**Exploratory Data Analysis on Hotel Booking's Data**

Amit Kumar,Anand Kumar,Birendera Ashish,
Kanishque Tyagi,Sukruth Reddy
Data science trainees,
AlmaBetter, Bangalore

## Abstract:

The project "Hotel Booking Analysis" is based on two major factors i.e. Bookings and Cancellations. The purpose of this analysis is to find out the factors that impact Booking and on how to minimize the No. of Cancellations.

In this analysis, we're provided with a dataset with some records. We did some inspection and basic data cleaning to avoid errors in the outcome. We divided the analysis into 7 segments which are : Cancellation Analysis, Monthly Trends, Duration of Stay, Geographical Analysis, ADR , Revenue and Customer Analysis. We've used different plots to visualize our analysis in the easiest way. This can be used to improve the No. of Bookings and can be used as a Customer's perspective.

Keywords: Bookings, Cancellations.

## 1.Problem Statement

Data was obtained directly from the hotels' PMS databases' servers by executing a T-SQL query on SQL Server Studio Manager, the integrated environment tool for managing Microsoft SQL databases . This query first collected the value or ID (in the case of foreign keys) of each variable in the BO(Booking Order) table.

This data set contains booking information for a City Hotel and a Resort Hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

The main object here is to do Exploratory Data Analysis (EDA) on the given data and find out various factors or parameters that affect these hotel bookings and also on how to minimize the number of cancellations

**Columns Names and Their Meanings**

- hotel: hotel type(H1 = Resort Hotel or H2 = City Hotel)
- is_canceled: Value indicating if the booking was canceled (1) or not (0)
- lead_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- arrival_date_year: Year of arrival date
- arrival_date_month: Month of arrival date
- arrival_date_week_number: Week number of year for arrival date
- arrival_date_day_of_month: Day of arrival date
- stays_in_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- stays_in_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- adults: Number of adults

- children: Number of children
- babies: Number of babies
- meal: Type of meal booked. Categories are presented in standard hospitality meal packages:
    1. Undefined/SC – no meal package
    2. BB – Bed & Breakfast
    3. HB – Half board (breakfast and one other meal – usually dinner)
    4. FB – Full board (breakfast, lunch and dinner)
- country: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- market_segment: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- distribution_channel: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- is_repeated_guest: Value indicating if the booking name was from a repeated guest (1) or not (0)
- previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
- previous_bookings_not_canceled: Number of previous bookings not cancelled by the customer prior to the current booking. reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- assigned_room_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- booking_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
- deposit_type: Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
    1. No Deposit – no deposit was made
    2. Non Refund – a deposit was made in the value of the total stay cost
    3. Refundable – a deposit was made with a value under the total cost of stay.
- agent: ID of the travel agency that made the booking
- company: ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons.
- days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type: Type of booking, assuming one of four categories:
    1. Contract - when the booking has an allotment or other type of contract associated to it
    2. Group – when the booking is associated to a group
    3. Transient – when the booking is not part of a group or contract, and is not associated to other transient booking
    4. Transient-party – when the booking is transient, but is associated to at least other transient booking.
- adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights. (measures the average rental revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using the ADR. Multiplying the ADR by the occupancy rate equals the revenue per available room.)
- 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 (e.g. twin bed or high floor)

- reservation_status: Reservation last status, assuming one of three categories:
    1. Canceled – booking was canceled by the customer
    2. Check-Out – customer has checked in but already departed
    3. No-Show – customer did not check-in and did inform the hotel of the reason why.
- reservation_status_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when the booking was canceled or when the customer checked-out of the hotel.

## 2. Introduction

- We were given a dataset of around 120k rows and 33 variables of Hotel Bookings, which contains various parameters of bookings which has been mentioned above.
- We have to analyze whole data, considering every factor which effected booking in anyway and come up with the conclusions with relevant analysis.

## 3. Exploratory Data Analysis

❖ Data Description

Fig-1 shows the Numerical Features present in the dataset. A total of 20 columns.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| is_canceled | 119390.0 | 0.370416 | 0.482918 | 0.00 | 0.00 | 0.000 | 1.0 | 1.0 |
| lead_time | 119390.0 | 104.011416 | 106.863097 | 0.00 | 18.00 | 69.000 | 160.0 | 737.0 |
| arrival_date_year | 119390.0 | 2016.156554 | 0.707476 | 2015.00 | 2016.00 | 2016.000 | 2017.0 | 2017.0 |
| arrival_date_week_number | 119390.0 | 27.165173 | 13.605138 | 1.00 | 16.00 | 28.000 | 38.0 | 53.0 |
| arrival_date_day_of_month | 119390.0 | 15.798241 | 8.780829 | 1.00 | 8.00 | 16.000 | 23.0 | 31.0 |
| stays_in_weekend_nights | 119390.0 | 0.927599 | 0.998613 | 0.00 | 0.00 | 1.000 | 2.0 | 19.0 |
| stays_in_week_nights | 119390.0 | 2.500302 | 1.908286 | 0.00 | 1.00 | 2.000 | 3.0 | 50.0 |
| adults | 119390.0 | 1.856403 | 0.579261 | 0.00 | 2.00 | 2.000 | 2.0 | 55.0 |
| children | 119386.0 | 0.103890 | 0.398561 | 0.00 | 0.00 | 0.000 | 0.0 | 10.0 |
| babies | 119390.0 | 0.007949 | 0.097436 | 0.00 | 0.00 | 0.000 | 0.0 | 10.0 |
| is_repeated_guest | 119390.0 | 0.031912 | 0.175767 | 0.00 | 0.00 | 0.000 | 0.0 | 1.0 |
| previous_cancellations | 119390.0 | 0.087118 | 0.844336 | 0.00 | 0.00 | 0.000 | 0.0 | 26.0 |
| previous_bookings_not_canceled | 119390.0 | 0.137097 | 1.497437 | 0.00 | 0.00 | 0.000 | 0.0 | 72.0 |
| booking_changes | 119390.0 | 0.221124 | 0.652306 | 0.00 | 0.00 | 0.000 | 0.0 | 21.0 |
| agent | 103050.0 | 86.693382 | 110.774548 | 1.00 | 9.00 | 14.000 | 229.0 | 535.0 |
| company | 6797.0 | 189.266735 | 131.655015 | 6.00 | 62.00 | 179.000 | 270.0 | 543.0 |
| days_in_waiting_list | 119390.0 | 2.321149 | 17.594721 | 0.00 | 0.00 | 0.000 | 0.0 | 391.0 |
| adr | 119390.0 | 101.831122 | 50.535790 | -6.38 | 69.29 | 94.575 | 126.0 | 5400.0 |
| required_car_parking_spaces | 119390.0 | 0.062518 | 0.245291 | 0.00 | 0.00 | 0.000 | 0.0 | 8.0 |
| total_of_special_requests | 119390.0 | 0.571363 | 0.792798 | 0.00 | 0.00 | 0.000 | 1.0 | 5.0 |

Fig-1

Fig-1 shows the Categorical Features present in the dataset. A total of 12 columns.
Unique refer the no. of unique values present referring the category

| | count | unique | top | freq |
|---|---|---|---|---|
| **hotel** | 119390 | 2 | City Hotel | 79330 |
| **arrival_date_month** | 119390 | 12 | August | 13877 |
| **meal** | 119390 | 5 | BB | 92310 |
| **country** | 118902 | 177 | PRT | 48590 |
| **market_segment** | 119390 | 8 | Online TA | 56477 |
| **distribution_channel** | 119390 | 5 | TA/TO | 97870 |
| **reserved_room_type** | 119390 | 10 | A | 85994 |
| **assigned_room_type** | 119390 | 12 | A | 74053 |
| **deposit_type** | 119390 | 3 | No Deposit | 104641 |
| **customer_type** | 119390 | 4 | Transient | 89613 |
| **reservation_status** | 119390 | 3 | Check-Out | 75166 |
| **reservation_status_date** | 119390 | 926 | 2015-10-21 | 1461 |

Fig-2

- Data Cleaning

  We did data cleaning on the dataset and found 4 features that consist of null values which are referred below :

  - ★ Company - 112593
  - ★ Agent - 16340
  - ★ Country - 488
  - ★ Children - 4

  As the 'company' column consists of more than 90% null values, therefore we didn't consider it in any of our findings and dropped it.

  For the 'agent' column we replaced the null values with Median values.

  For the 'country' column we replaced the 'null' values with the Moderate value i.e 'PRT' (Portugal).

  For the 'children' column we replaced the 'null' values with the Median values.

- Cancellation Analysis

  We analyzed different features with respect to cancellation analysis.

  - ➔ Hotel Type
  - ➔ Arrival Month
  - ➔ Lead time
  - ➔ Market Segment
  - ➔ Deposit Type

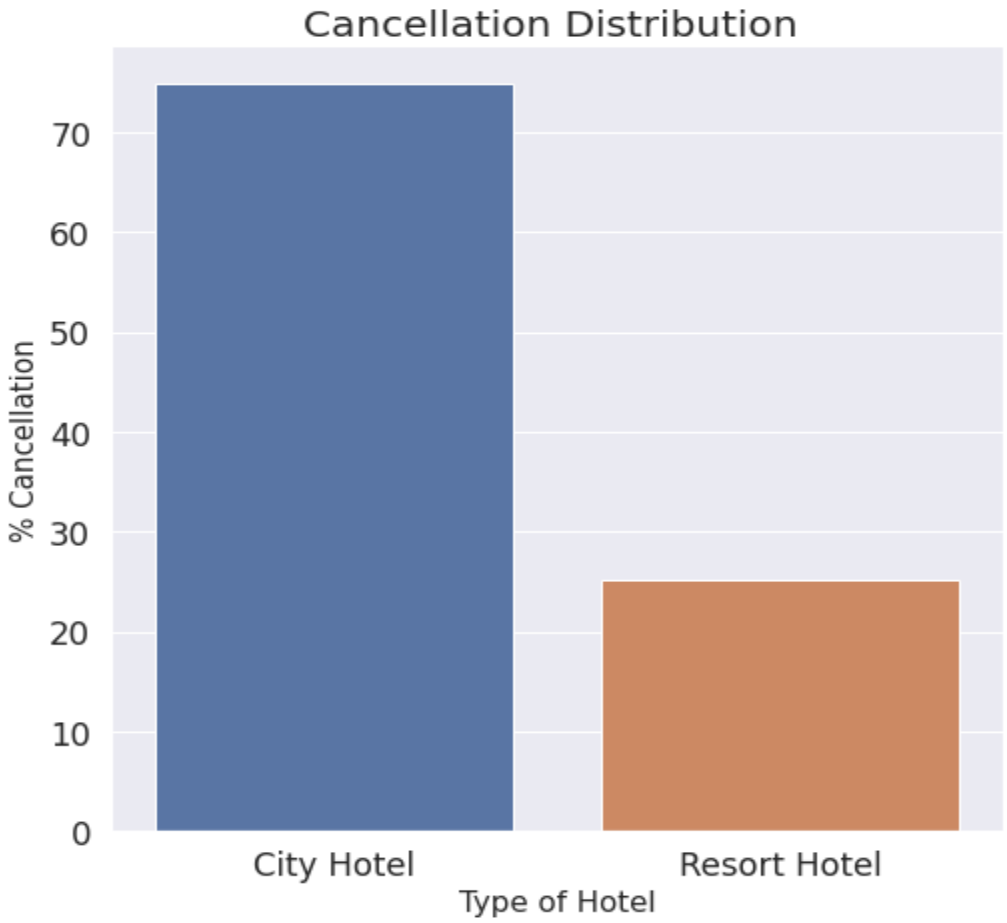



Fig -1

[ Fig - 1 ] Most bookings & cancellations are from the City Hotel followed by the Resort Hotel.

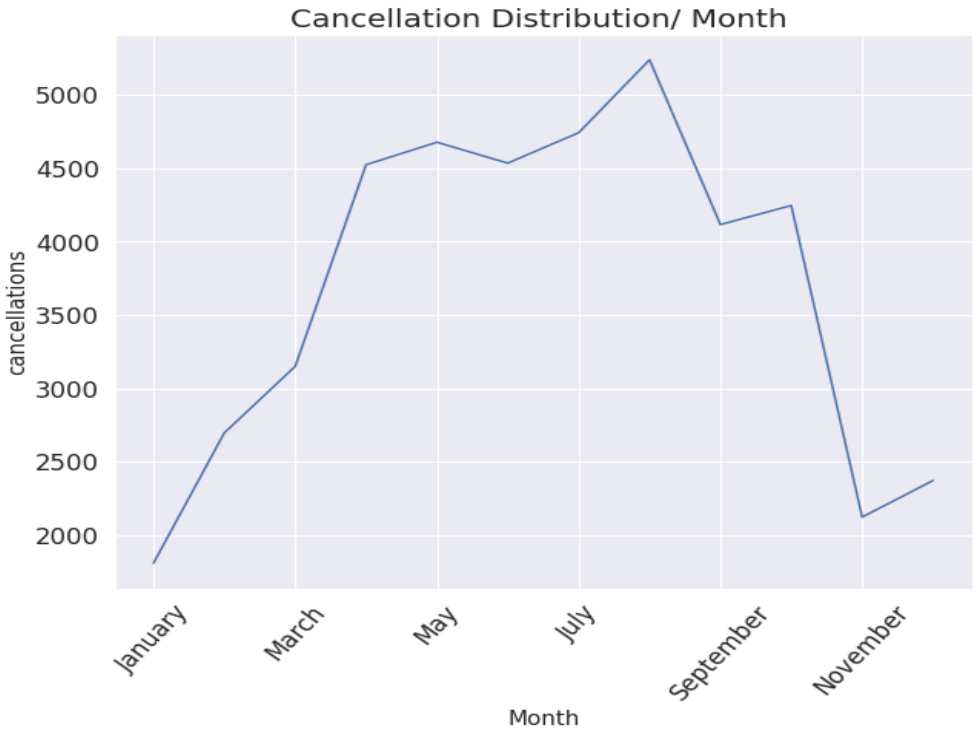



Fig-2

[ Fig - 2 ] Most cancellations were observed during the month of August and least cancellations were observed during the month of January.

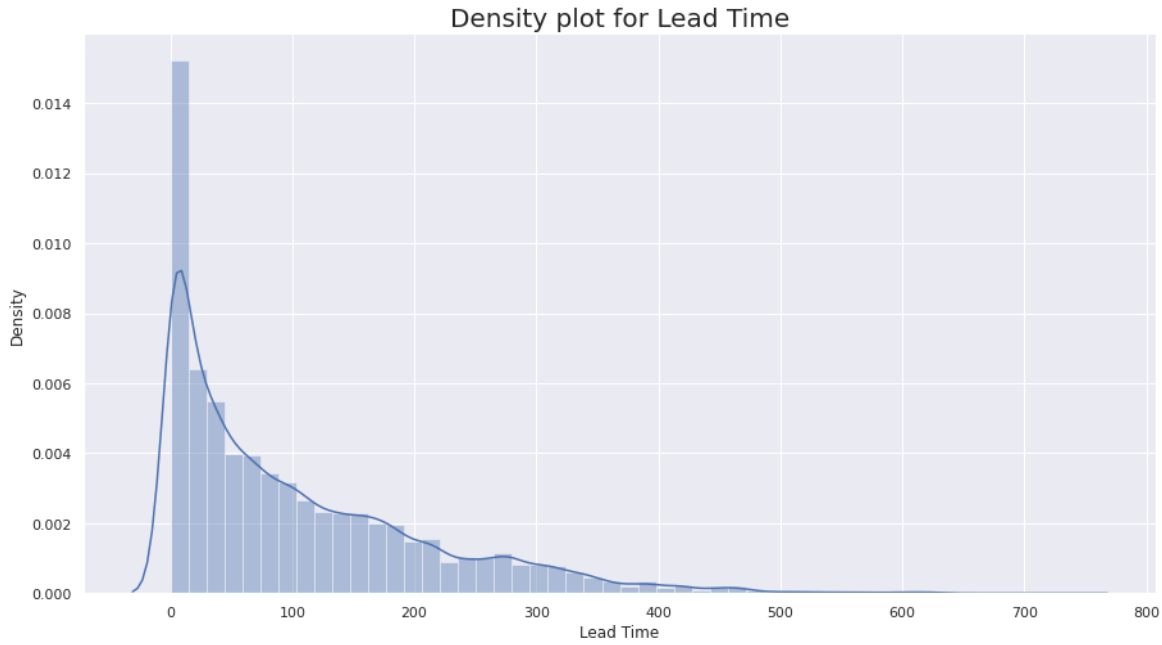



Fig - 3

[ Fig - 3 ] Shows the density plot of lead time when bookings were cancelled. Acc. to this we can see that most no. of customers checked-in the same day they booked the room.

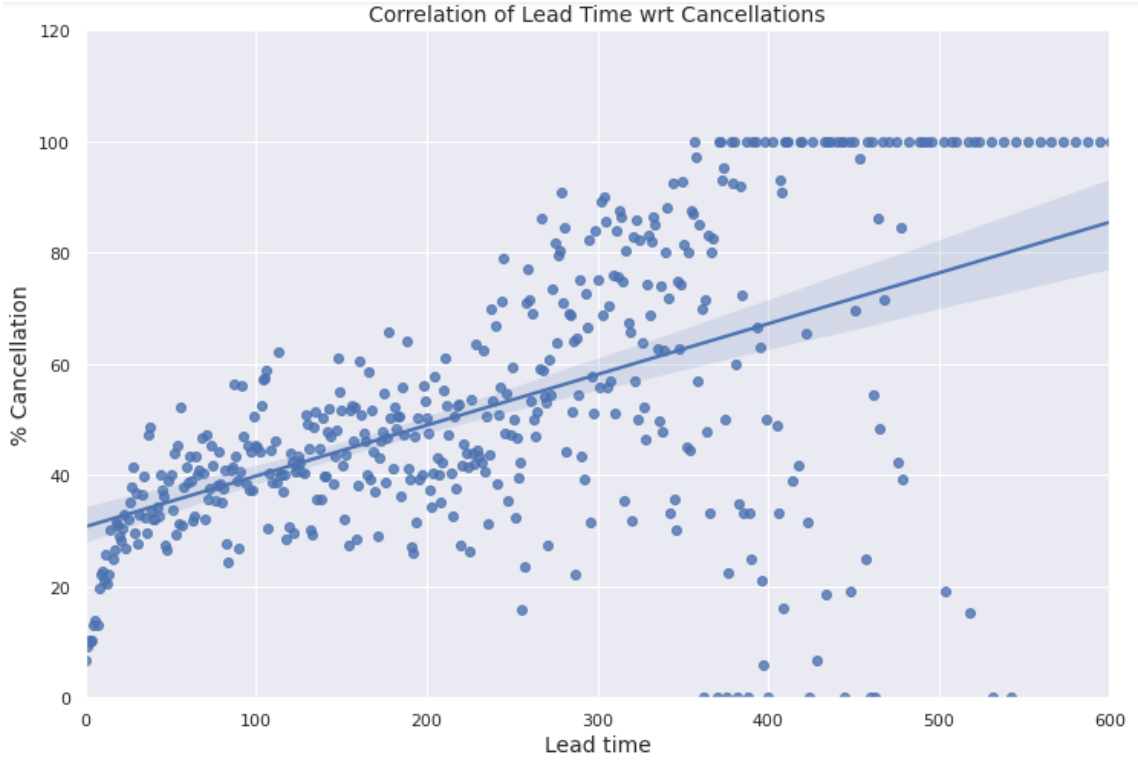



Fig - 4

[ Fig - 4 ] Correlation of lead time with respect to Cancellations - shows that with increase in lead time the no of cancellations also increase .

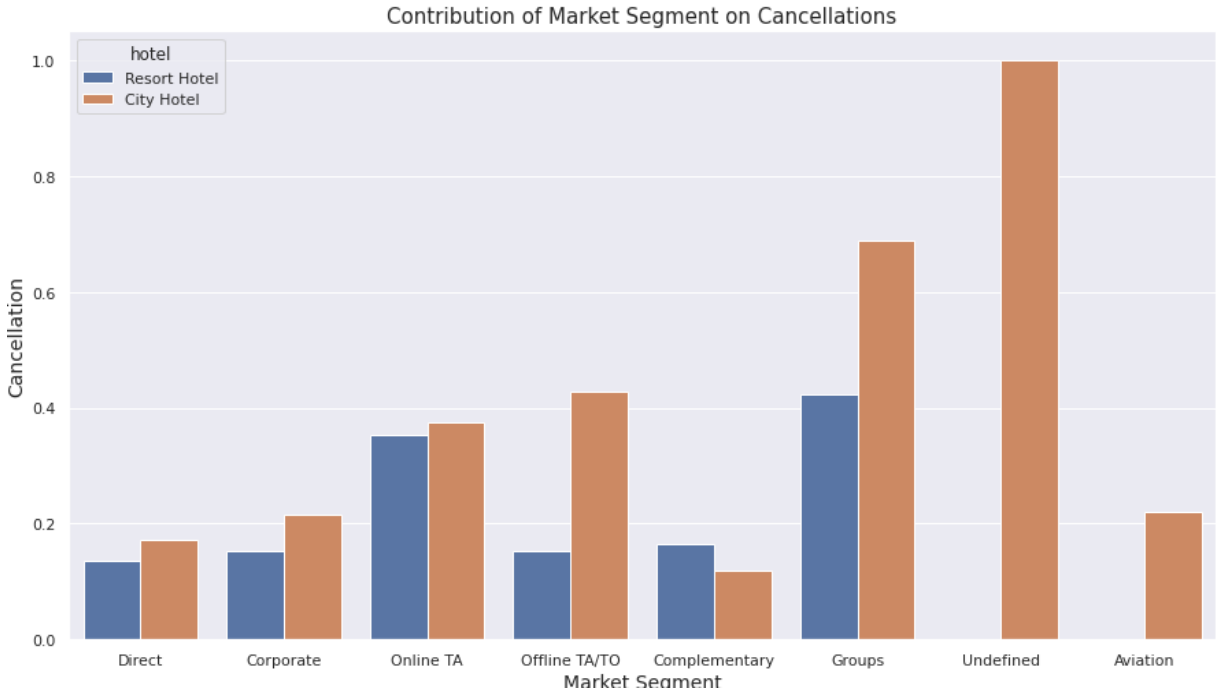



Fig - 5

[ Fig - 5 ] This shows the no. of cancellations with respect to the market segment. The ‘undefined’ market segment favors the most no of cancellations for City Hotel while ‘Groups’ market segment favors the most no. of cancellations for Resort Hotel.

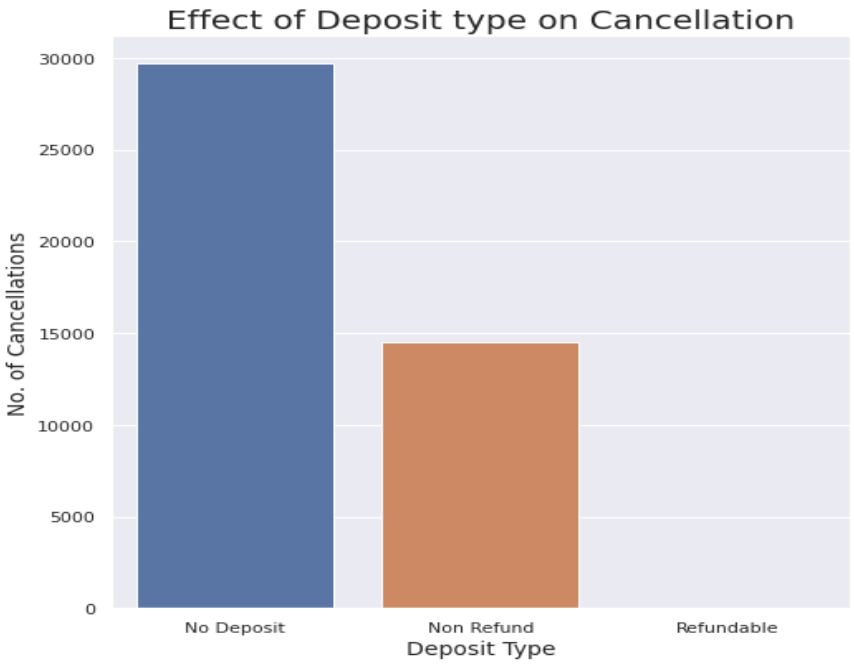



Fig - 6

[ Fig - 6 ] From the above bar graph we can see that the bookings of Customers with **'No Deposit'** type got cancelled. Maximum no of bookings that got cancelled are of **'No deposit'** type followed by the **‘Non-Refund’** type.

❖ Monthly Trends

Here, we analyzed the no. of bookings with respect to months.

1. In Fig-1, we can see that most bookings are made in the month of August for both types of hotels.
2. Fig-2 tells us more about the arrivals of customers per month for different types of hotel. The pattern for arrivals is almost the same for each month. Arrivals are more in City Hotel as compared to Resort Hotel.
   The possible reason can be due to the price range of Resort Hotel might be higher than City Hotel.

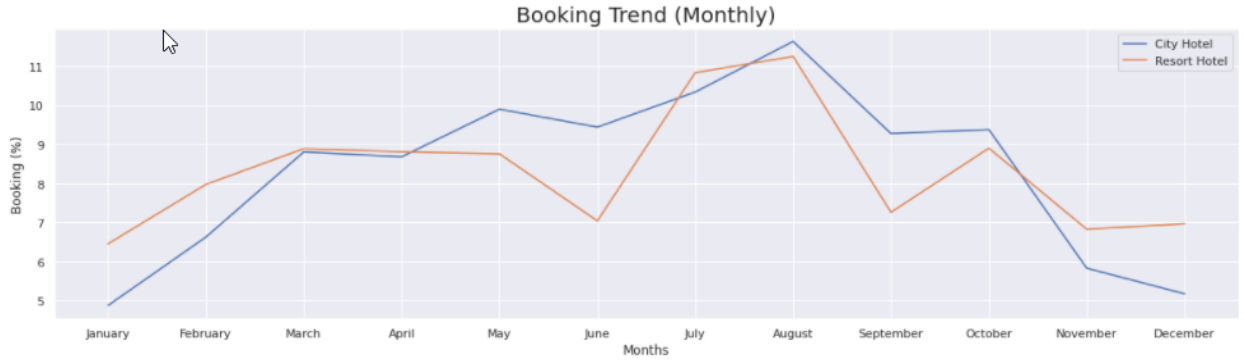



Fig-1

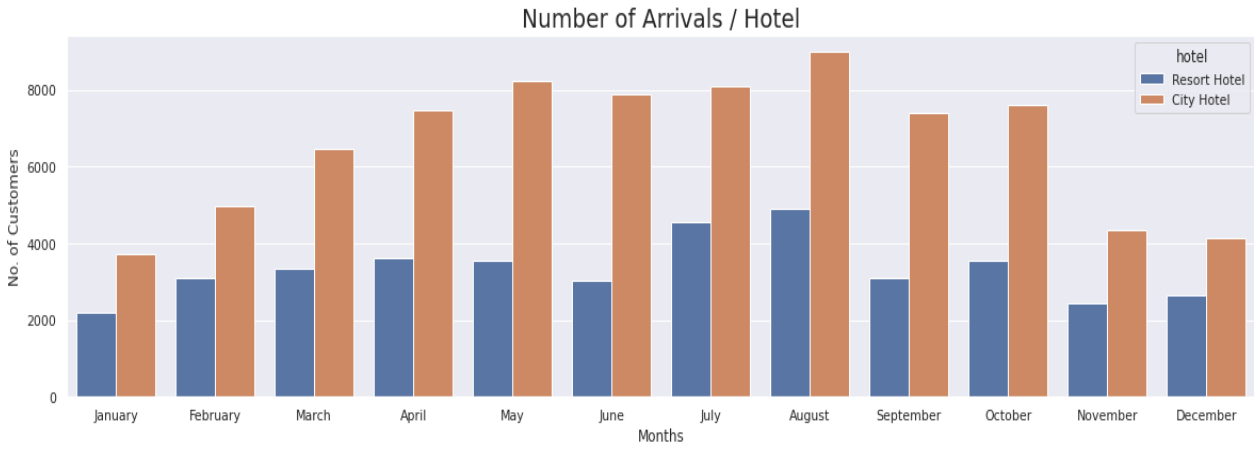



Fig-2

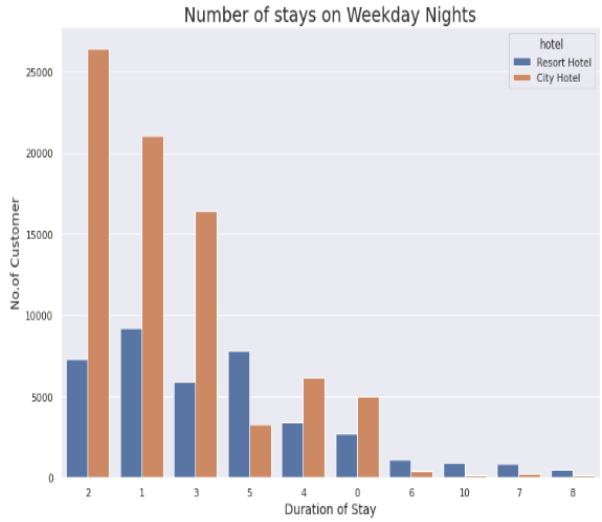
❖ Duration of Stay

In this particular segment, we analyzed the duration of stay with reference to various factors such as:

1. Hotel wise stay
2. Weekdays and Weekend nights stays
3. No of Adults staying in the hotel.


Fig-1

Fig-1 depicts the number of nights the customer stayed in both types of hotel. The maximum no. of stays is 1 for Resort Type Hotel and 3 for City Type hotel.



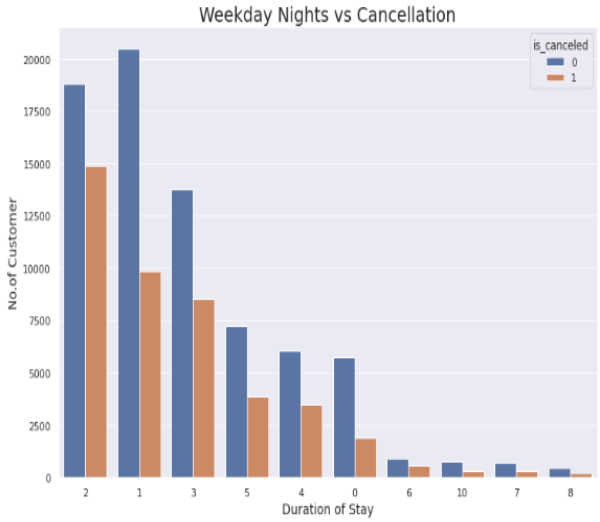
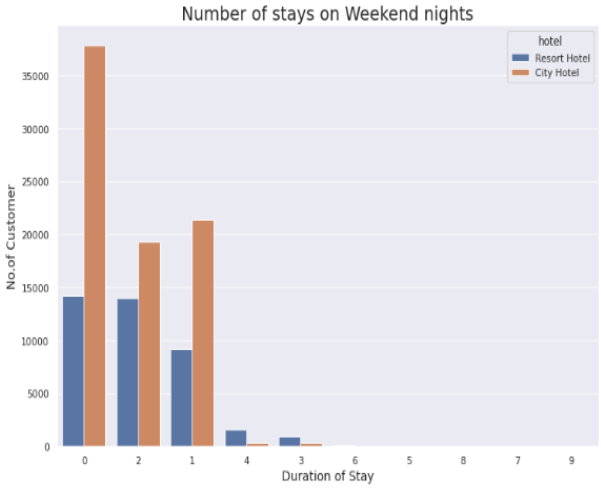




Fig-2



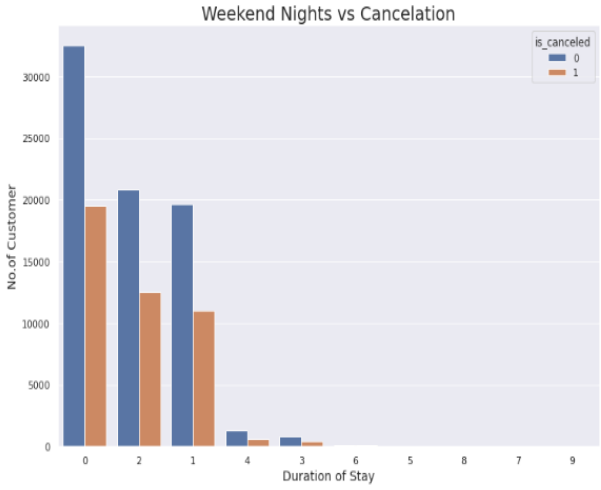




Fig-3

Fig-2 and Fig-3 depicts the no of weekday and weekend nights. It's very clear that weekend nights are more preferred by the customers as they are more free on weekends.

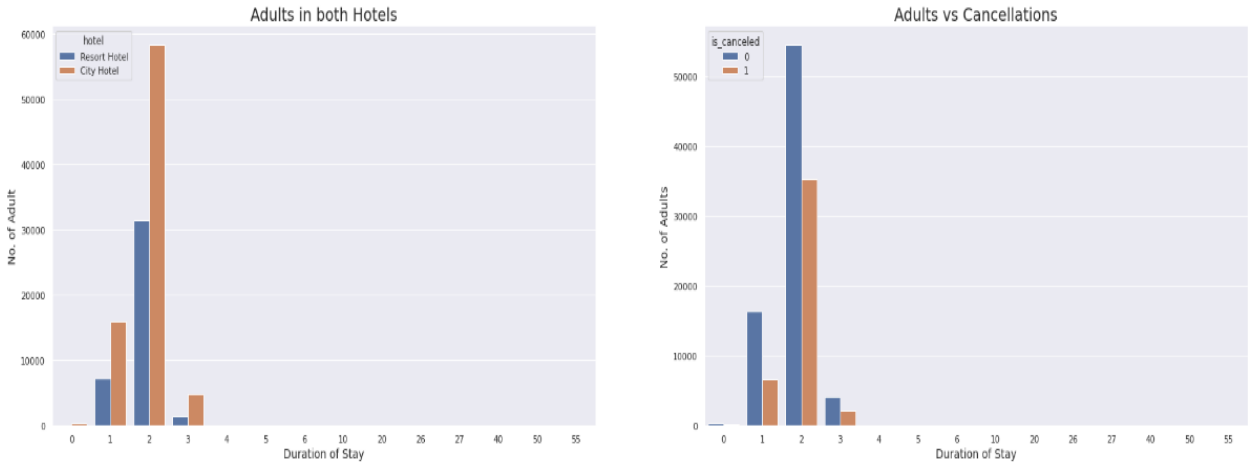



Fig-4

Fig-4 shows the no. of Adults staying in each type of hotel. The maximum no. of adults is 2 for both hotel types. With this, we can assume that the hotel is more preferred by the couples.

- Geographical Analysis

Geographical Analysis shows which is the busiest country from which shows the maximum no of customers are booking the hotels.

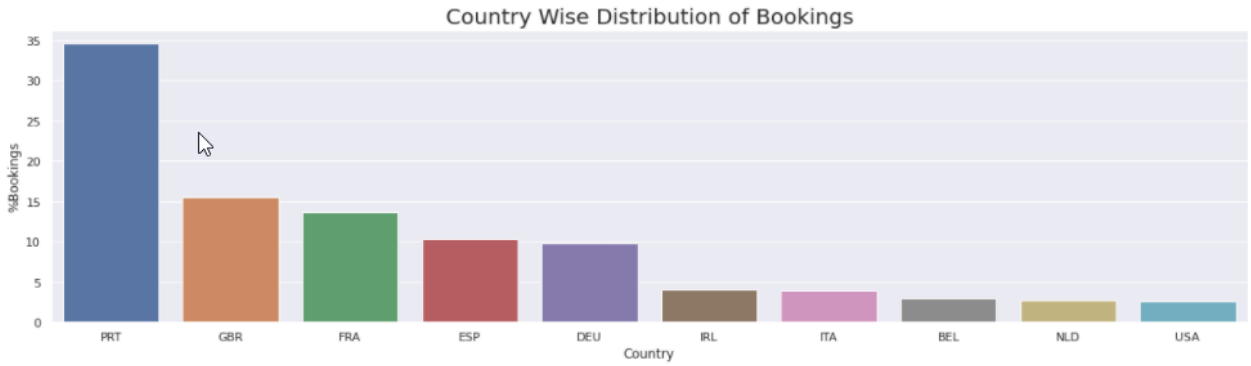



Fig-1

Fig-1 shows the TOP 10 countries from where the customers booked the Hotels. ‘PRT’ (Portugal) is the busiest country from all the countries which is responsible for most no. of bookings.

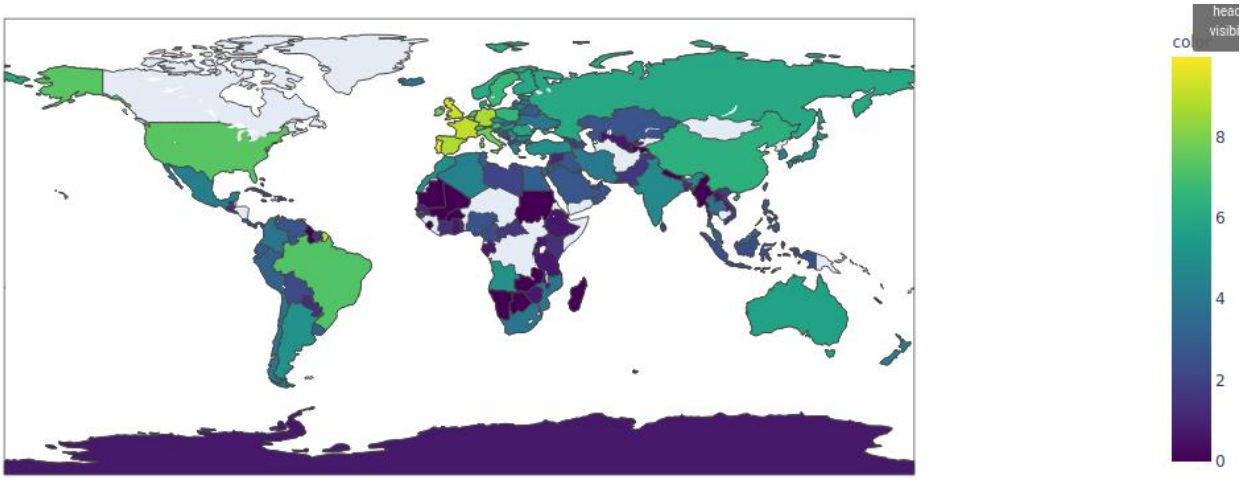


Fig - 2

[ Fig - 2 ]The above figure shows the various countries the travellers have come from to the hotels where purple signifies least/no travellers and yellow signifies the most travellers. The lighter green is the part of the world with most no of Travellers.

❖ ADR Analysis

ADR is Average daily Rate measures the average rental revenue earned for an occupied room per day.

- ★ Fig-1 shows the density plot of ADR for City Hotels. The maximum ADR for a City Hotel is around 100 Euros.
- ★ Fig-2 shows the ADR of City Hotel with respect to Months. The Maximum ADR is generated in May Month.

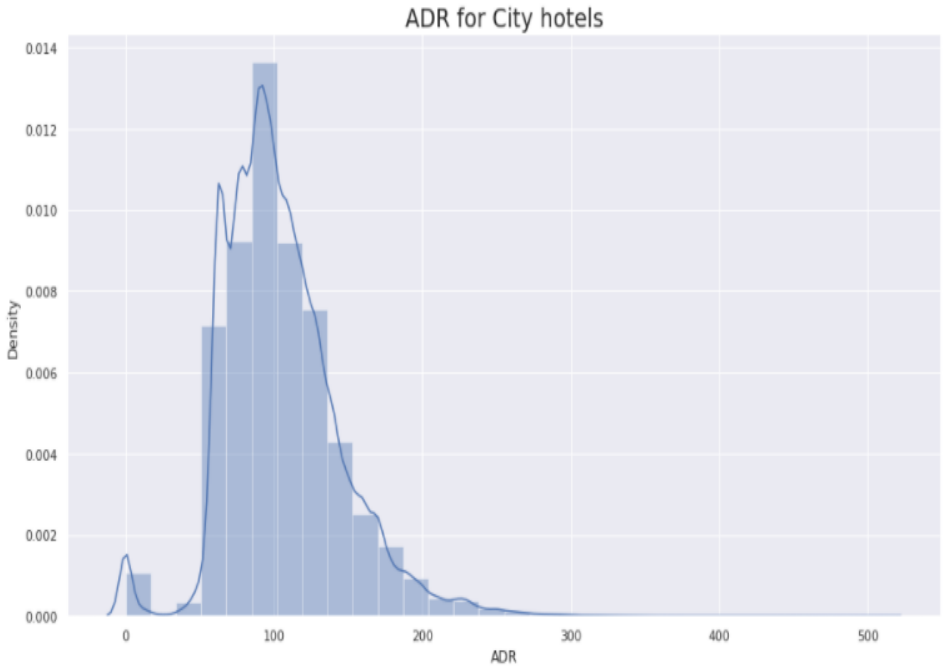



Fig-1

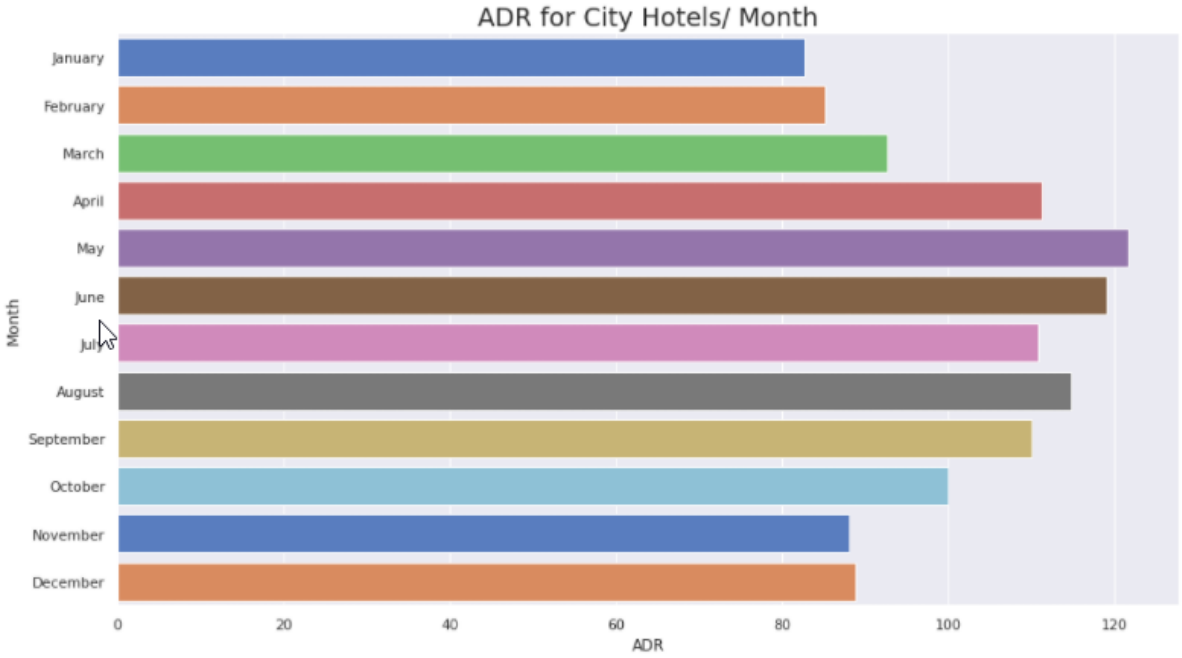



Fig-2

- ★ Fig-3 shows the density plot of ADR for Resort Hotels. The maximum ADR for a Resort Hotel is around 75 Euros.
- ★ Fig-4 shows the ADR of Resort Hotel with respect to Months. The Maximum ADR is generated in August Month.

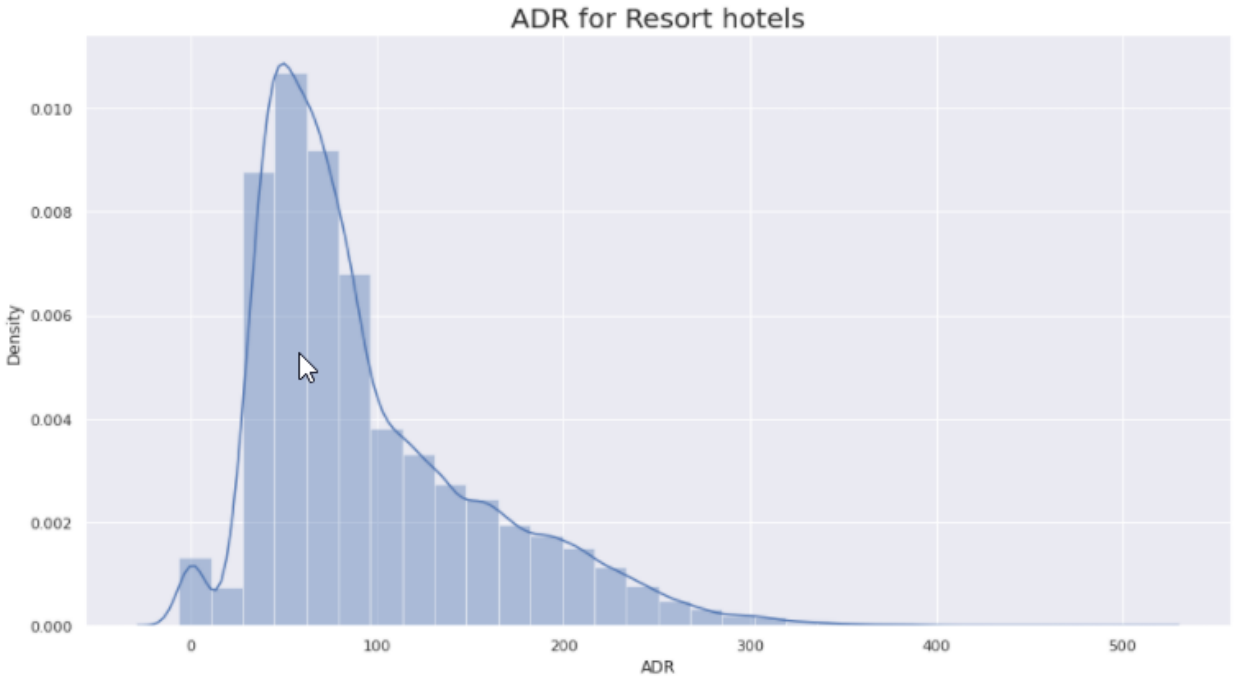



Fig-3

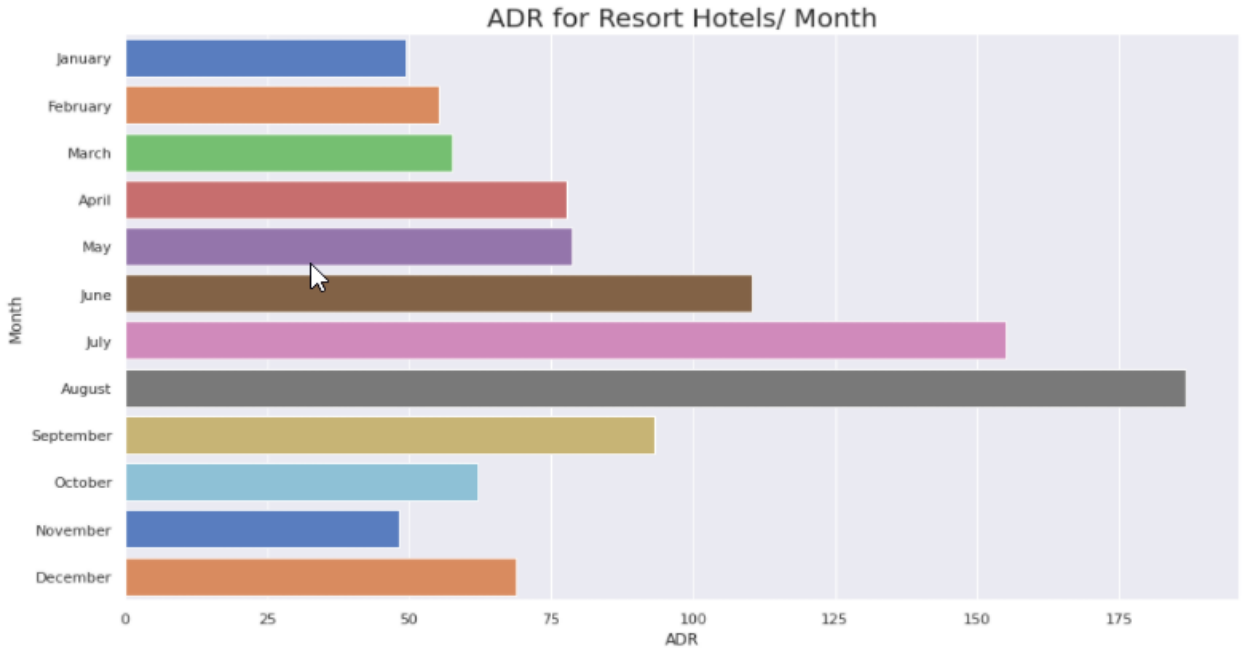



Fig-4

★ Fig-5 depicts the ADR on the basis of Room types. We have 9 types of rooms in both hotel type.
The maximum ADR for Resort Type hotel comes from ‘H’ type of room,
While that for City Hotel comes from ‘G’ room type.

This shows that most no. of customers prefer ‘G’ room type and that could be the factor where no. of cancellation occurs due to ‘G’ room type.

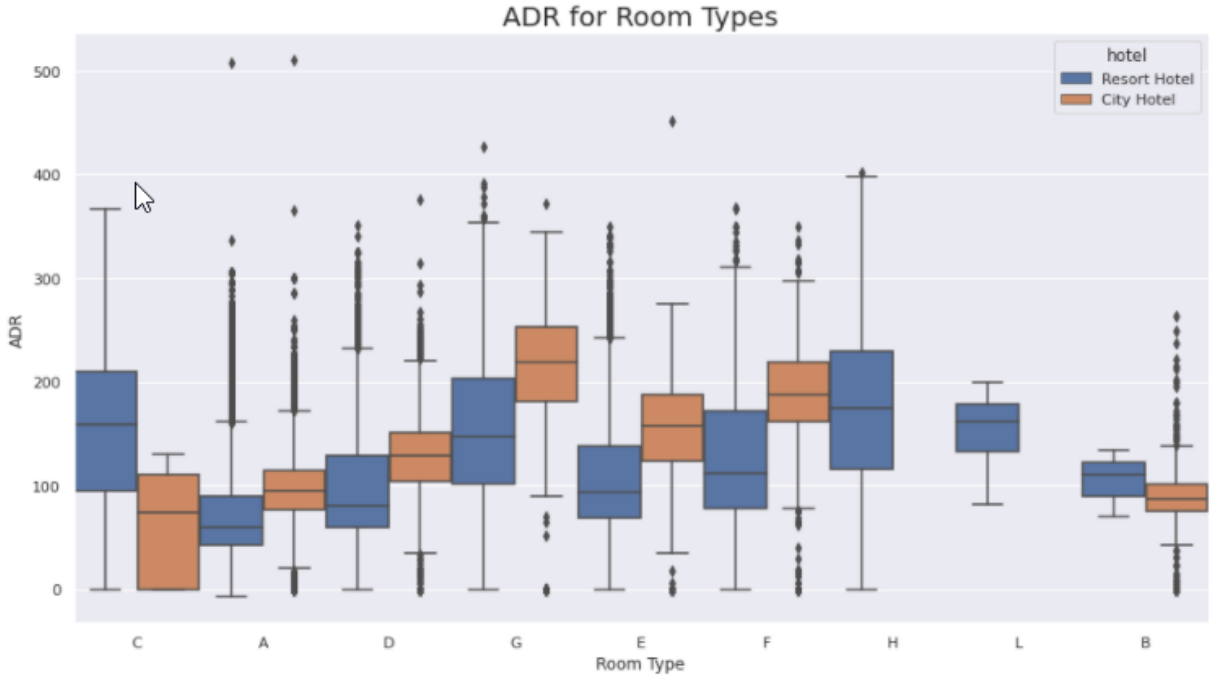



Fig-5

❖ Revenue Analysis

This particular segment tells us about the Revenue system in each month for both types of Hotel.

From Fig-1 and Fig-2, we can see that the maximum amount of Revenue comes in the most of August. This is due to the fact that there are most no. of bookings in August month.

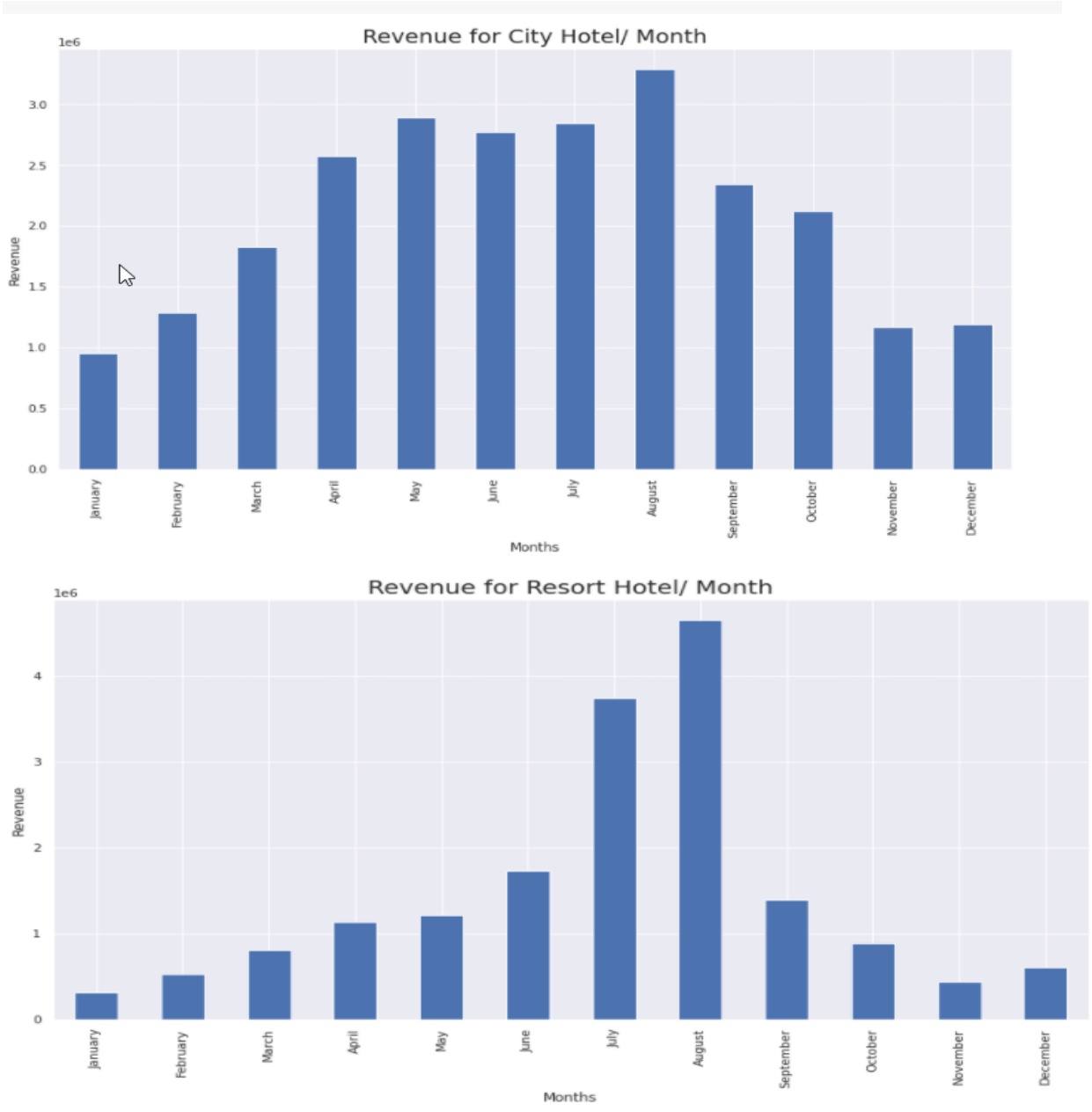
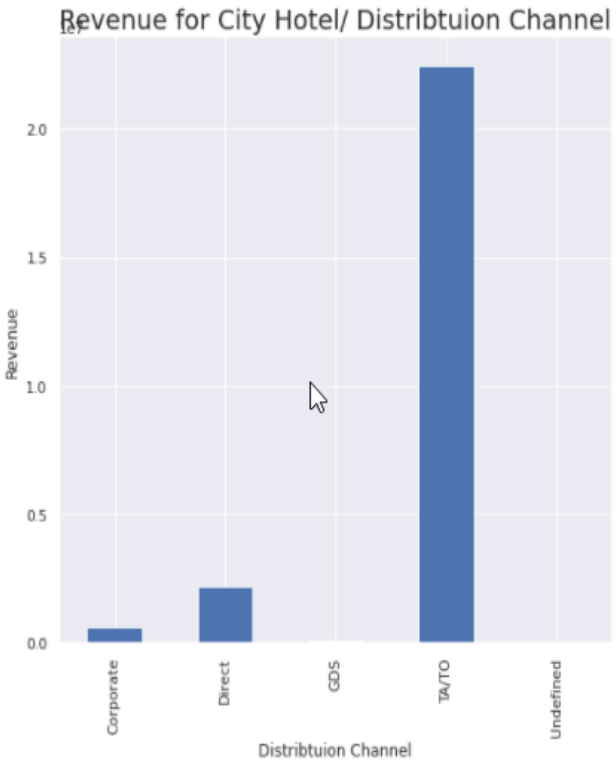



Fig-1 & Fig-2



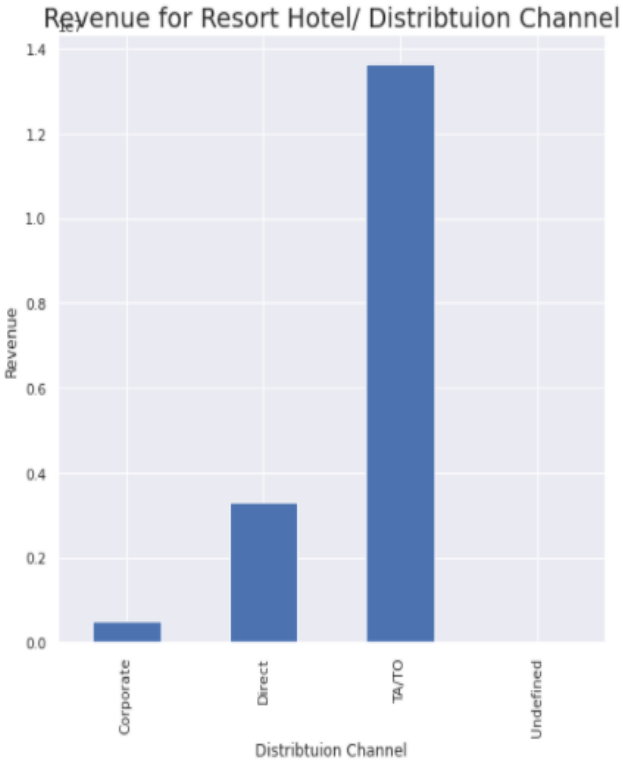



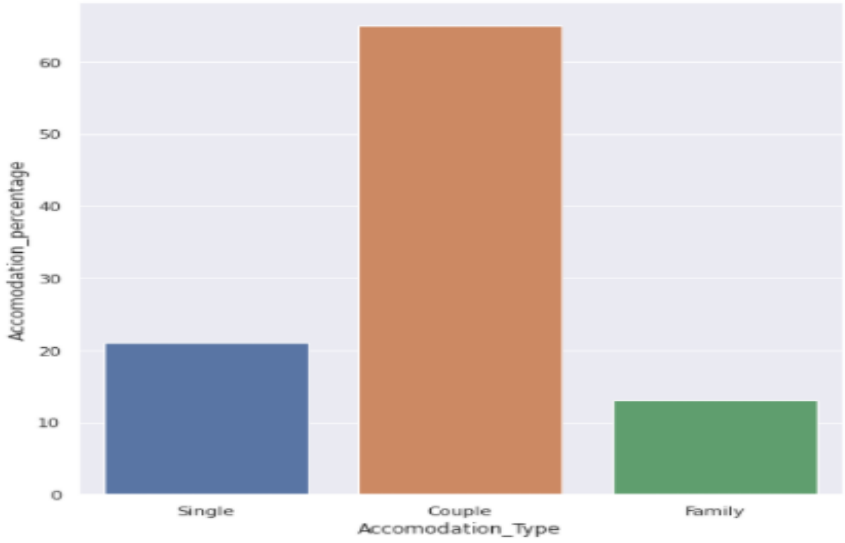

Fig - 3 [ Fig - 4 ]

[ Fig - 3 ] and [ Fig - 4 ] The graph shows the revenue generated for City Hotel by each individual distribution channel, we can see that TA(Travel Agents) or TO(Travel Operator) is generating the highest revenue.

❖ Customer Analysis

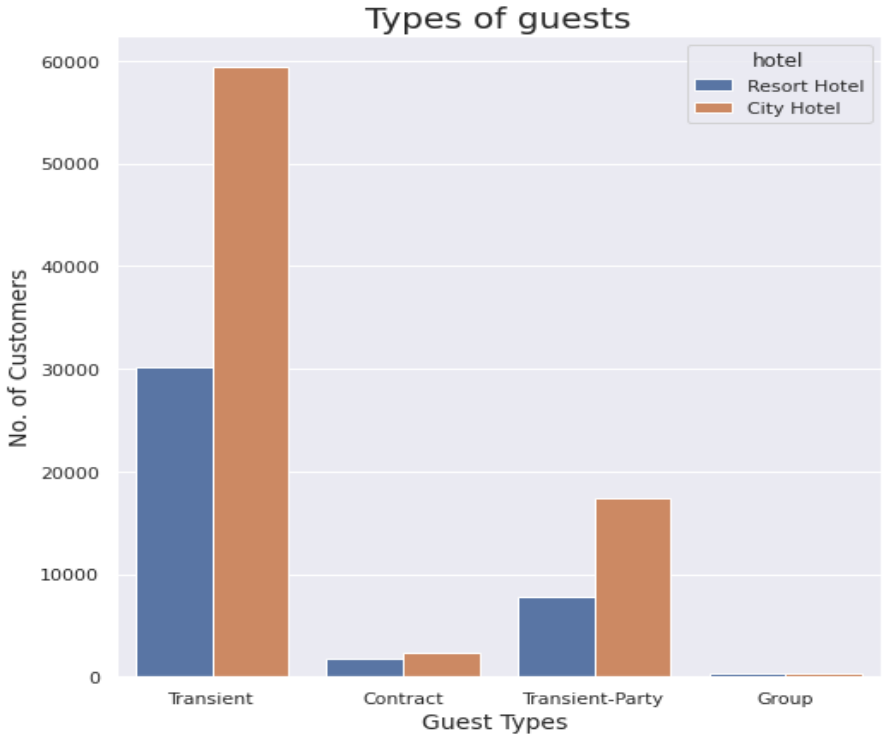
★ Accommodation Type

For the part of Customer Analysis, we've made some assumptions for accomodation type. We considered 3 types of accommodation mentioned below:
Single - 1 Adult, 0 Children , 0 Babies
Couple - 2 Adult, 0 Children , 0 Babies
Family - Adult + Children + Babies > 2



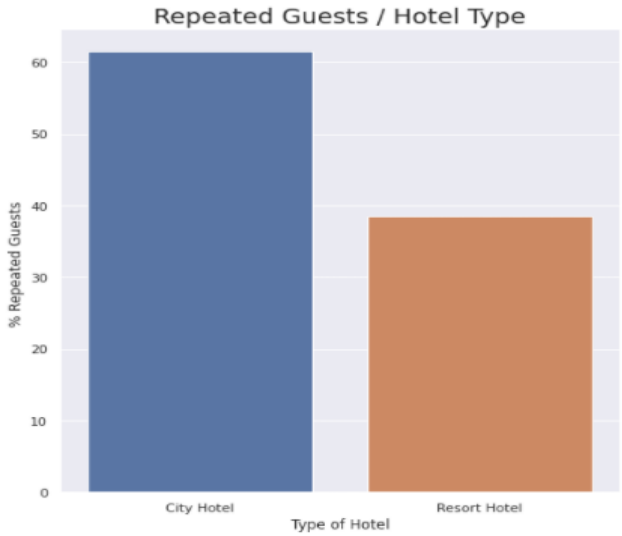
★ Guest Type

There are 4 types of guest and from the bar plot we can see that the maximum number of bookings are made by the Transient Guest type for both types of hotels.



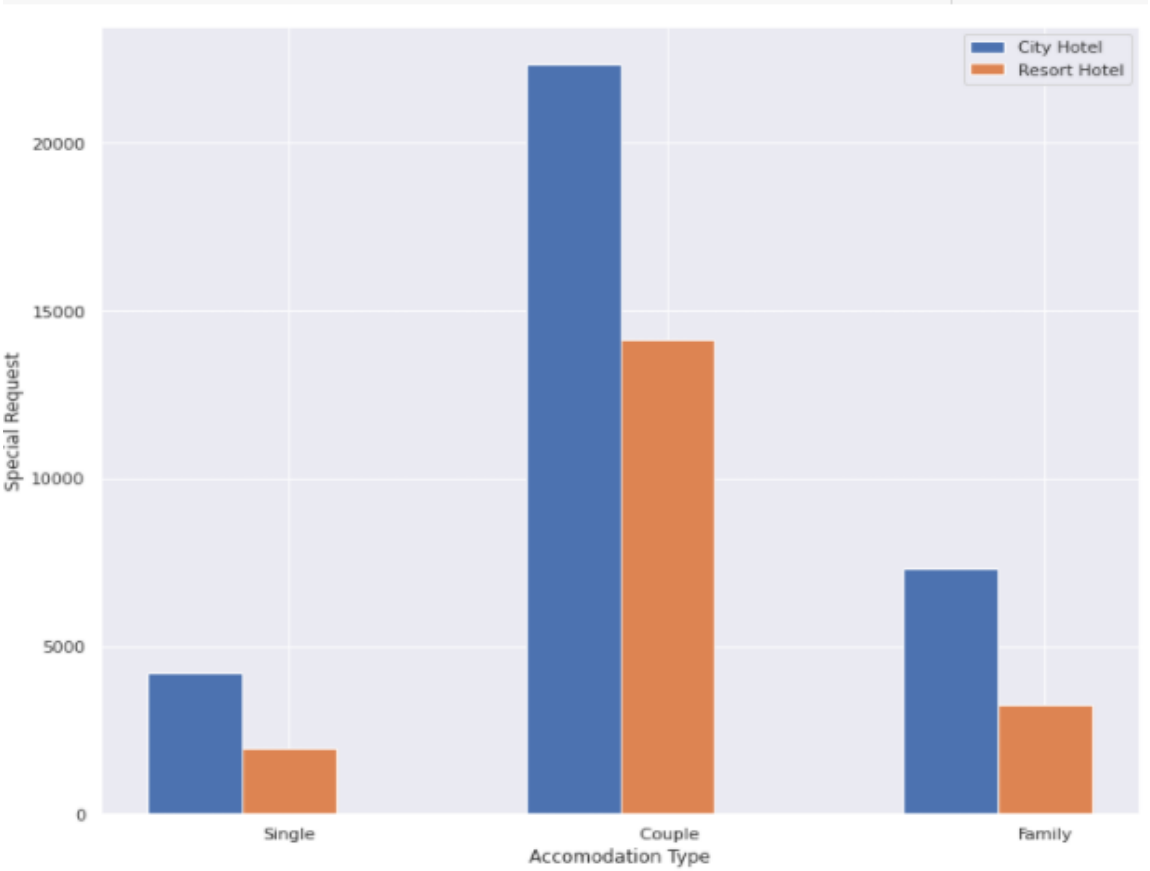

★ No. of Repeated Guests

From the graph below, we can say that there are more no of repeated guests for City Hotel than Resort Hotel.
There are more customers that prefer City Hotel on top of Resort Hotel.



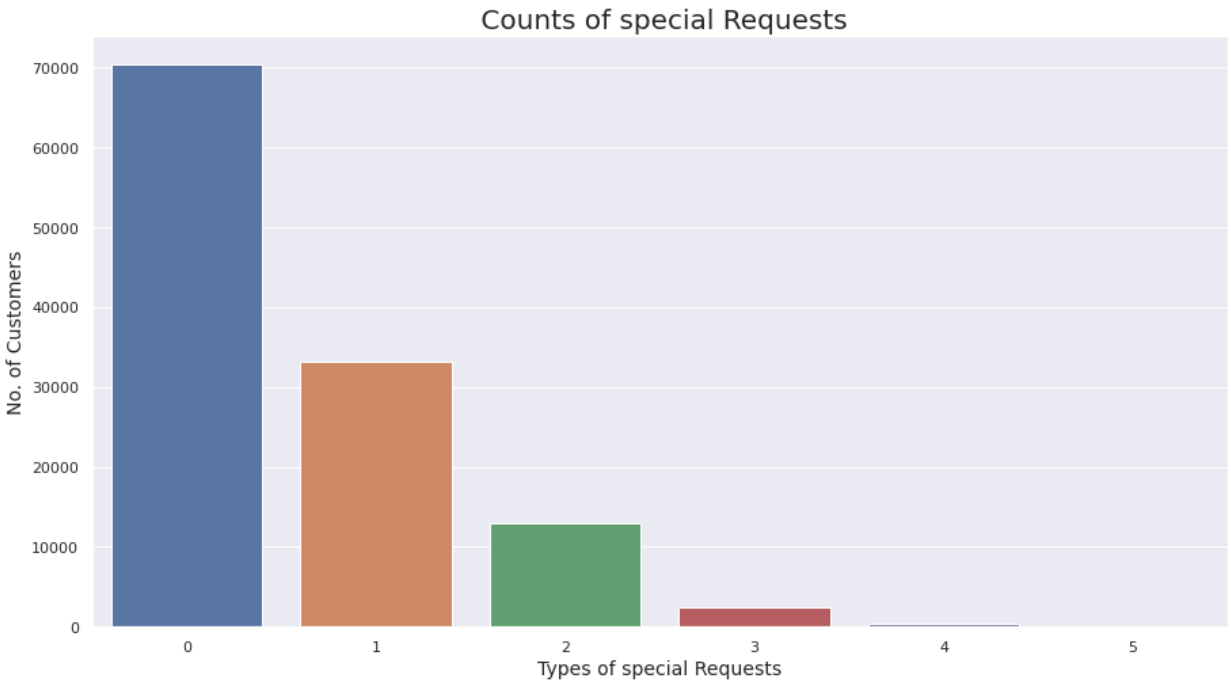
★ Special Requests for different Accommodation type

Special Requests here signify (e.g. twin bed or high floor) some personalized change in his/her booked room. From the figure below we can see that couples have made the most requests in both the hotels and singles have made the least requests.



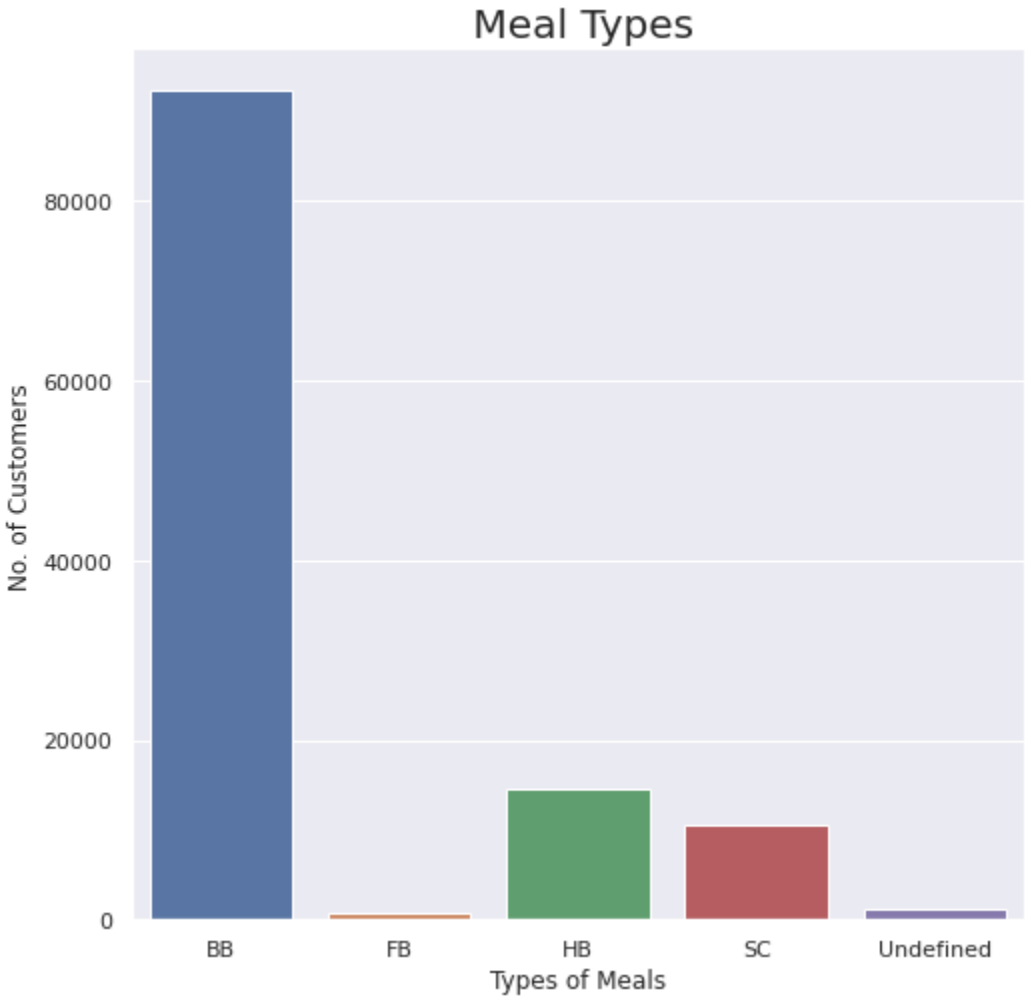

★ Special requests across various room types

The fig below signifies shows that room type E had the highest special requests and room type L had the least .




★ Most Prefered Meal Type

From the figure below it can be inferred that meal type BB(bed and Breakfast) was the most preferred and FB (Full Board) was the least.



## 4. Conclusion

- From this analysis we came to the conclusion that there were some factors which effect bookings and cancellations in a significant way, whereas some of the factors like Meal type, Day of booking, did not matter that much which we dropped for our analysis.
- **Inferences:**
    a. People prefer City Hotels more as compared to resort hotel, as these hotels are cheaper than resort hotels, whereas if people are coming for longer duration they prefer resort hotels over city hotels.
    b. Most of the people who travelled were from European countries, who preferred to travel more in summer and spring season which increased the revenue of hotels significantly in these months. If any hotel want to increase their yearly turnover, they can hike their prices in these months as demand is more in these months
    c. Couples travel more than solo and family type of guest, so hotels can come up with more couple friendly policies which can help them in increasing their revenue.
    d. Average Daily Rate/Revenue of hotels increases if the family comes to travel.
    e. Lead Time is one of the major factor of cancellations, more the lead time, more is the cancellations, so to counter this what hotels can do is they can only allow bookings one or two month prior to guest arrival.