Capstone Project

Hotel Booking Analysis

Midhun R

Points for Discussion

- Business Task
- Data Summary
- Data Cleaning
- Exploratory Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Correlation Analysis
- Conclusion

Business Task

A dataset has been given with information regarding bookings of two hotels, City Hotel and Resort Hotel, from July 2015 to August 2017. The objective of this project is to explore the given dataset and discover the factors which govern the bookings.

This will be carried out by performing univariate analysis, bivariate analysis and correlation analysis of the dataset. This is undertaken as an individual project.

Data Summary

- Number of records (rows): 119390
- Number of features (columns): 32
- Out of 119390 rows, 31994 rows are duplicate rows.
- Out of 32 columns, 4 columns have missing values, 3 columns require conversion of data type, 4 extra columns are needed to be added and in 2 columns Boolean data must be converted to corresponding string data.
- These irregularities will be handled later during data cleaning step.

Data Summary (Contd.)

1. hotel: Name of the hotel (Resort Hotel or City Hotel).
2. is_canceled: If the booking was canceled (1) or not (0).
3. lead_time: Number of days before the actual arrival of the guests.
4. arrival_date_year: Year of arrival date.
5. arrival_date_month: Month of arrival date.
6. arrival_date_week_number: Week number of year for arrival date.
7. arrival_date_day_of_month: Day of month for arrival date.
8. stays_in_weekend_nights: Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests.
9. stays_in_week_nights: Number of weeknights (Monday to Friday) spent at the hotel by the guests.
10. adults: Number of adults among guests.
11. children: Number of children among guests.

Data Summary (Contd.)

12. babies: Number of babies among guests.
13. meal: Type of meal booked.
14. country: Country of guests.
15. market_segment: Designation of market segment.
16. distribution_channel: Name of booking distribution channel.
17. is_repeated_guest: If the booking was from a repeated guest (1) or not (0).
18. previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.
19. previous_bookings_not_canceled: Number of previous bookings not cancelled by the customer prior to the current booking.
20. reserved_room_type: Code of room type reserved.
21. assigned_room_type: Code of room type assigned.

Data Summary (Contd.)

22. booking_changes: Number of changes/amendments made to the booking.
23. deposit_type: Type of the deposit made by the guest
24. agent: ID of travel agent who made the booking.
25. company: ID of the company that made the booking.
26. days_in_waiting_list: Number of days the booking was in the waiting list.
27. customer_type: Type of customer
28. adr: Average Daily Rate, as defined by dividing the sum of all lodging transactions by the total number of staying nights.
29. required_car_parking_spaces: Number of car parking spaces required by the customer.
30. total_of_special_requests: Number of special requests made by the customer.
31. reservation_status: Reservation status (Canceled, Check-Out or No-Show).
32. reservation_status_date: Date at which the last reservation status was updated.

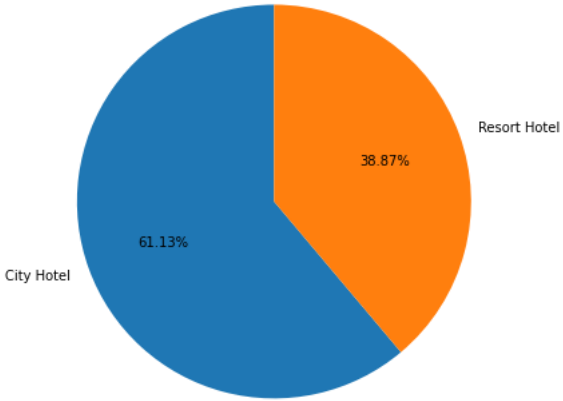
Data Cleaning

- 31994 duplicate rows were dropped.
- About 94% of data in ‘company’ is null, so the column was dropped.
- About 14% of data in ‘agent’ is null, they were replaced with mode.
- Only 0.5% of data in ‘country’ is null, they were replaced with ‘Others’.
- Only 4 rows in ‘children’ had null value, they were replaced with 0.

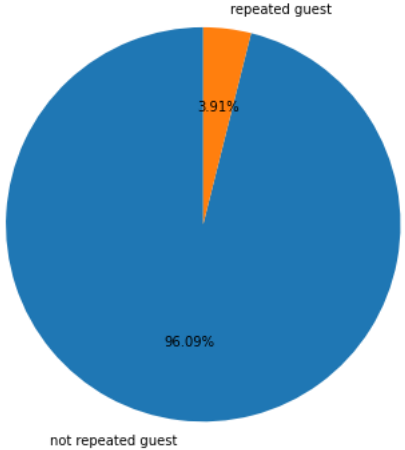
Data Cleaning (Contd.)

- ‘children’ and ‘agent’ are converted to integer
- 'reservation_status_date’ is converted to date.
- Four new columns ‘total_stays_in_nights’, ‘total_guests’, ‘revenue’ and ‘is_reserved_room_type_assigned’ are created from existing columns.
- Boolean data of ‘is_canceled’ and ‘is_repeated_guest’ is converted to string for easy representation.

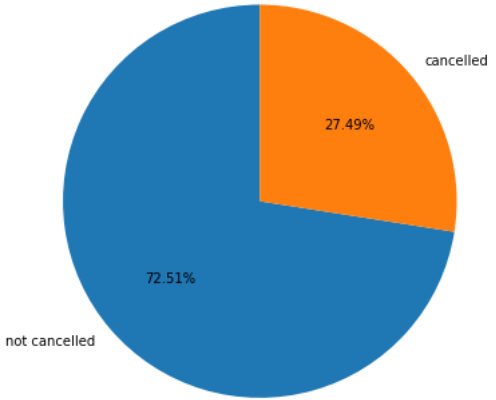
Univariate Analysis



Number of bookings for City Hotel is 1.6 times more than that of Resort Hotel.

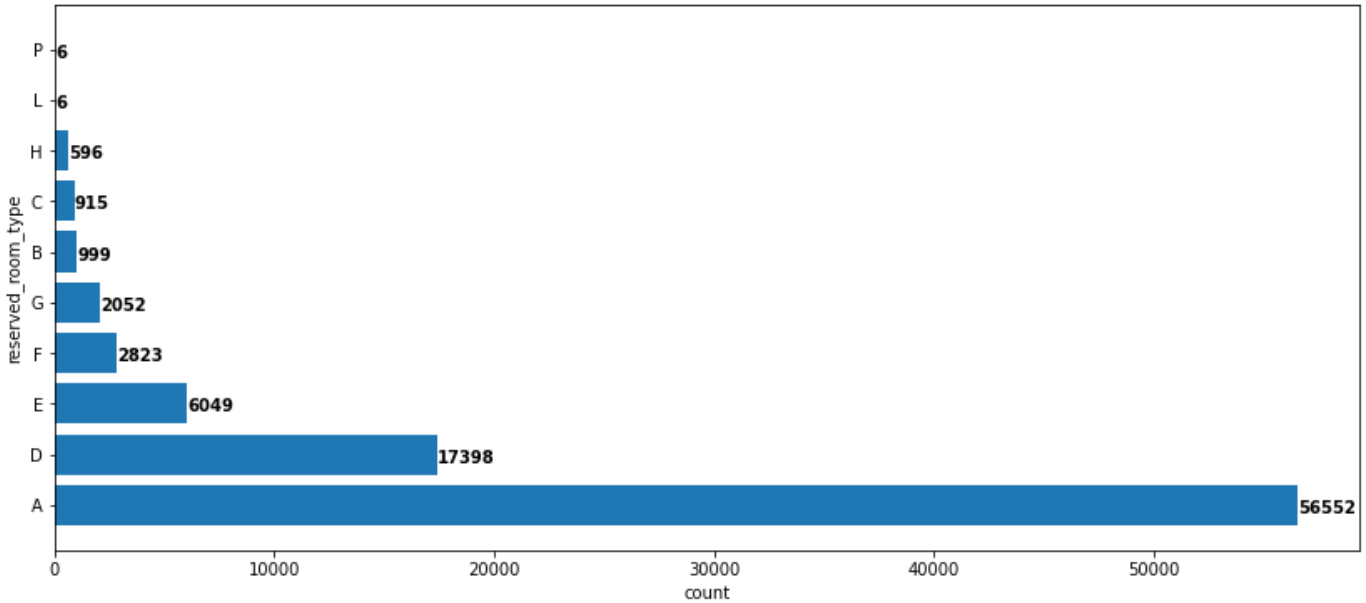


Only a very small percentage of bookings are made by repeated guests.



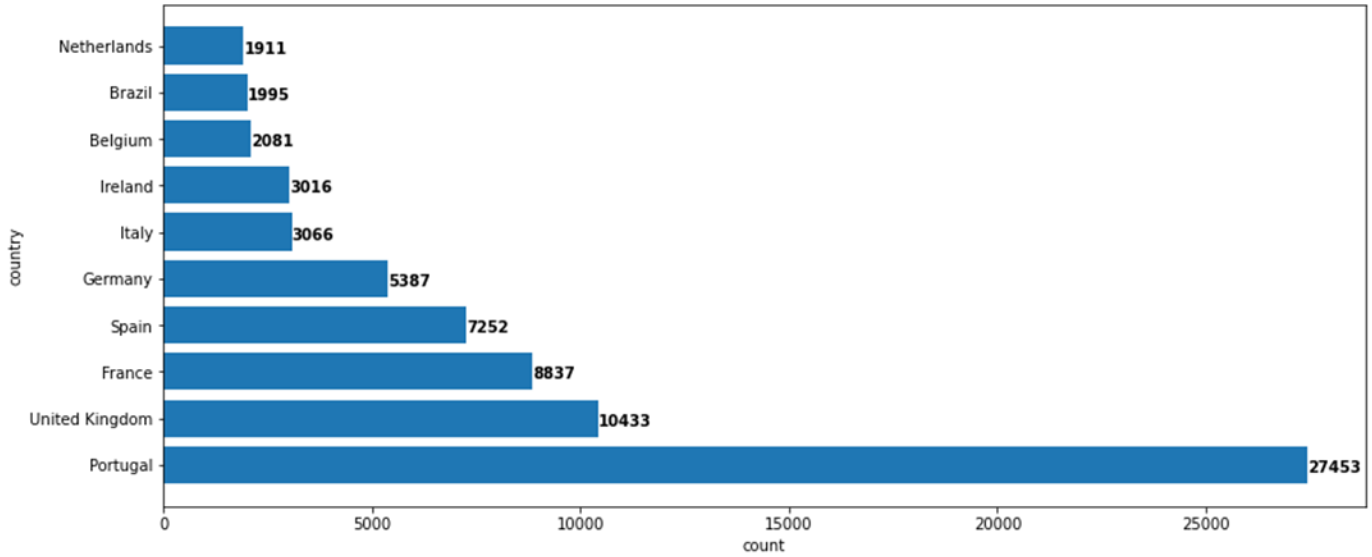
Around one-fourth of the total bookings get cancelled.

Univariate Analysis (Contd.)



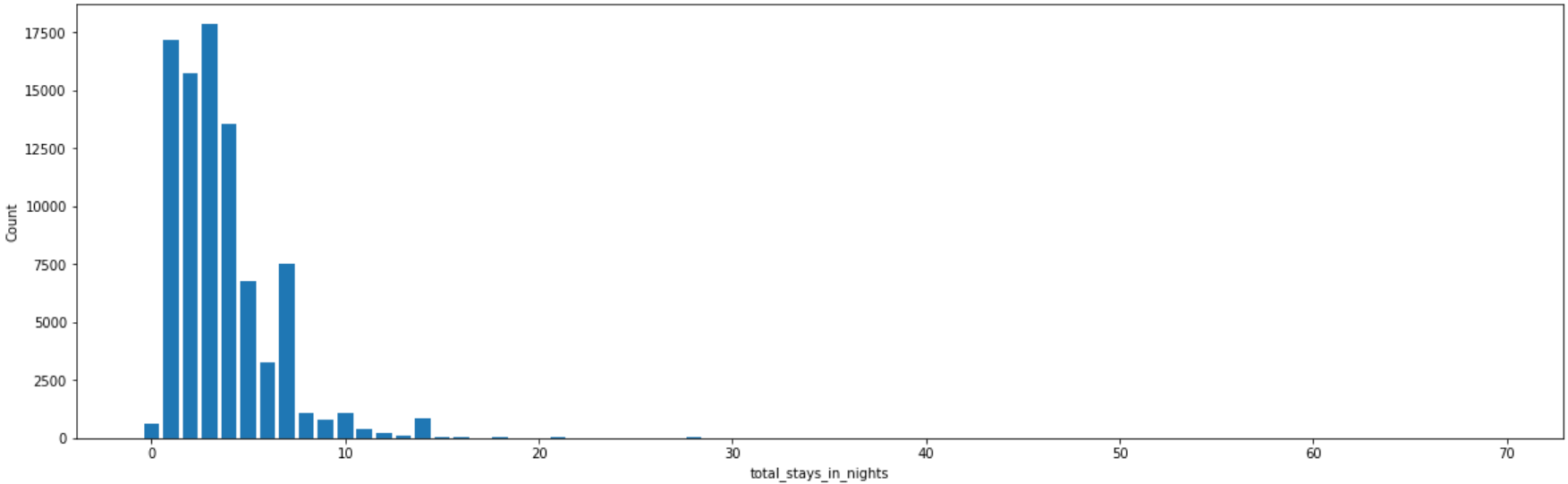
Room Type A has around 25000 more bookings than all other room types combined.

Univariate Analysis (Contd.)



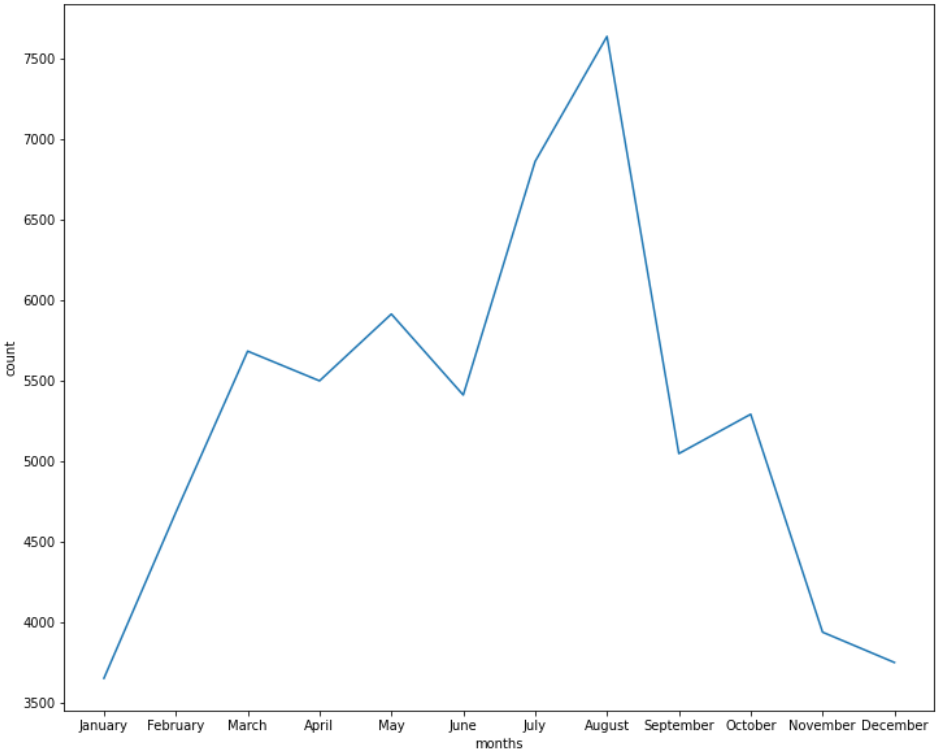
Most number of bookings was made from Portugal.

Univariate Analysis (Contd.)



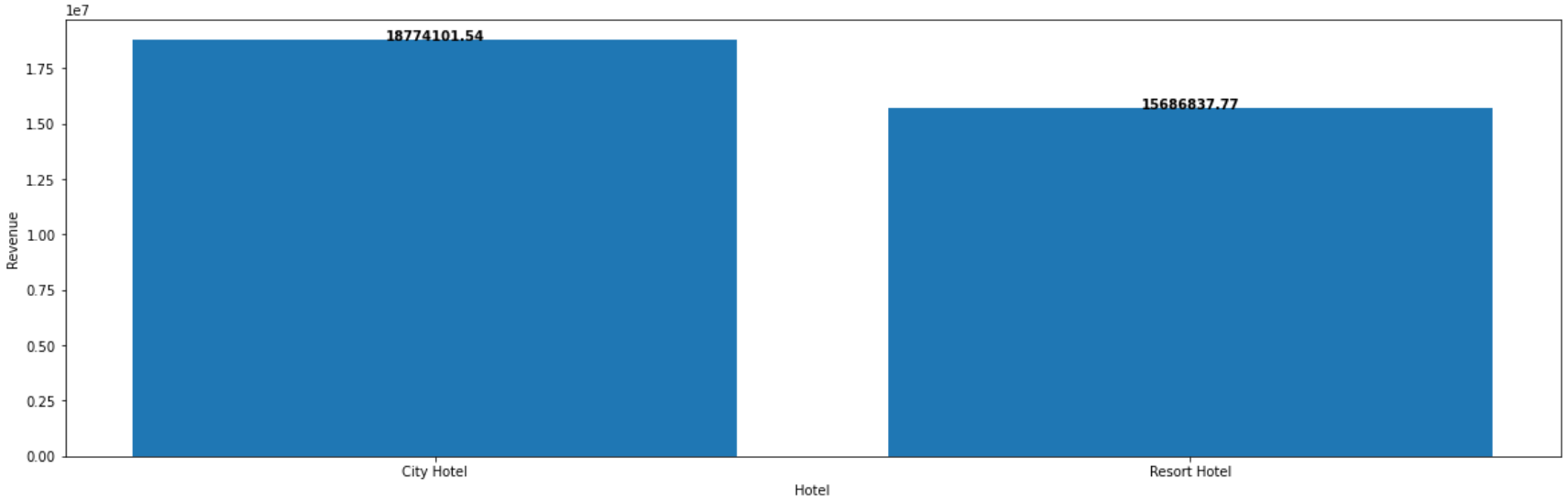
The length of stay for which the number of bookings was greater than 10,000 was for 1-4 days only.

Univariate Analysis (Contd.)



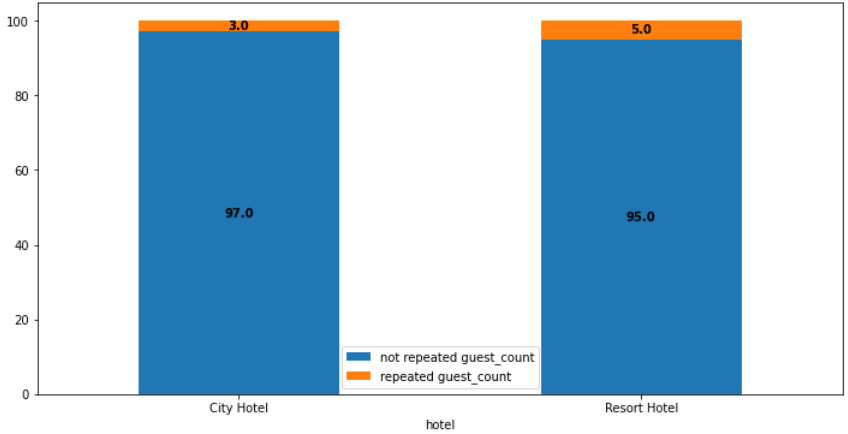
July and August has the most number of bookings.

Bivariate Analysis

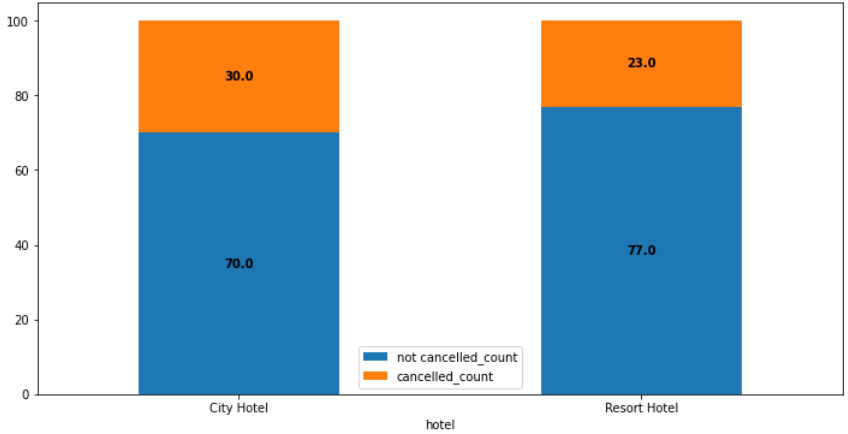


The revenue generated by the City Hotel is just 1.2 times than that of the Resort Hotel.

Bivariate Analysis (Contd.)

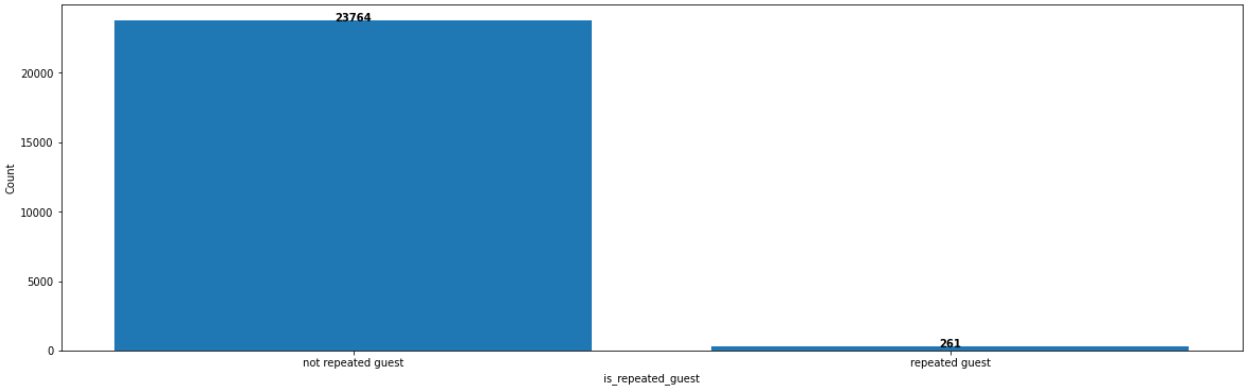


The percentage of repeated guests in Resort Hotel is greater than that of City Hotel.

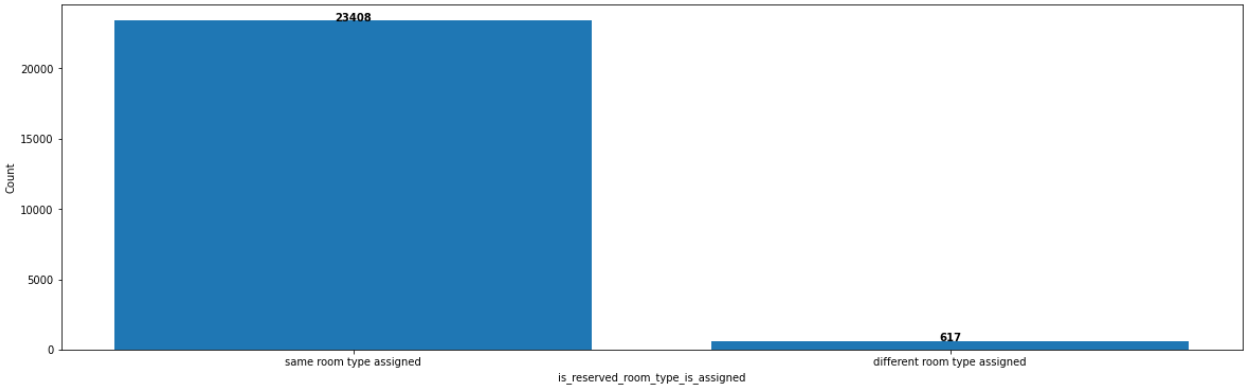


The percentage of cancelled bookings in City Hotel is greater than that of Resort Hotel.

Bivariate Analysis (Contd.)

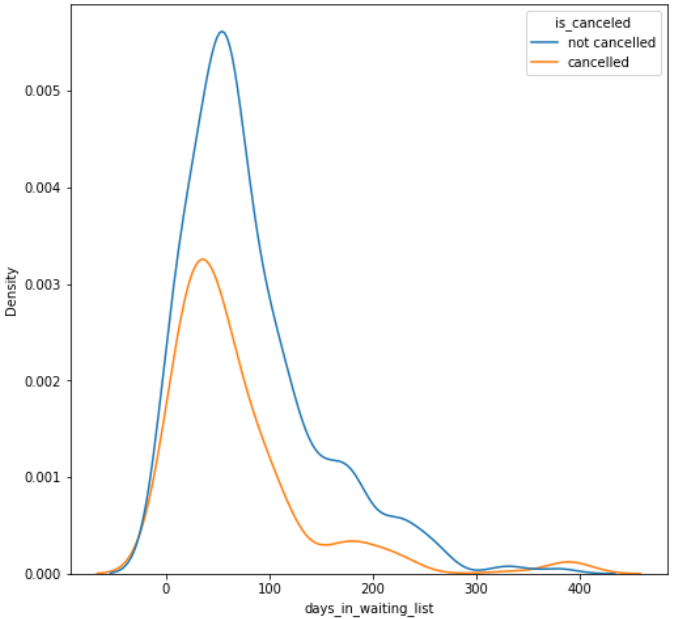


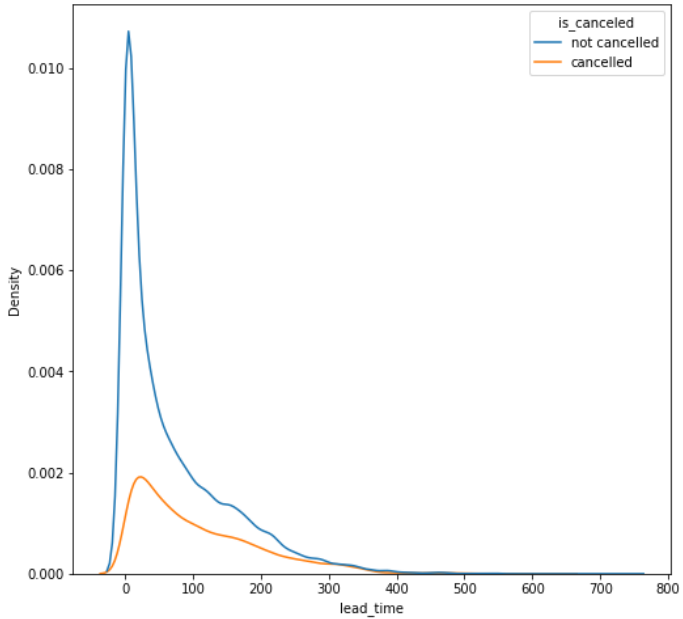
The number of bookings cancelled by new guests is 91 times more than that of repeated guests.



The number of cancelled bookings where same room type is assigned is 38 times greater than that where different room type is assigned.

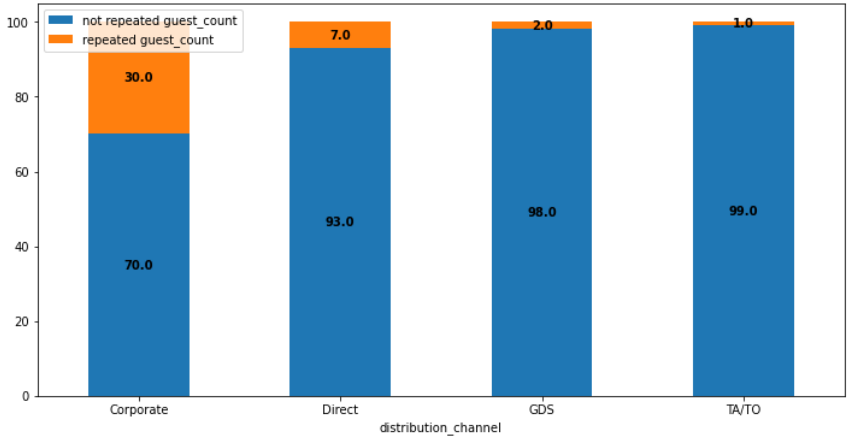
Bivariate Analysis (Contd.)



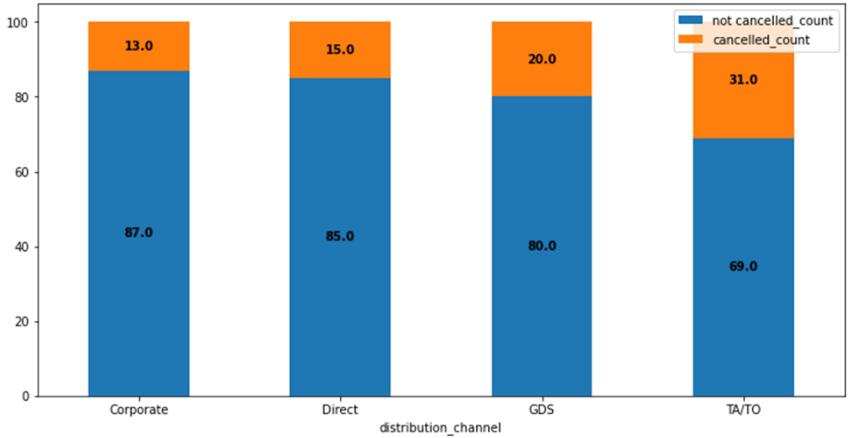


Density plot of lead time and number of days in waiting list peaks at the same point for both cancelled and non-cancelled bookings and they also attain almost identical shape.

Bivariate Analysis (Contd.)

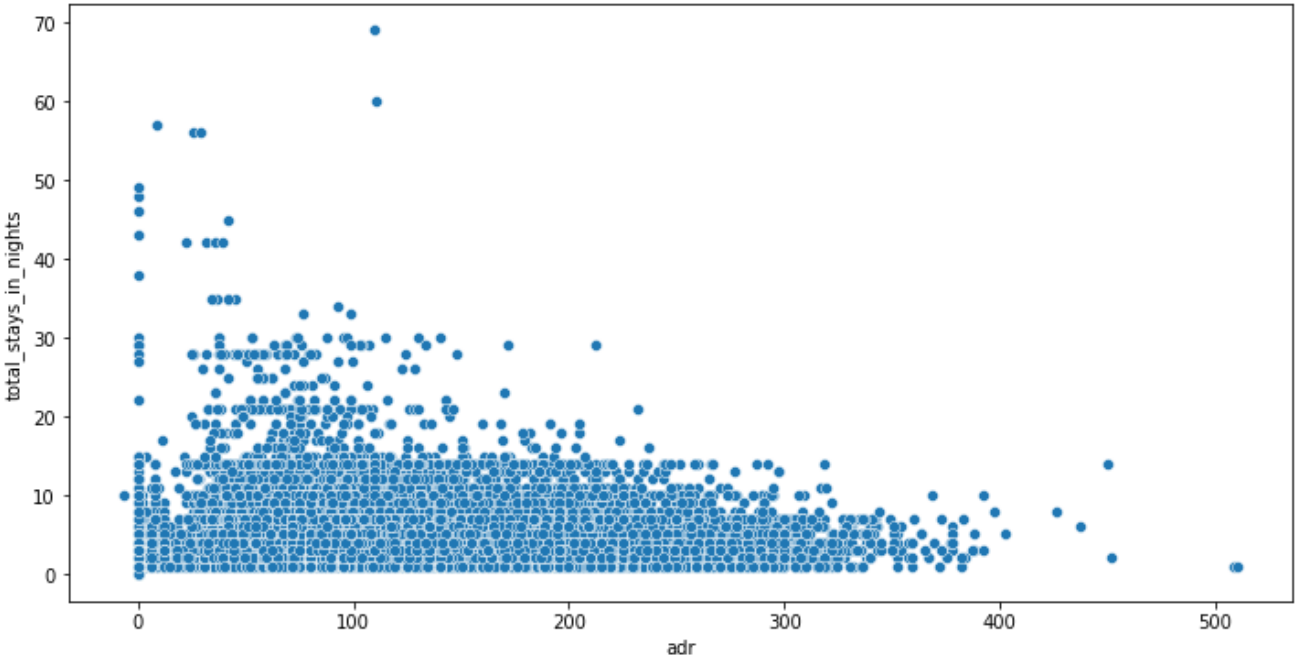


Corporate has the most percentage of repeated guests while TA/TO has the least.



TA/TO has the most percentage of cancelled bookings while Corporate has the least.

Bivariate Analysis (Contd.)



The length of stay decreases as ADR increases.

Correlation Analysis

- Lead time and length of stay have a good correlation with each other. Thus, we may conclude that generally visits planned with longer stays are booked earlier than those planned with shorter stay.
- ADR and total number of guests also have good correlation with each other. This may be because of the increase in expenditure with the increase in the count of guests.

Conclusion

- City Hotel seems to be more preferred among travelers and it also generates more revenue.
- Most number of bookings are made in July and August.
- Room Type A is the most preferred room type among travelers.
- Most number of bookings are made from Portugal.
- Most of the guest stays for 1-4 days in the hotels.
- Number of repeated guests is very low. Resort Hotel retains a greater percentage of guests.
- Around one-fourth total bookings gets cancelled. More cancellations are from City Hotel.
- New guests tend to cancel bookings more than repeated customers.
- Lead time, number of days in waiting list or assignation of reserved room to customer does not affect cancellation of bookings.
- Corporate has the most percentage of repeated guests while TA/TO has the least whereas in the case of cancelled bookings TA/TO has the most percentage while Corporate has the least.
- The length of the stay decreases as ADR increases probably to reduce the cost.
- Visits planned with longer stays are booked earlier than those planned with shorter stay.