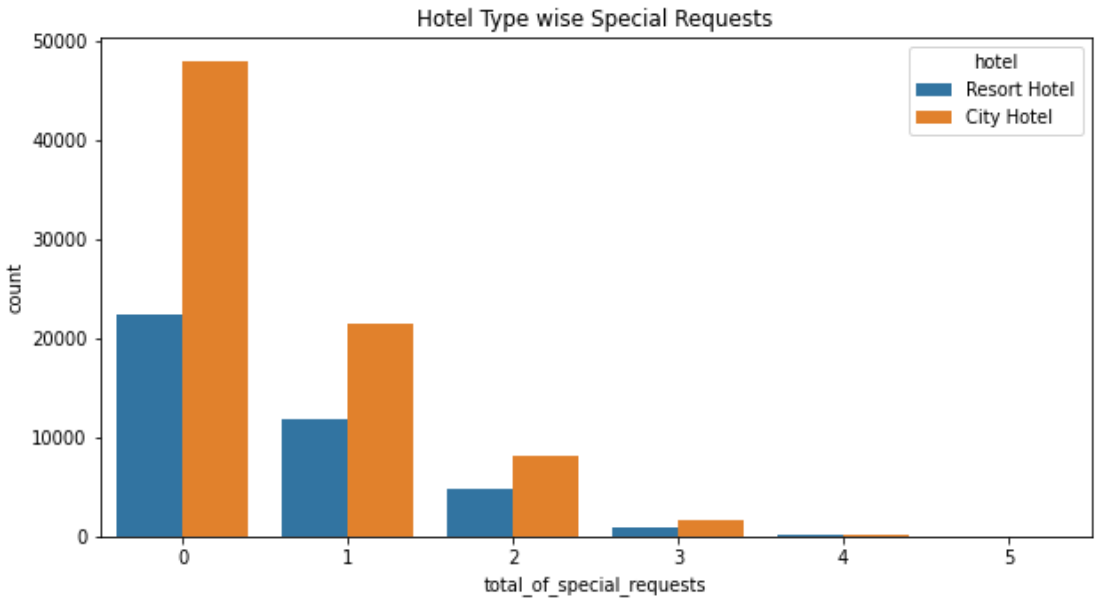


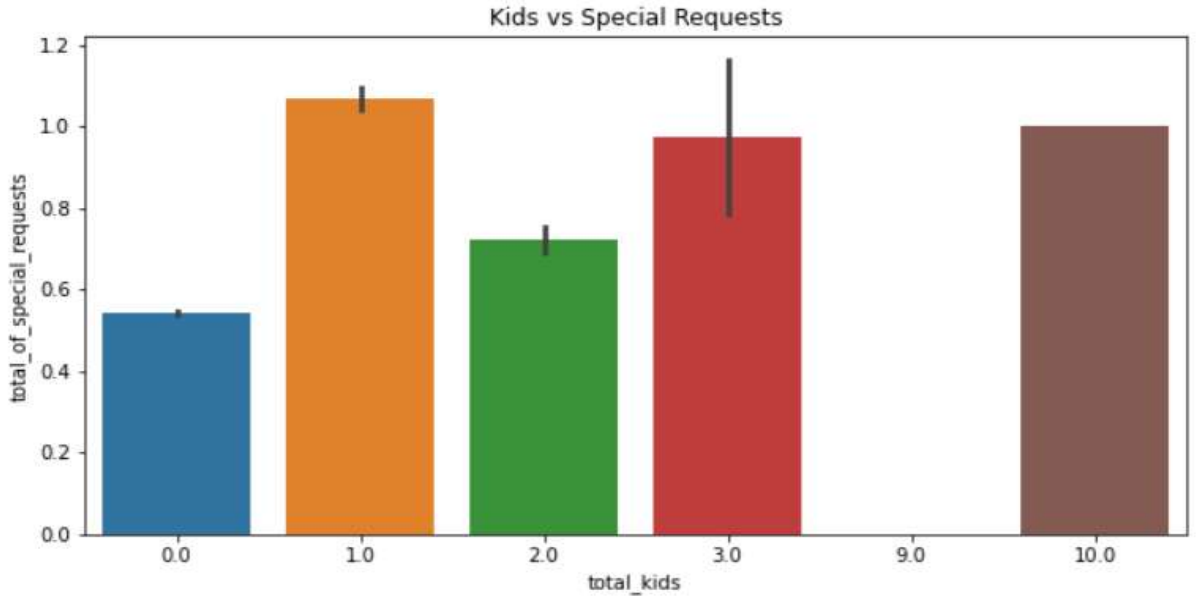
- Travel Agents (TA)/Tour Operators (TO) is the distribution channel that has experienced more cancellations from their total bookings, followed by Direct bookings and Corporate bookings.
- Online Travel Agents (TA) is the market segment that has experienced more cancellations from their total bookings, followed by groups.
- From the analysis, it was also observed that the majority of the bookings that were canceled belong to the no-deposit segment. But almost 99% of the bookings with a non-refundable deposit type experienced cancellations. Hence, the deposit type has no role to play in the cancellation of bookings.

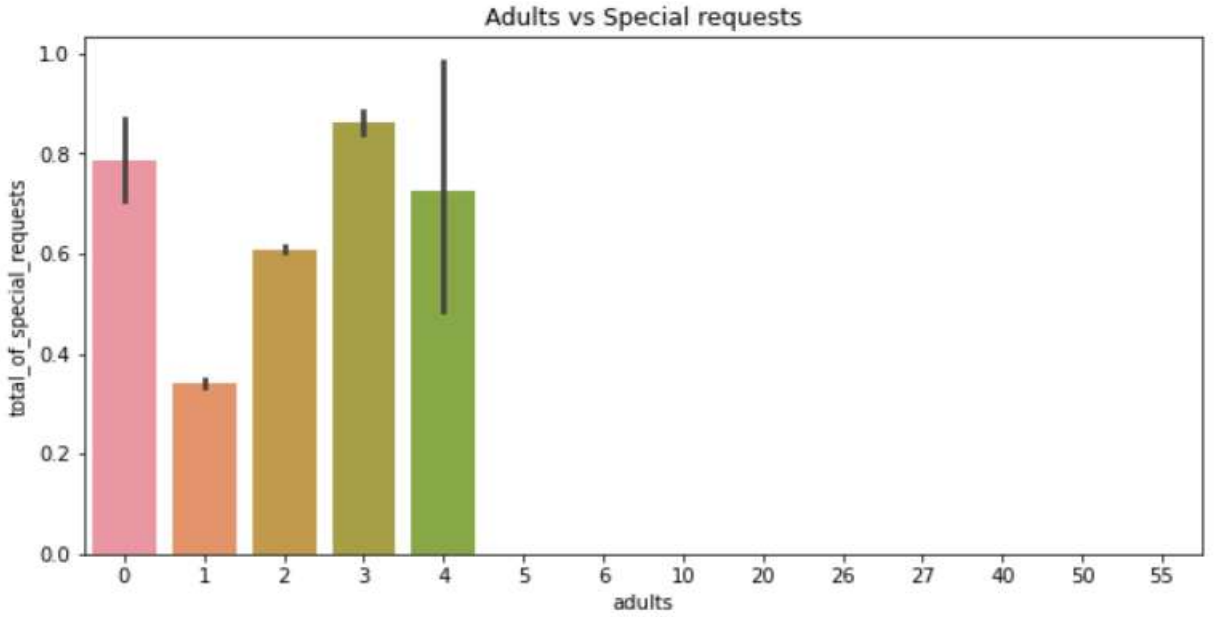
Table 4: The cancellation % pertainingg to bookings with different deposit types

Deposit Type	Total Cancellations	Cancelled % among total deposit type cancellations	Total Bookings	Cancelled % among bookings made
No Deposit	29669	67.13 %	104461	28.0 %
Non Refund	14493	32.79 %	14586	99.0 %
Refundable	36	0.08 %	162	22.0 %

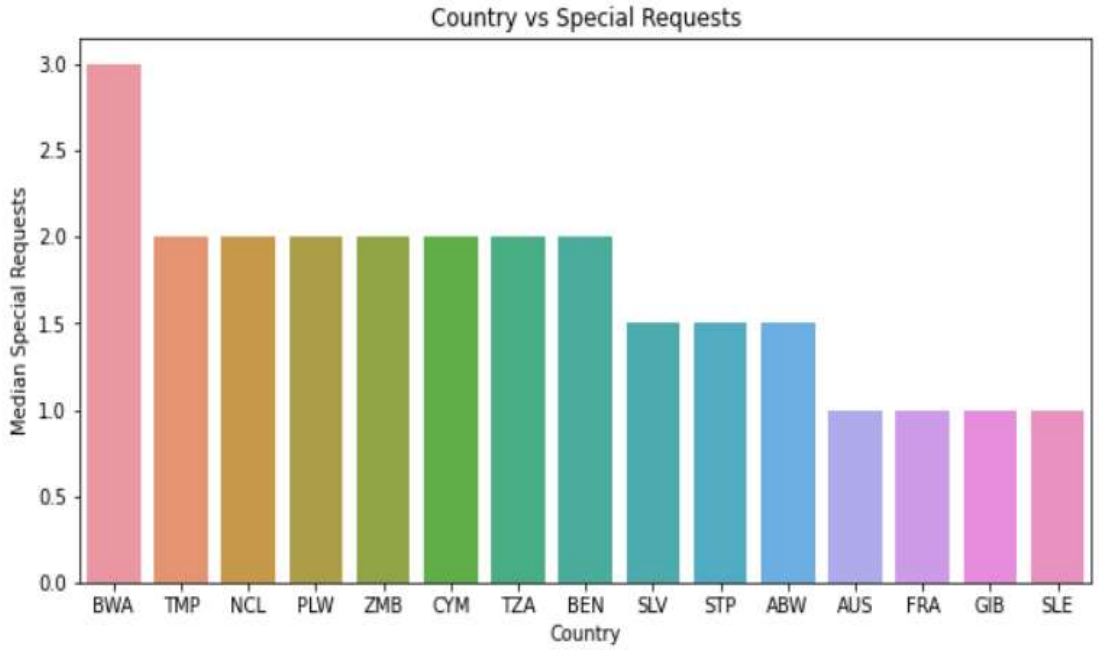
Analysis Based on Special Requests

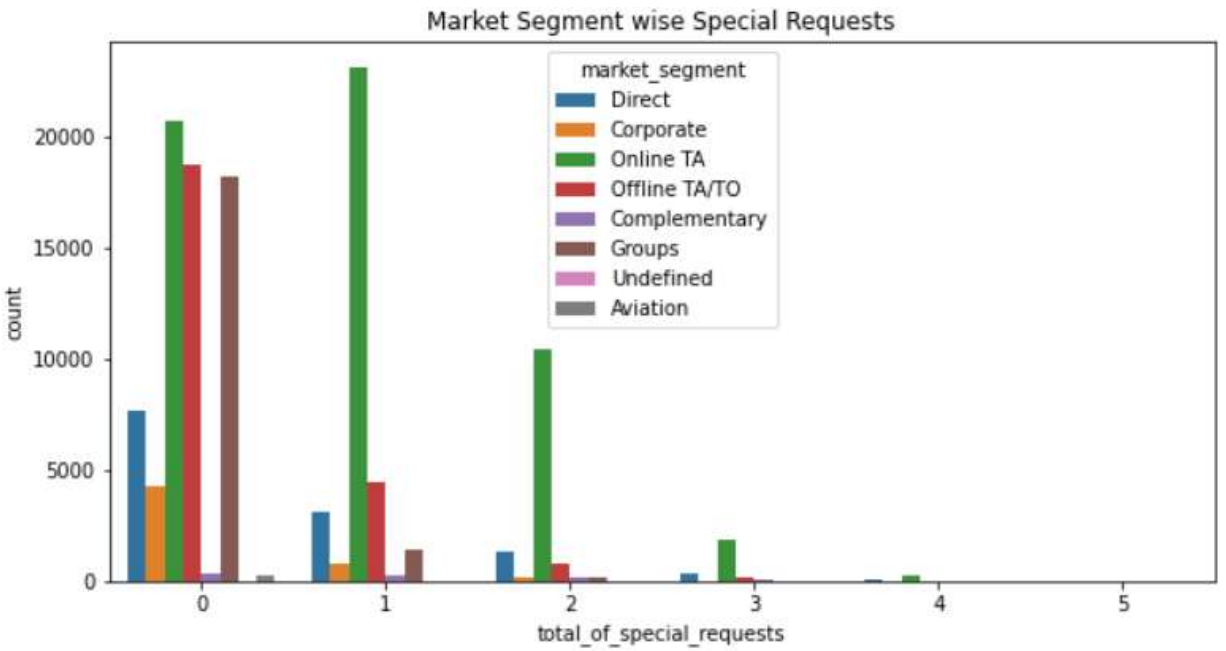






- The majority of the bookings were made with no special requests for both hotels.
- From the bookings with special requests, the city hotel has a greater number of bookings with special requests.
- Almost all of the kids' segments had special requests. Hence the presence of kids (children, and babies) could be the reason for the requirement of special requests.
- Adults greater than 2 have a higher distribution of bookings with special requests.

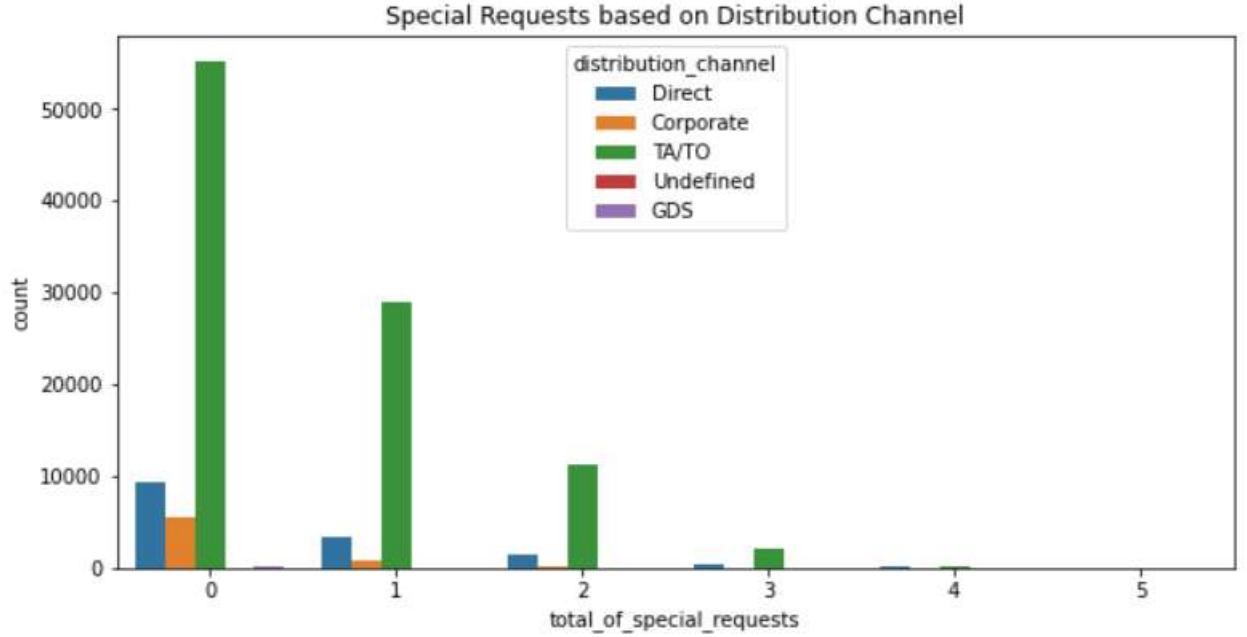


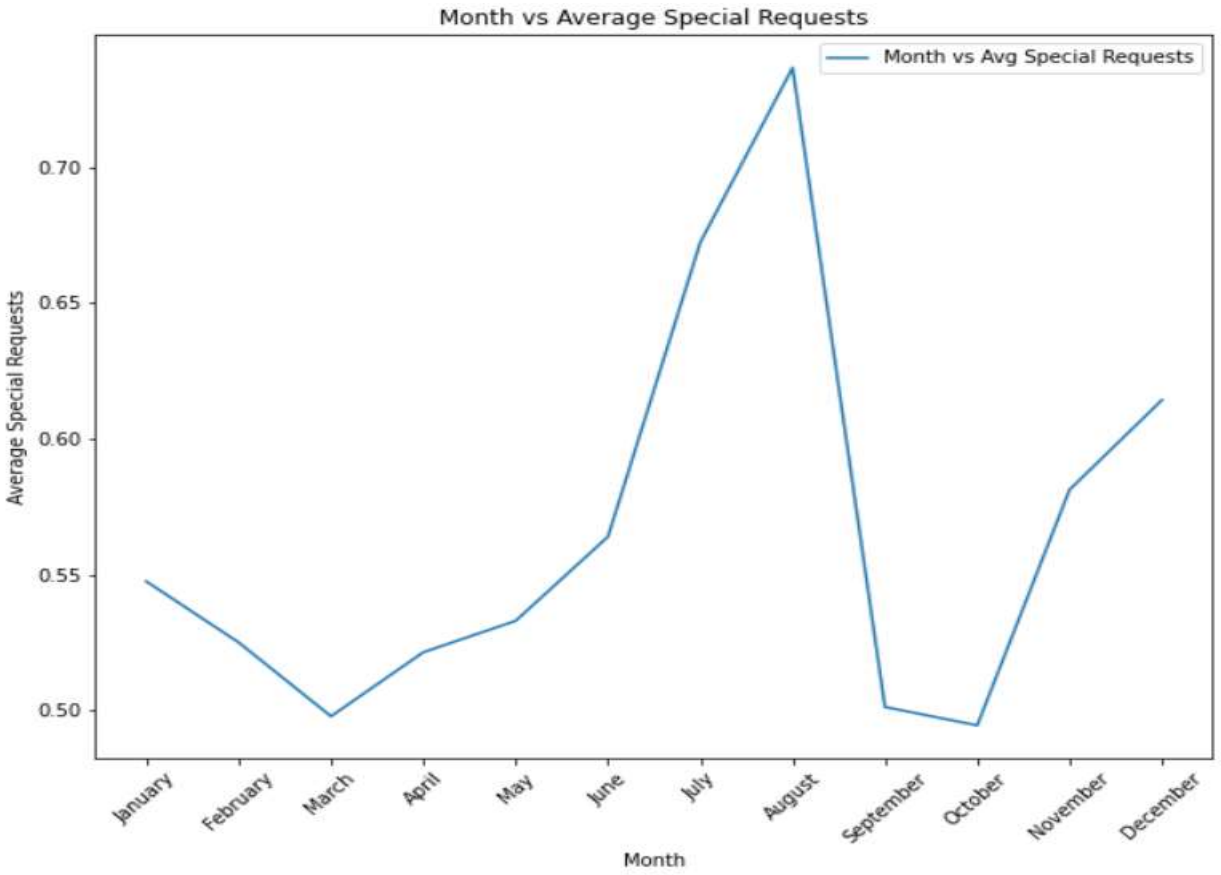


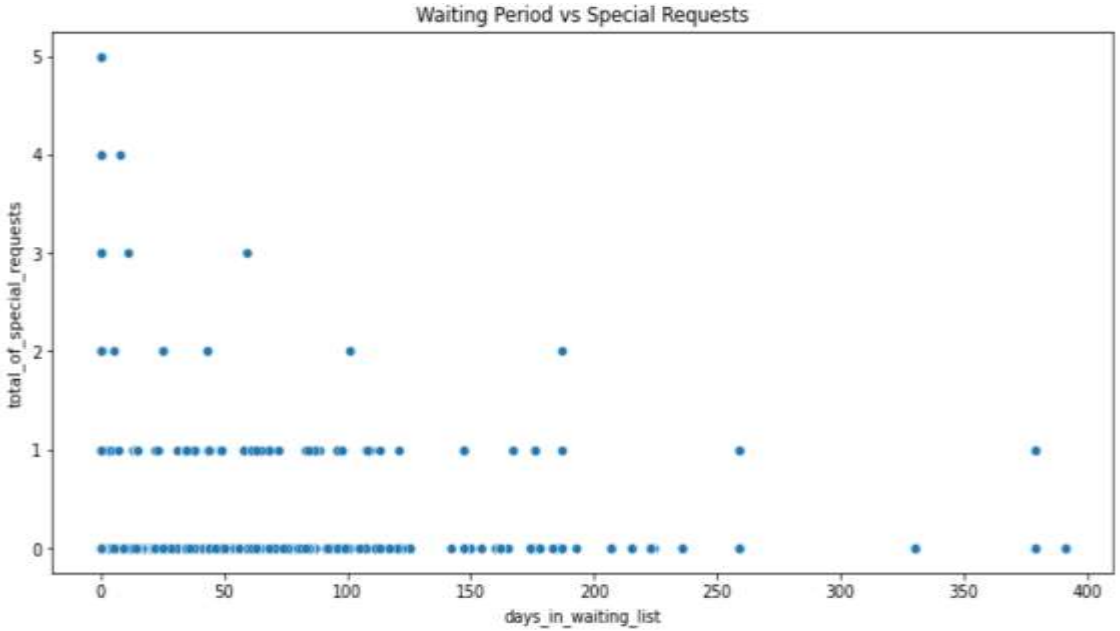
- The higher average number of special requests were received mostly from the African countries namely Botswana with an average of 3 requests, followed by countries such as Zambia, Tanzania, Benin, and more.

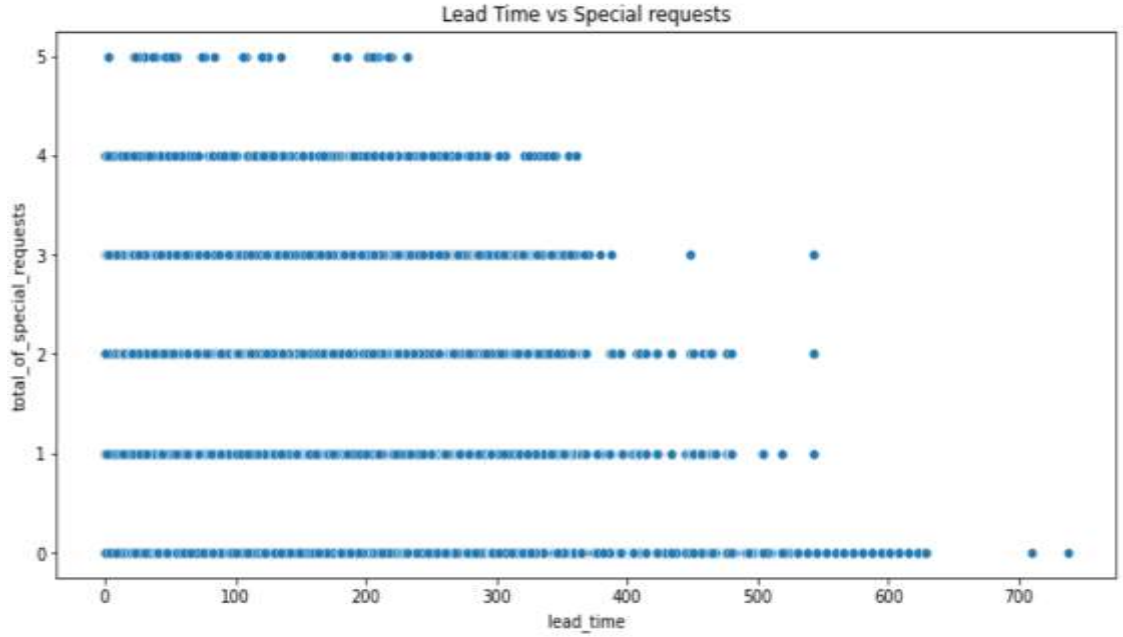
- Although, for all the segments majority of the bookings had no special requests, except for the Online Travel Agents segment which had a majority of bookings with 1 special request.

- Although, for all the distribution channels majority of the bookings had no special requests. Among the special requests made Online Travel Agents/ Tour Operators (TO) channel experienced the majority of bookings with 1 special request.





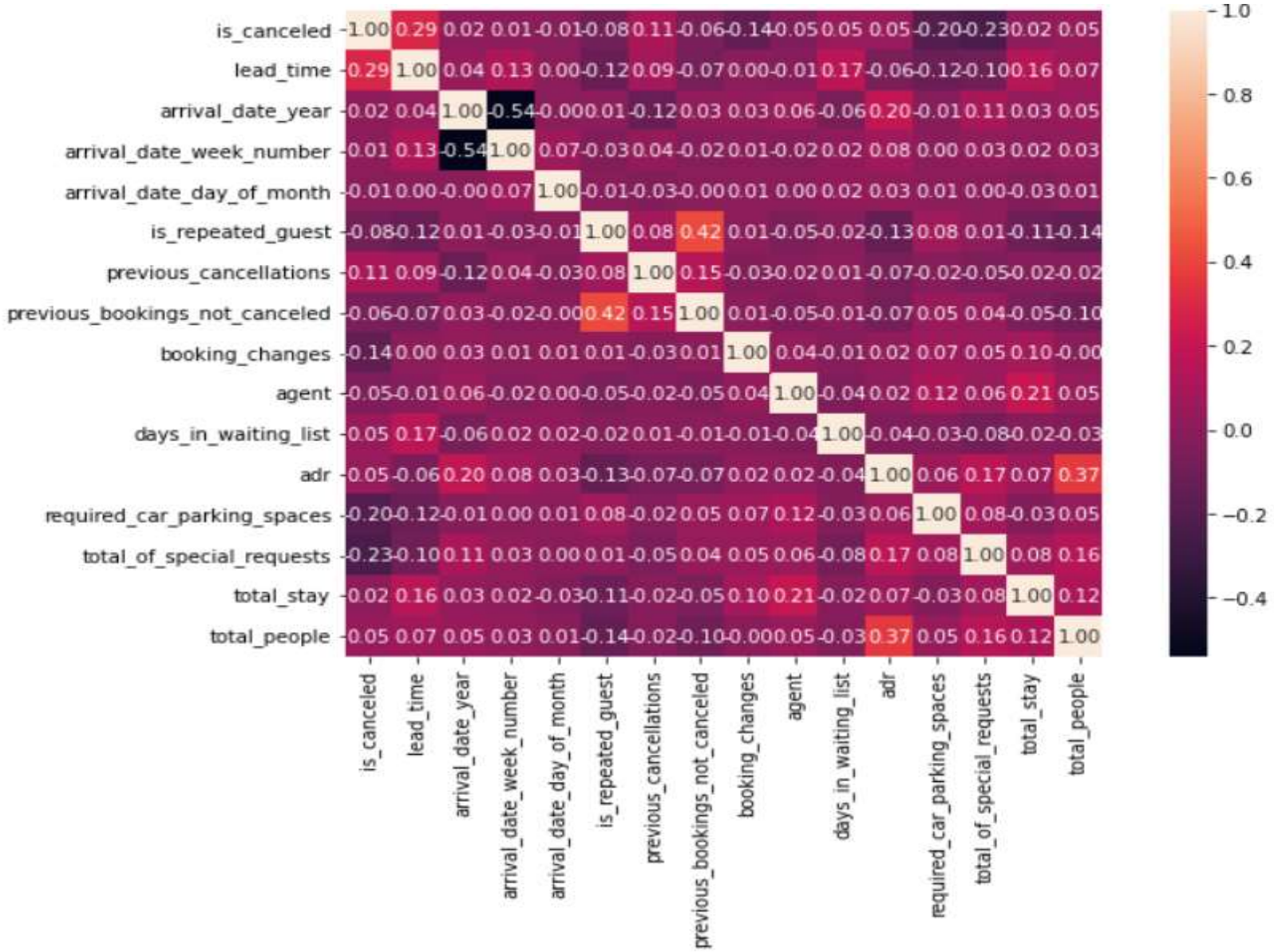




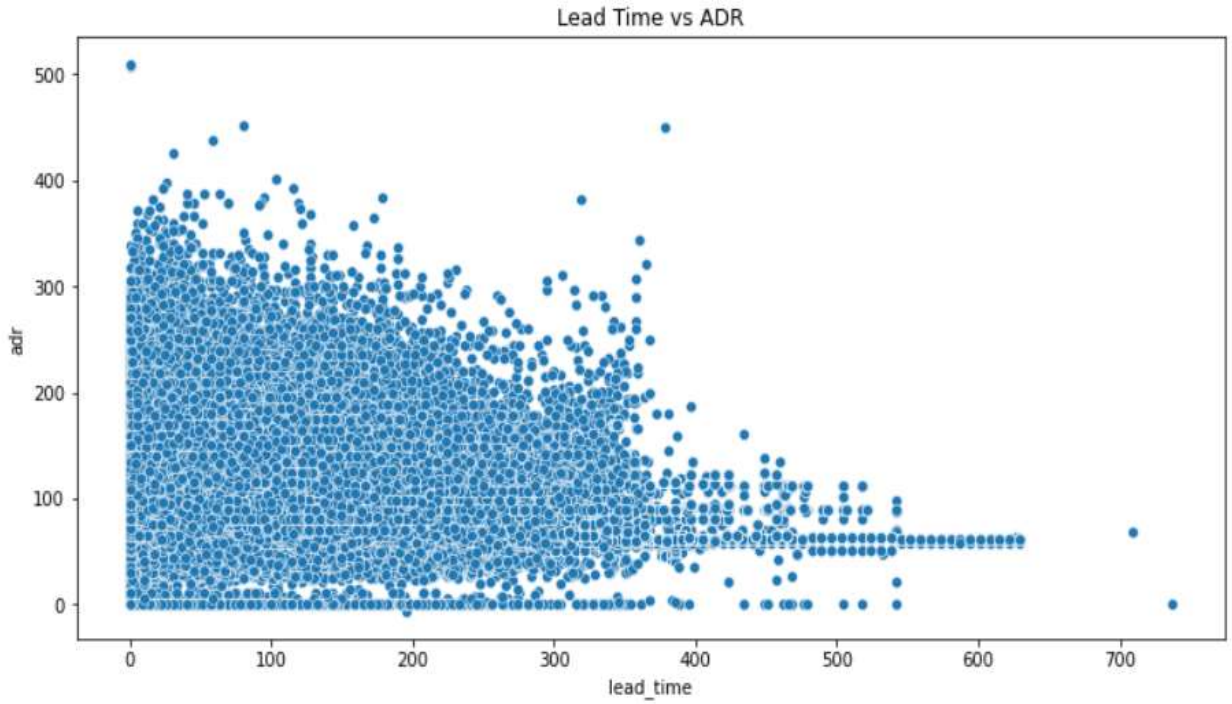
- The higher average number of special requests were received in August, and July with an average of >0.7 and >0.65 respectively.
- As the waiting period is increasing the number of special requests was declining. Most of the special requests are for bookings with less than a waiting period of 100 days.
- The greater the lead time is, the lesser the special requests received by the hotels. Lead time with less than 400 days had most of the special requests.

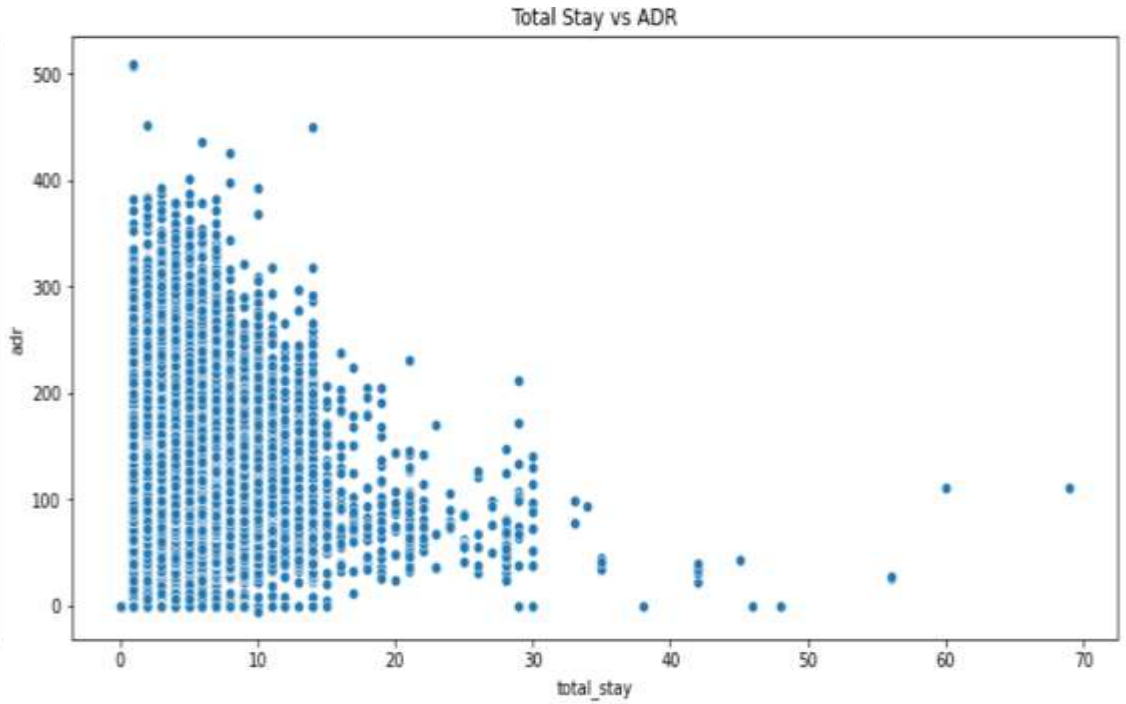
Correlation Heatmap

- is_canceled and lead time have a slight correlation. This means that the higher the lead time of the customer the higher the cancellation rate.
- is_repeated_guest and previous_bookings_not_canceled show a slight positive correlation as it is obvious that with no previous cancellations the guest would be a repeated one
- ADR is slightly correlated with total_people, which makes sense as more people means more revenue and therefore more ADR.

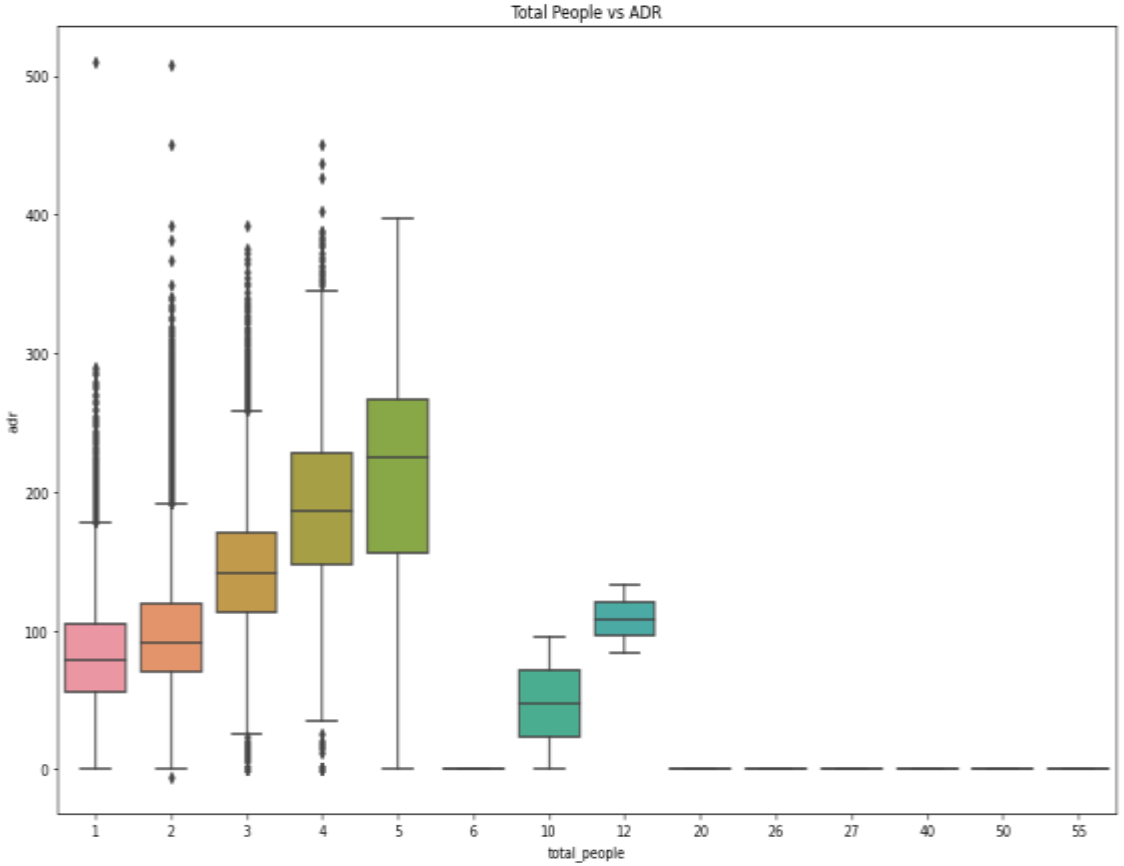


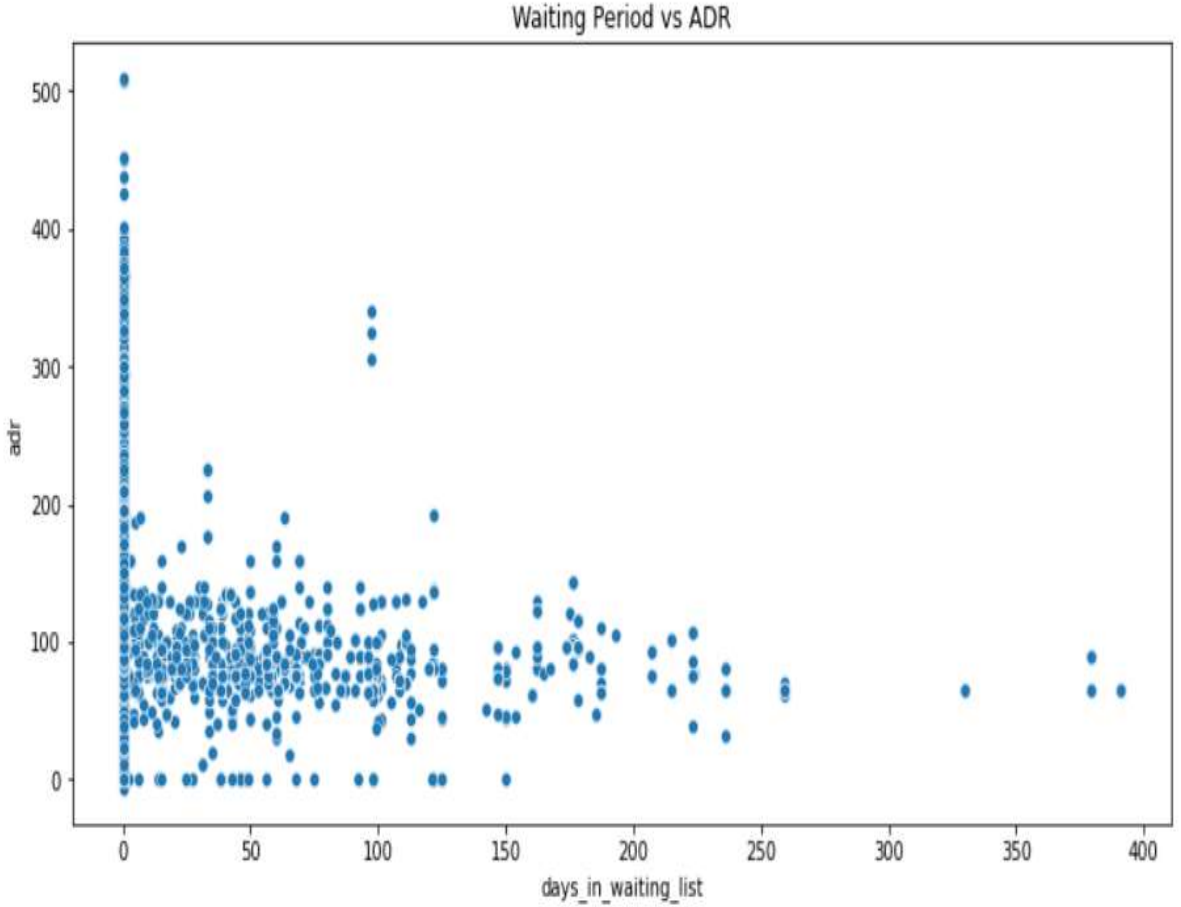
Optimal Stay for Customer





- ADR is decreasing as the lead time, and total stay length are increasing.
- Most of the bookings were made with a lead time in the range of 0 to 400 days, total stay length in the range of 0 to 15 days, and a waiting period in the range of 0 to 100 days.
- For lead time less than 350 days, ADR varies greatly but it is comparatively lesser for bookings with greater lead time. Therefore, customers can get better deals for bookings with a greater lead time of >350 days.
- ADR varies greatly for shorter stays but for longer stays (> 15 days) ADR is comparatively very less. Therefore, customers can get better deals for longer stays (> 15 days).





- The ADR for 2 people is around 100 and for people >= 3, the ADR is greater than 200. Thus a booking for a single person or a couple would be a better deal for the customers.
- For bookings with >= 6 persons in a single booking, the ADR is considerably low which could be a better deal for customers.
- Customers can get a better deal if the booking had a waiting period. Apart from most of the bookings with a no waiting period the ADR falls below 200.

Conclusion:

- The majority of the bookings from the dataset are from the year 2016.
- Of the bookings and cancellations made, the majority of them were for a city hotel.
- Most of the bookings made were from the European countries, predominantly from Portugal while the higher average special requests were received from African countries.
- Bed and Breakfast was the most preferred meal type and the majority of the guests had no requirement for car parking spaces.
- For a shorter stay length (less than 5 days) city hotels are predominant, whereas for longer stay length (greater than 7 days) resort hotels are preferred.
- Guests used different channels for booking their stay and among them, the preferred way is TA/TO. Also, the majority of the cancellations were made via channel TA/TO.
- Hotels can work to increase outreach on GDS channels to get higher revenue-generating deals.
- The bookings made via TA/TO channel are booked way ahead of the actual arrival date.
- Cancellation of bookings is not affected by not getting the same room as reserved, deposit type, longer lead time, and longer waiting period. Although a slightly lesser ADR is observed for bookings with a different room allotted than the reserved one.
- July and August are the busier and most profitable months for both hotels.
- Within a month, for both hotels, arrival increases during weekends and ADR gradually increases as the month ends.
- Most of the special requests are received by the city hotel and from the channel TA/TO. In July - August most of the requests were received. More number of people (kids and adults) in a particular booking results in more number of special requests.
- For customers, generally, longer stays (more than 15 days) can result in better deals in terms of low ADR.

Thank You...