Capstone Project
Exploratory Data Analysis: Hotel Booking

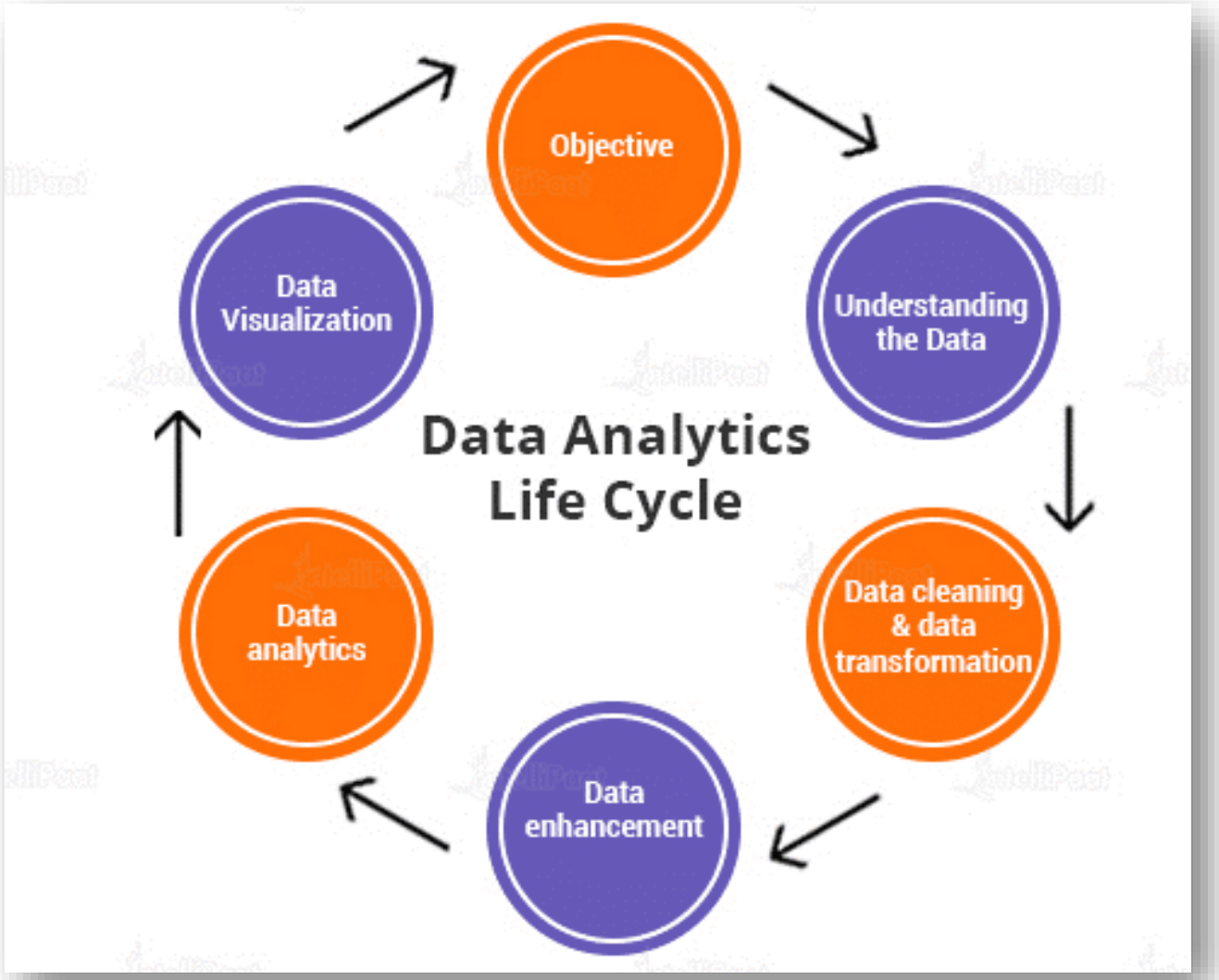
FAISAL JAMAL

AlmaBetter

Table of contents

AlmaBetter

Data Analytics Life Cycle



Problem Statement/ Objective

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

This project aims to create meaningful estimators from the data set we have and to perform Exploratory so as to get business insights and use them to rectify problems and boost organization.

Understanding the Data

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 38 other things within 119390 rows.

Their are total 32 variables. Most columns have the data type object, either because they contain values of different types or contain empty values (NaN).

The variables in dataset are:

'hotel', 'is_canceled', 'lead_time', 'arrival_date_year', 'arrival_date_month', 'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'meal', 'country', 'market_segment', 'distribution_channel', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'assigned_room_type', 'booking_changes', 'deposit_type', 'agent', 'company', 'days_in_waiting_list', 'customer_type', 'adr', 'required_car_parking_spaces', 'total_of_special_requests', 'reservation_status', 'reservation_status_date'

Understanding the Data (contd..)

The unique values in relavent categorical variable are

- hotel ['Resort Hotel' 'City Hotel'] .
- arrival_date_year [2015 2016 2017] .
- meal ['BB' 'FB' 'HB' 'SC' 'Undefined'] .
- market_segment ['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
- 'Undefined' 'Aviation'] .
- distribution_channel ['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS'] .
- reserved_room_type ['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B'] .
- deposit_type ['No Deposit' 'Refundable' 'Non Refund'] .
- reservation_status ['Check-Out' 'Canceled' 'No-Show'] .

Understanding the Data (contd..)

Dataset contains three type of data type:

- **Categorical:**

'hotel', 'meal', 'country', 'market_segment', 'distribution_channel', 'arrival_date_day_of_month', 'reserved_room_type', 'assigned_room_type' , 'reservation_status', 'deposit_type', 'customer_type'

- **Numerical:**

'lead_time', 'arrival_date_year', 'arrival_date_month', 'arrival_date_week_number', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'previous_cancellations', 'previous_bookings_not_canceled', 'booking_changes', 'agent', 'company', 'days_in_waiting_list', 'adr', 'required_car_parking_spaces', 'total_of_special_requests', 'reservation_status_date'

- **Binary:**

'is_canceled', 'is_repeated_guest'

AlmaBetter

Data Cleaning and Data Transformation

The columns containing the null values are company (112593), agent (16340), country (488),children (4).

We dropped the column company, remove 4 rows of children containing null value, fill null value of country with other, and filled null values of agent with 0.

There are total 31994 duplicate values that we dropped from the dataset.

```
[ ]  df_hotel.isnull().sum().sort_values(ascending = False)[0:6]

company                 82137
agent                   12193
country                   452
children                    4
reserved_room_type          0
assigned_room_type          0
dtype: int64
```

```
[ ]  print(len(df_hotel[df_hotel.duplicated()]))

31994
```

```
[ ]  df_hotel.drop_duplicates(inplace = True)
```

AlmaBetter

Data Enhancement

- Added column of total number of guests by adding the adults, children and babies
- Added column total night stay by adding stay in week nights and stay in weekend nights
- Use arrival year, month, day of month to add arrival date and arrival day column.

```
# Adding a new column total number of guests
df_hotel['total_number_of_guests'] = (df_hotel['adults'] +df_hotel['babies']+ df_hotel['children']).astype(int)

# Adding a new column total number of night stay
df_hotel['total_night_stay'] = (df_hotel['stays_in_week_nights']+df_hotel['stays_in_weekend_nights']).astype(int)
```

Agenda

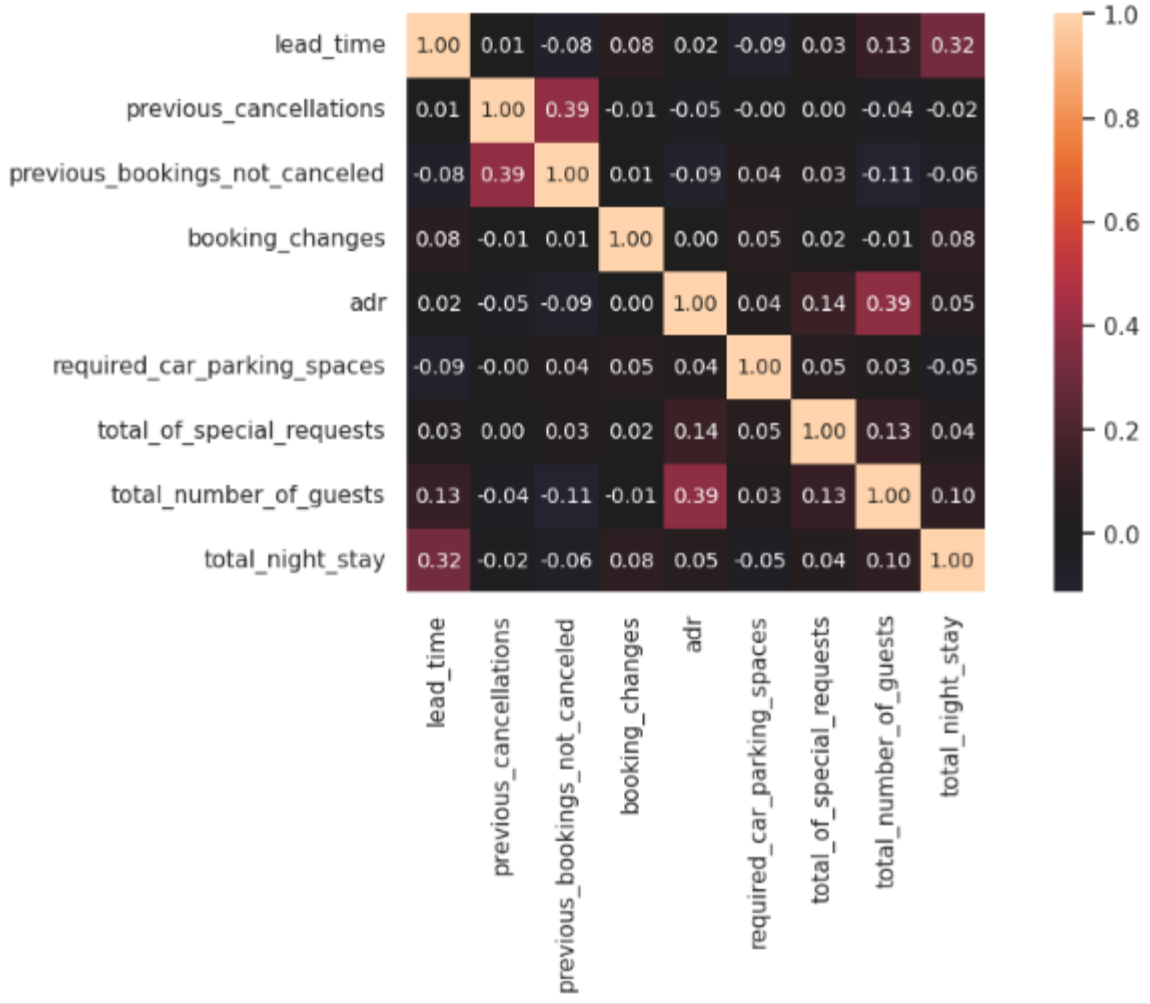
With the exploratory data analysis of the hotel booking dataset we will try to answer the following questions:

- Where do the guests come from?
- How many bookings were canceled?
- Which are the busiest month?
- Bookings by market segment and distribution channels
- Hotel type guests prefer.
- How long do people stay at the hotels?
- Repeated guest.
- The number of nights spent at hotels.

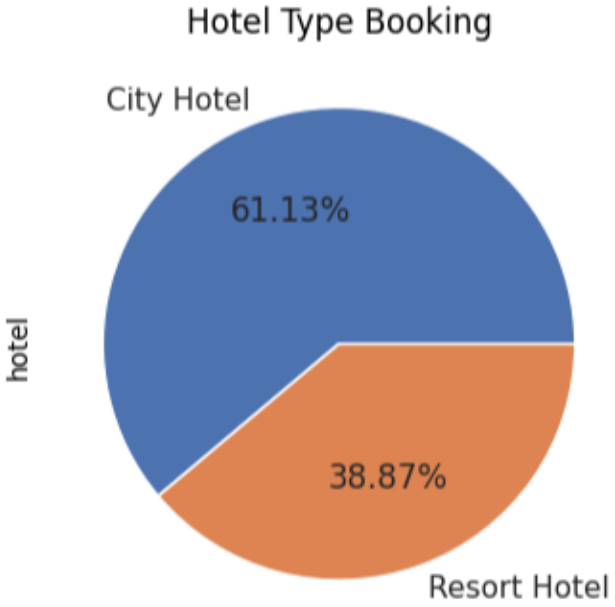
Correlation Heatmap

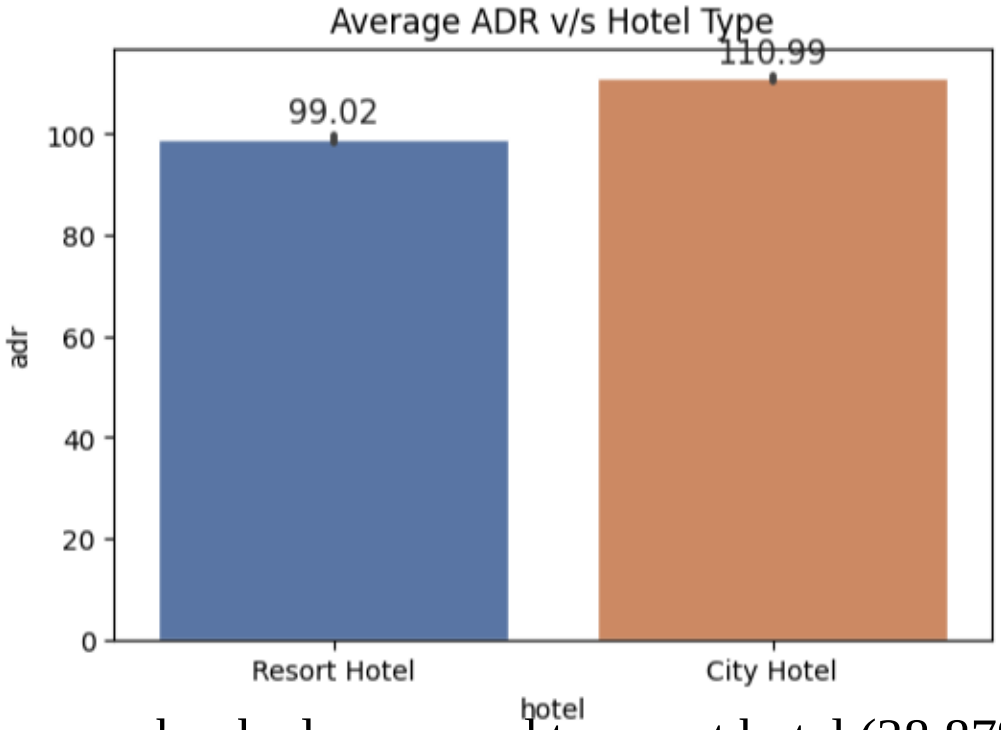
The heat map is selected as it can show the correlation between a number of variable at once. From the above heat map we can find that

- Total stay length and lead time are slightly
- correlated(0.32).This may means that for longer hotel stays, people generally plan little before the actual arrival.
- ADR is slightly correlated(0.39) with total_people, which makes sense as more no. of people means more service to deliver, therefore more adr.
- Moreover the correlation between total number of guests & previous booking not cancelled is small and negetive -0.11.



Hotel wise Analysis





- We can conclude that city hotels (61.13%) are more booked compared to resort hotel (38.87%).
- The average ADR for city hotel is more than resort hotel which is around 11.
- As resort hotels are comparatively expensive hence people books city hotel more frequently.

Time Wise Analysis Of Hotel Booking

While doing time-wise analysis of given hotel booking dataset, we answered following questions:

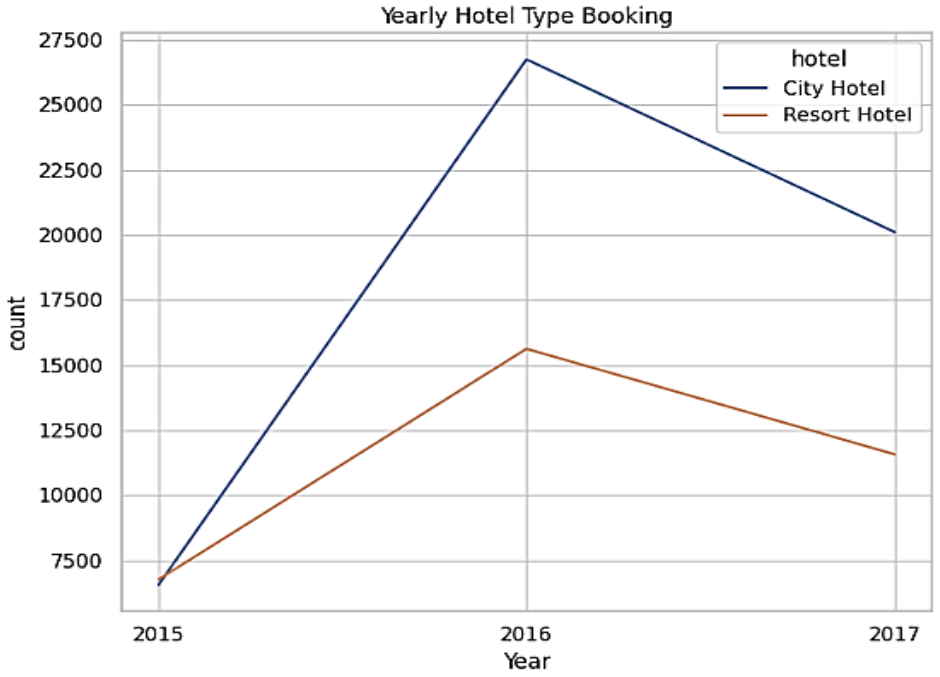
(1) What are the most busy months for hotels?

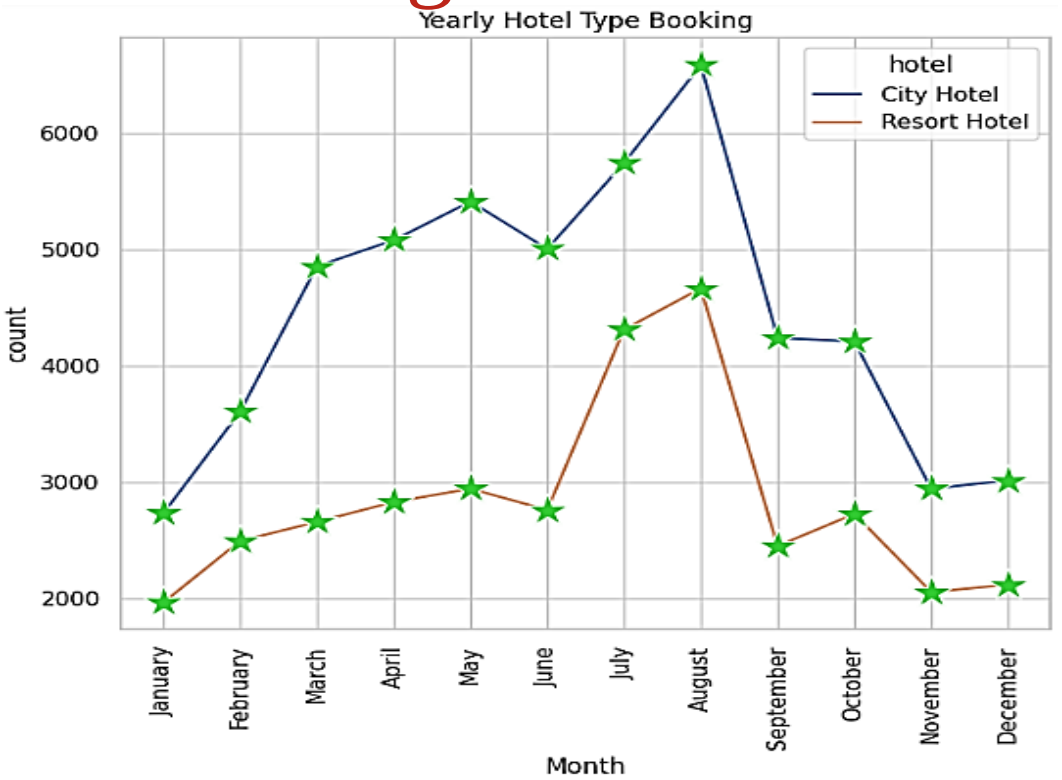
(2) In which months hotels charges higher adr?

(3) How does booking numbers and adr changes within a month?

(4) How does bookings varies along year for different types of customers

Time Wise Analysis Of Hotel Booking



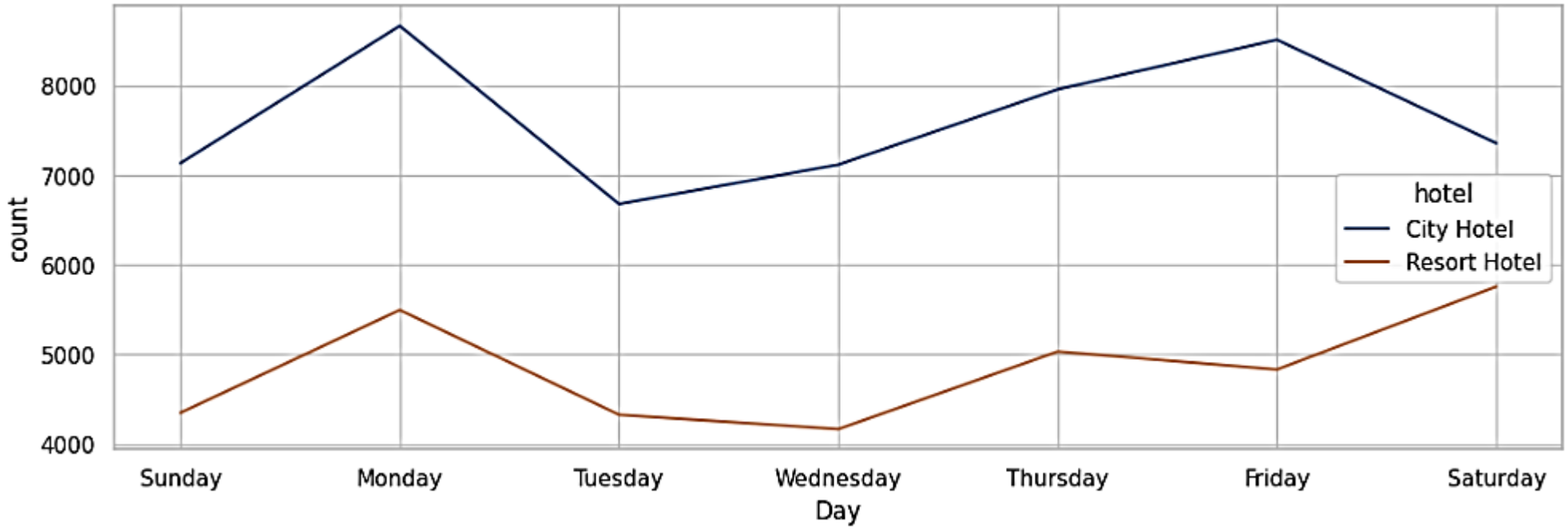


This should also be noted that we only had data of last six months for the year 2015.

The hotel bookings has increased during the three year period, with maximum in 2016 and a little decrease in 2017.

Maximum booking are in the month of July and August the reason behind it can be holiday season in various countries and summer break in colleges and schools. Moreover the least are in the month of January, November and December the reason can be extreme cold and the Christmas holidays which people enjoy at their family homes.

Time Wise Analysis Of Hotel Booking

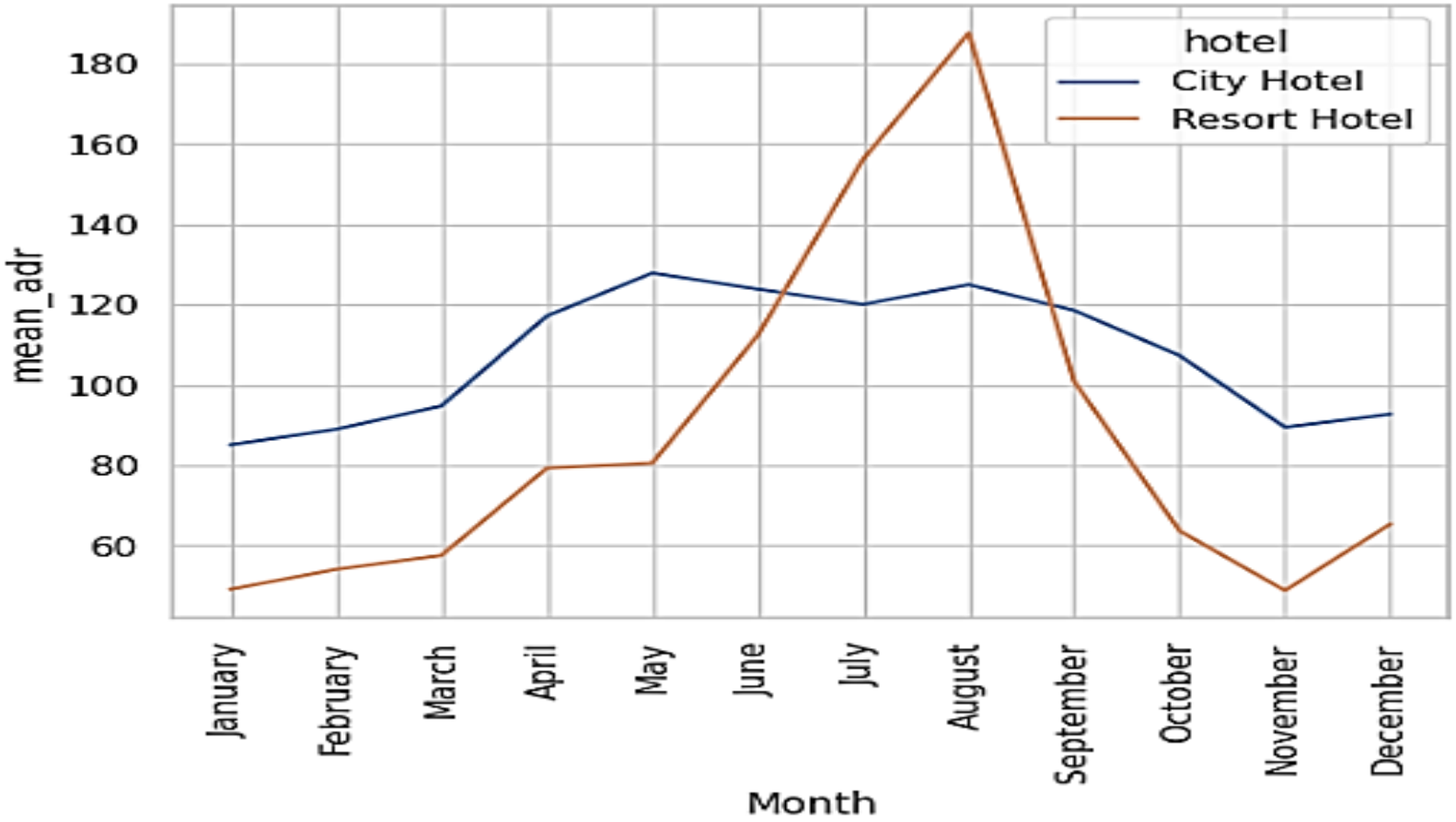


For city hotels most busy days are Monday, Friday and least is Tuesday.

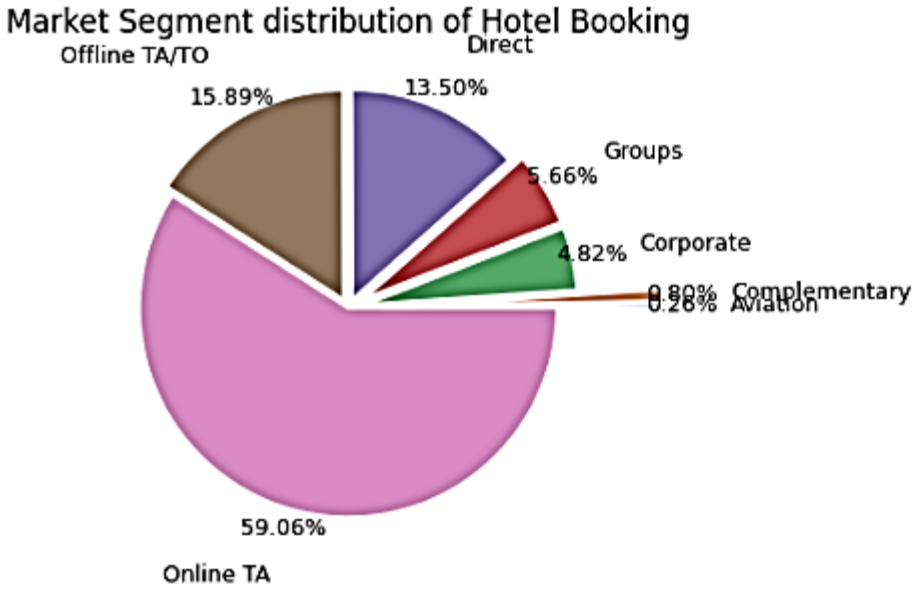
For resort hotels most busy days are Saturday, Monday and least is Wednesday.

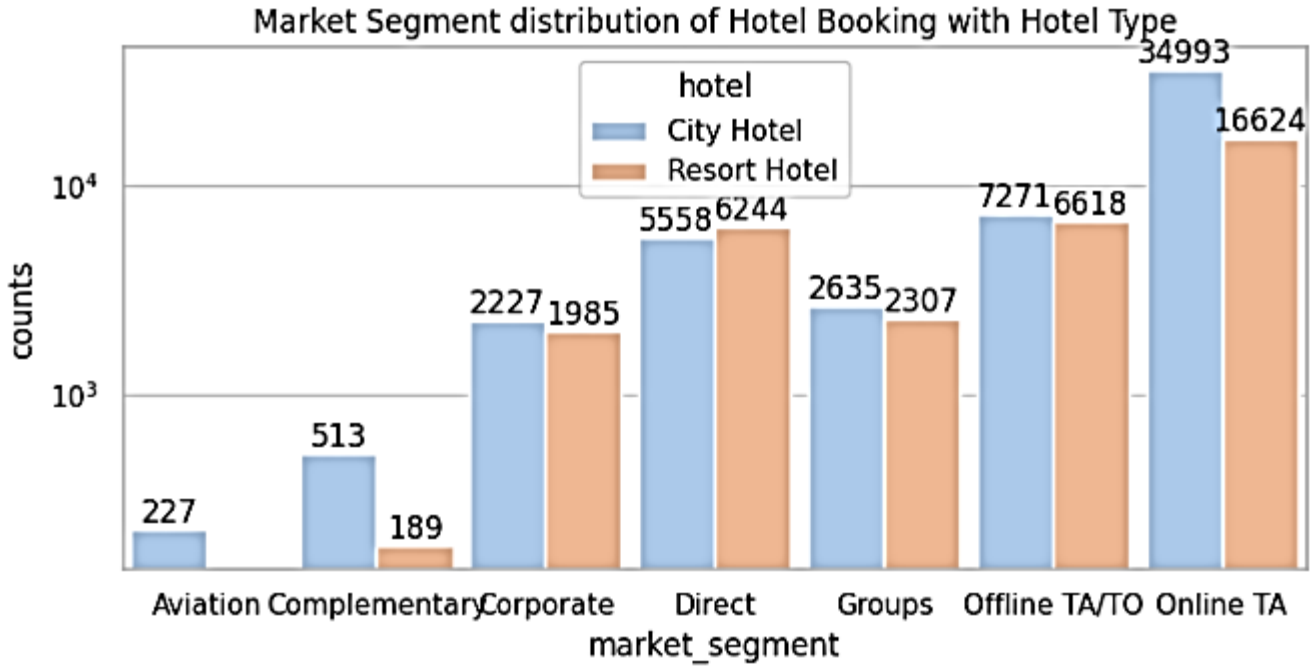
Time Wise Analysis Of Hotel Booking

The revenue aspect looks different, the Resort Hotels receives more revenue with respect to City Hotel. From May to August there was rapid increase in adr. August recorded the highest



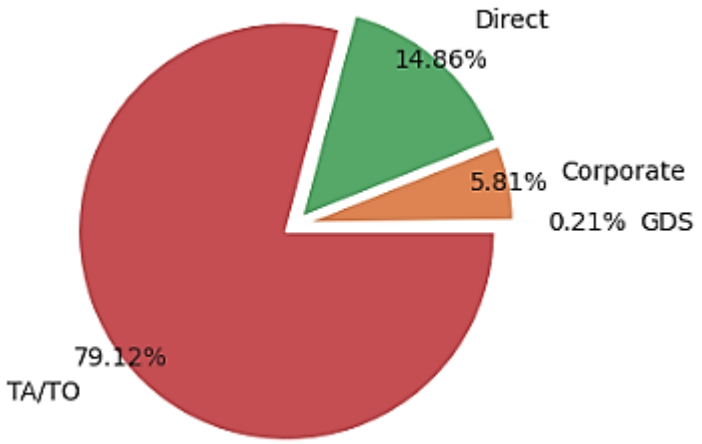
Market Segment wise Analysis

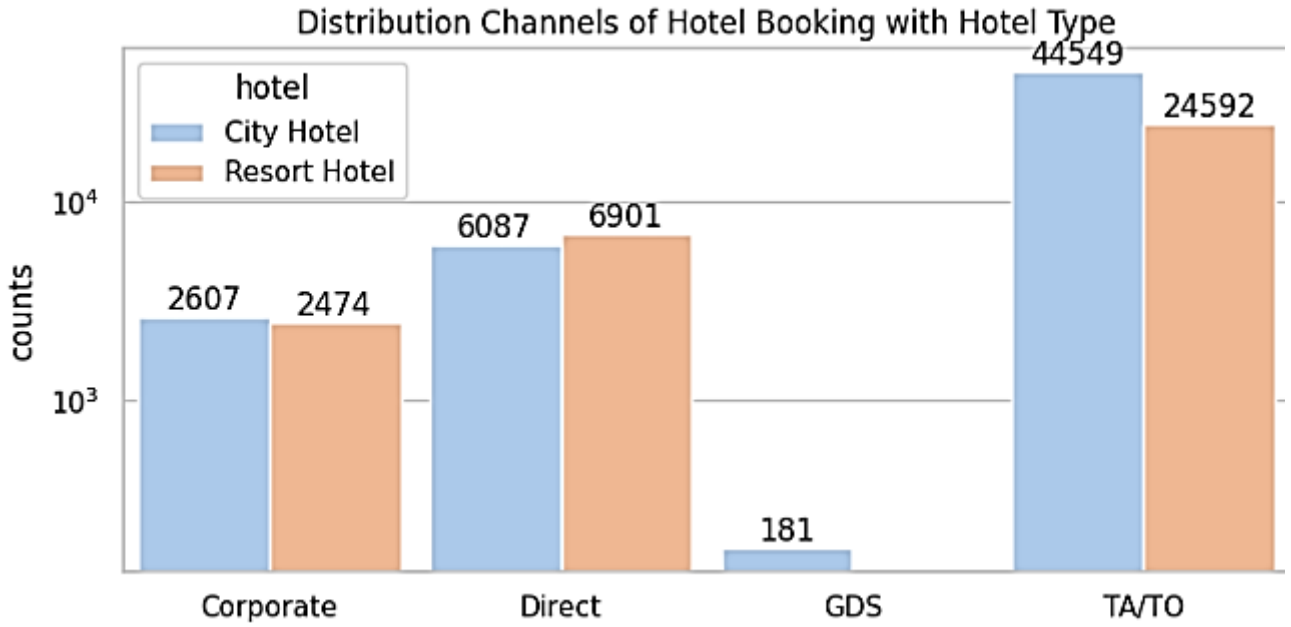




- The most bookings are from online TA around 60% followed by offline TA/TO (15.89%) and then Direct booking. Furthermore the least hotel booking are from complementary segment(0.80%) and Aviation which is only 0.26%.
- From the bar chart that follows same trends as pie chat and depict market segment with hotel type. One thing can be seen that their is no resort hotel book through Aviation.

Distribution Channel wise Analysis





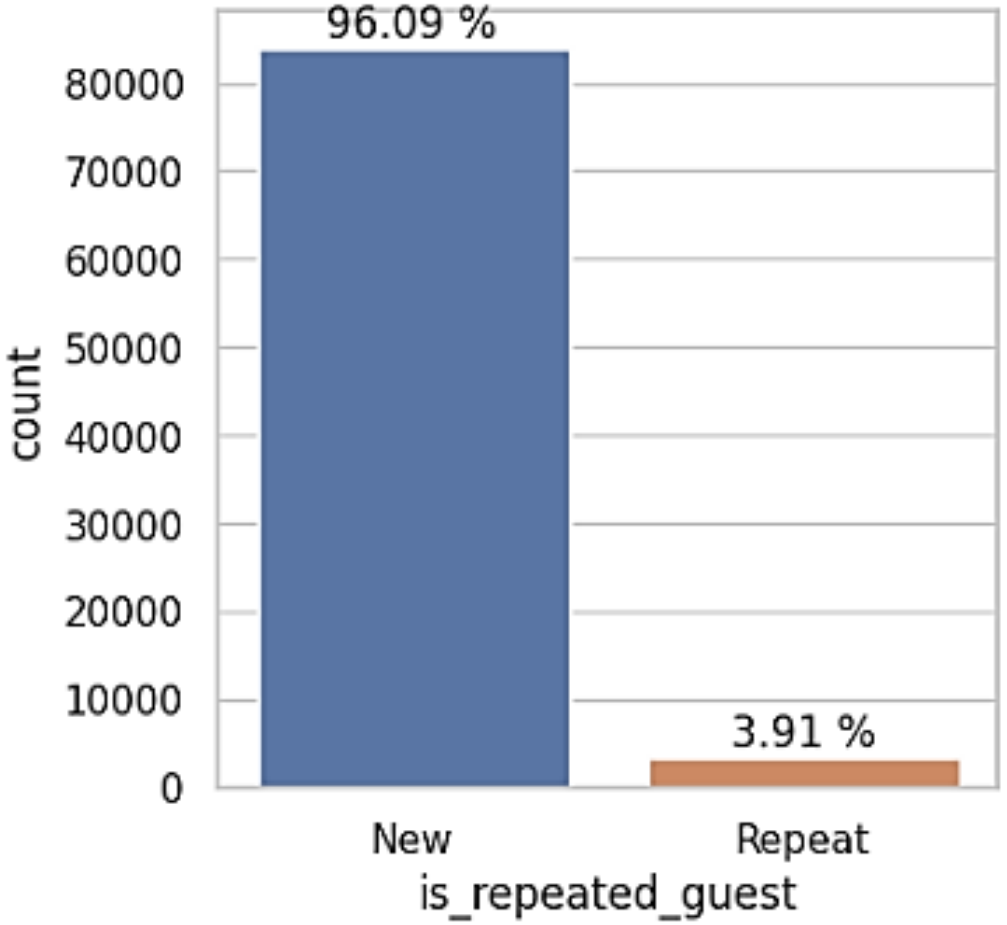
- Out of all the hotel booking just a little less than 80% is done by TA/TO (Travel Agent/ Tour Operator) hence we shold be focused on this. After this Direct (14.86%), Corporate(5.82%) and GDS (0.21%) distribution channel is used.
- The bar chart for the distribution channels hotel type wise shows the distribution of the hotel booking on hotel type. It can be seen that GDS is not opted by anyone for resort hotel.

Customers Retention Analysis

Low number of repeated guests.

Only 3.91% out of all are repeated customers

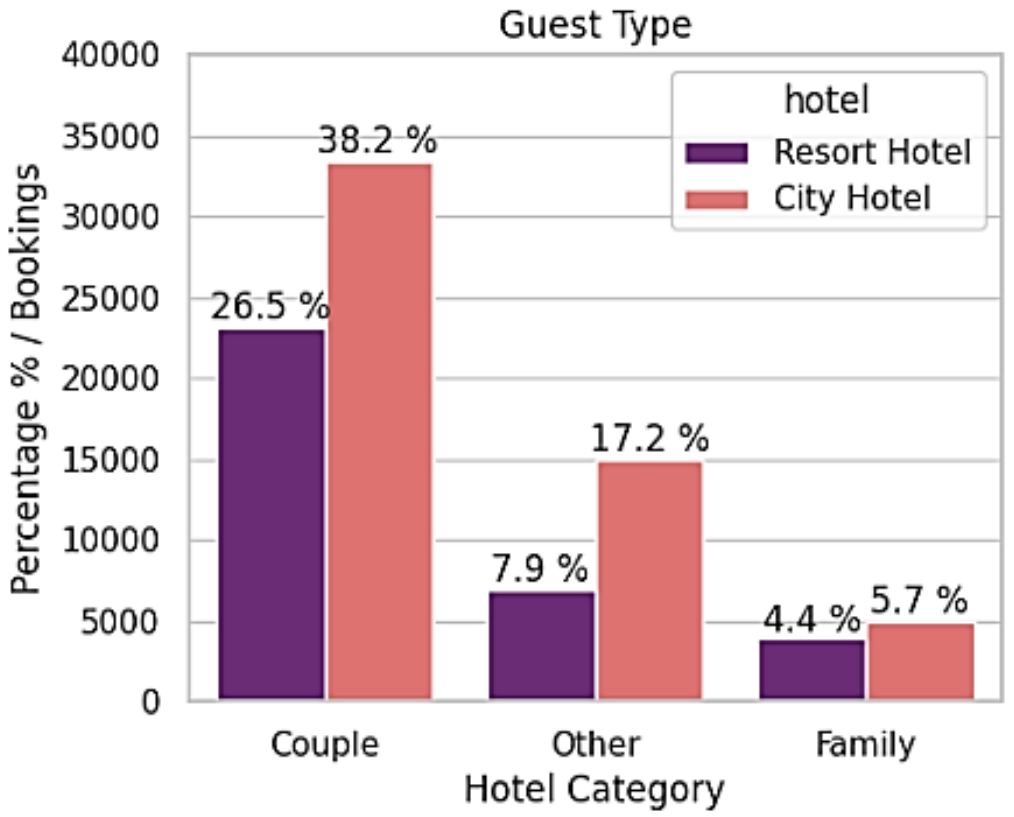
A need to target repeated guests since retaining a customer is way less expensive than gaining a new one.



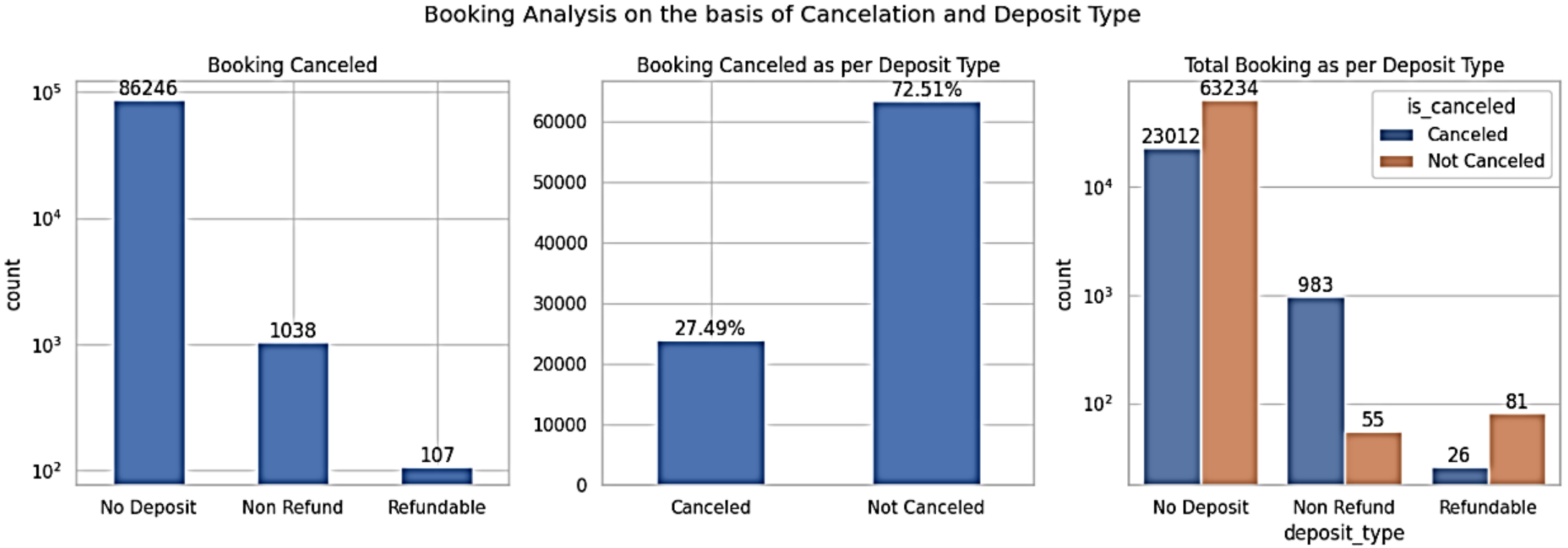
Hotel preference analysis for different categories of customers

The assumptions for the above are:

- When total number of guests are two we categories them as Couple.
- When total adults are not 2 and without children or babies are categories as other.
- Guests with children or babies are categories as family.
- As seen in the above chart most number of guests are couples then other followed by family.
- With all of them booking city hotel more frequently but the preference for family type in resort and city hotels is very close.

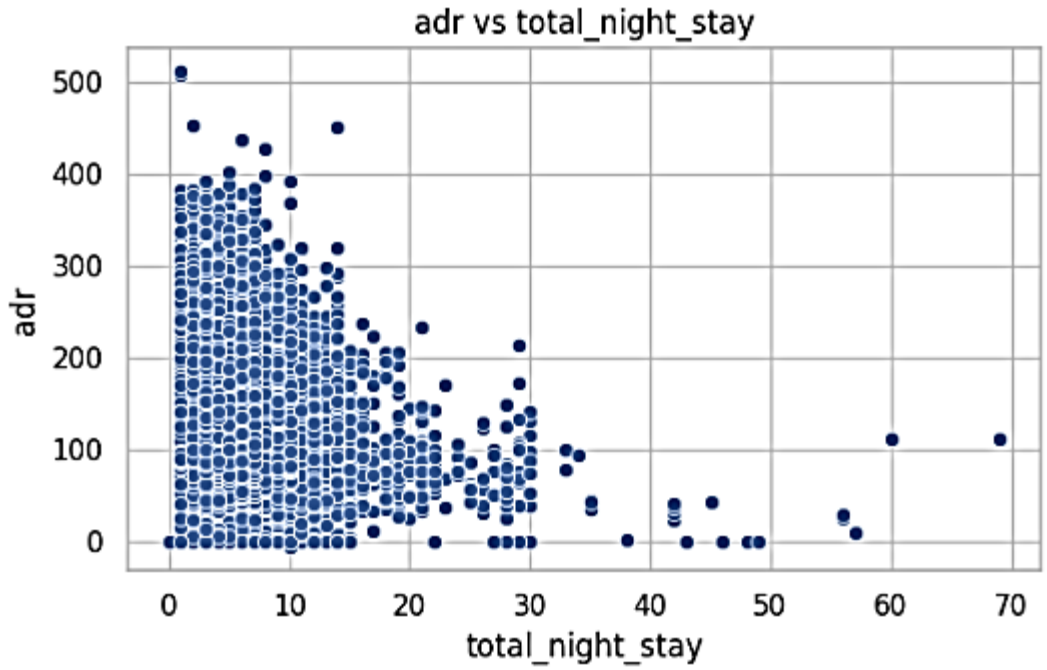


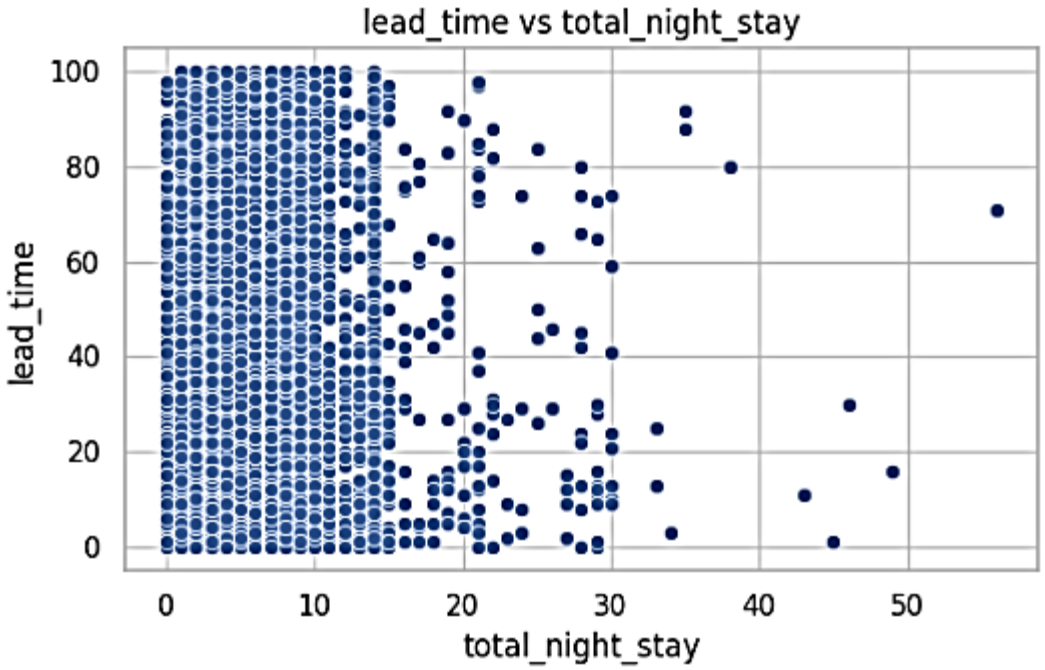
Analysis on the basis of Cancelation and Deposit Type



- Most bookings are of type No Deposit followed by Non Refund and Refundable.
- Out of all the booking total 27.49% are being canceled which is quite a high cancel rate driving revenues downward.
- Most canceled booking are from No Deposite category as it contributes to most bookings(86246).
- 90% booking in non refund type are being canceled even though it will not reduce revenue but need to check the cause of same.

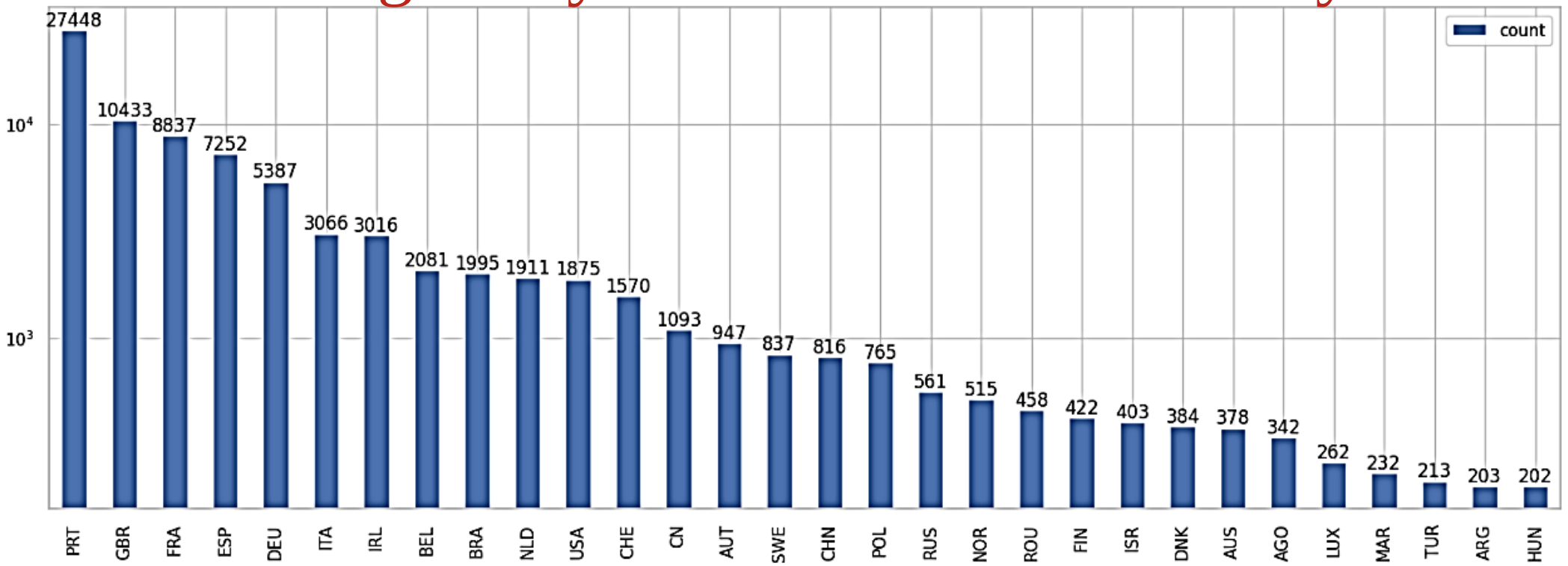
Analysis of Total Night Stay with ADR and lead time





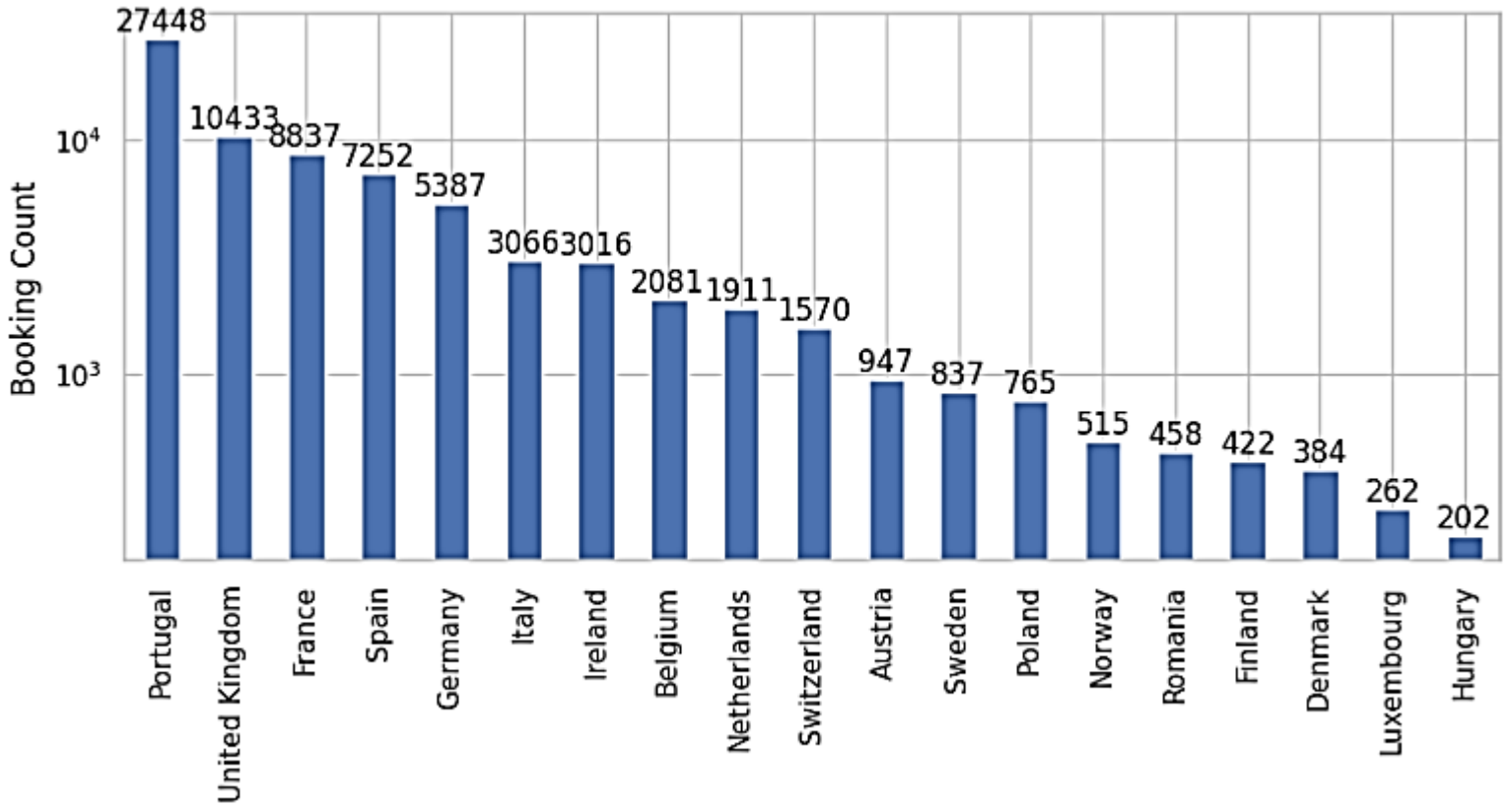
- As from the graph the ADR decreases with increase in the number of night stay showing how customers will get a better deal for longer stayes.
- For lead time and total night stay their is no such corelation for practical situations (night stay< 50days & lead time < 100days) but for whole dataset it shows a inverse relationship.

Hotel Booking Analysis on the basis of country



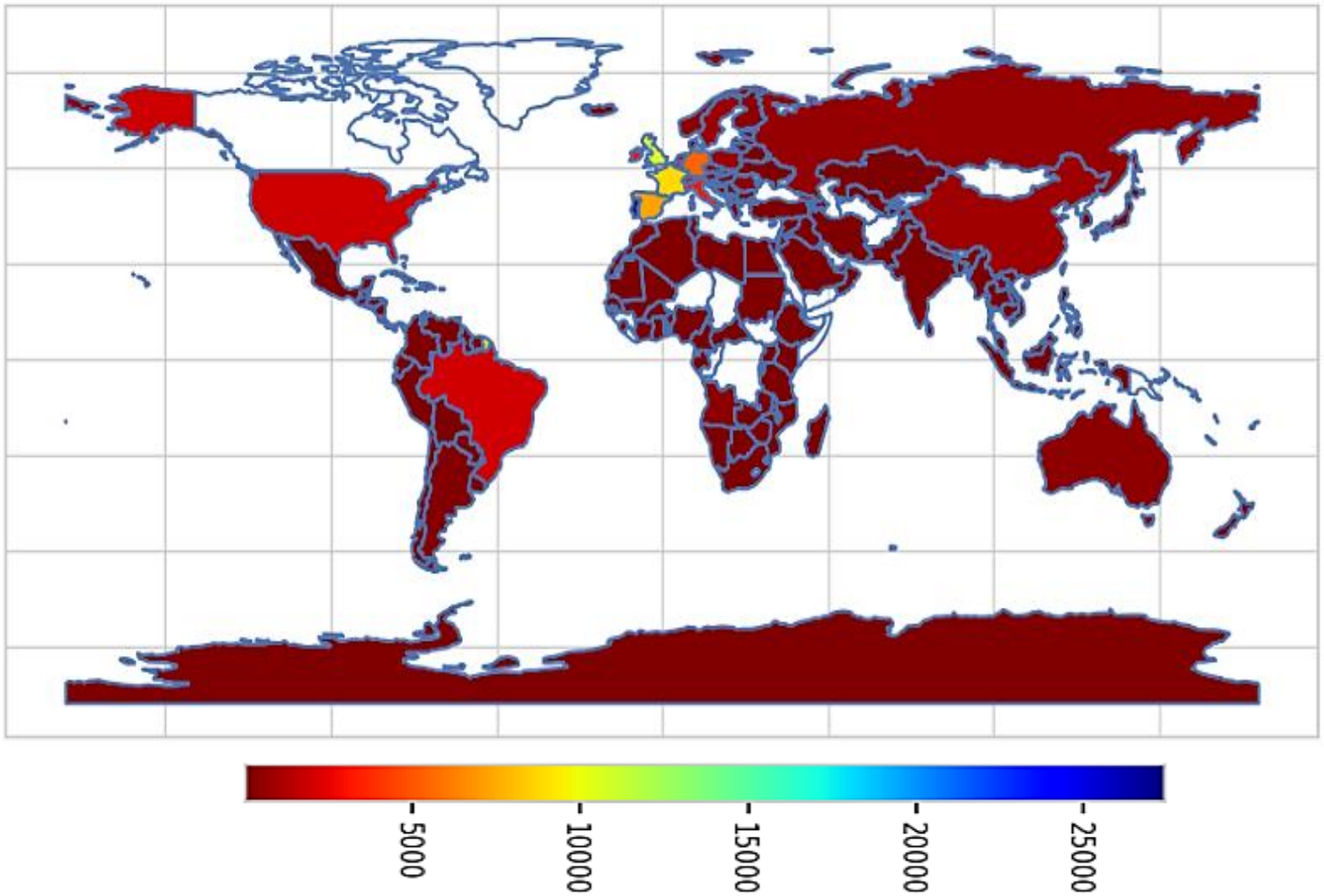
- Portugal, United Kingdom, France, Spain are the countries with most customers.
- Moreover the above chart gives us the total hotel bookings in different countries in descending order.

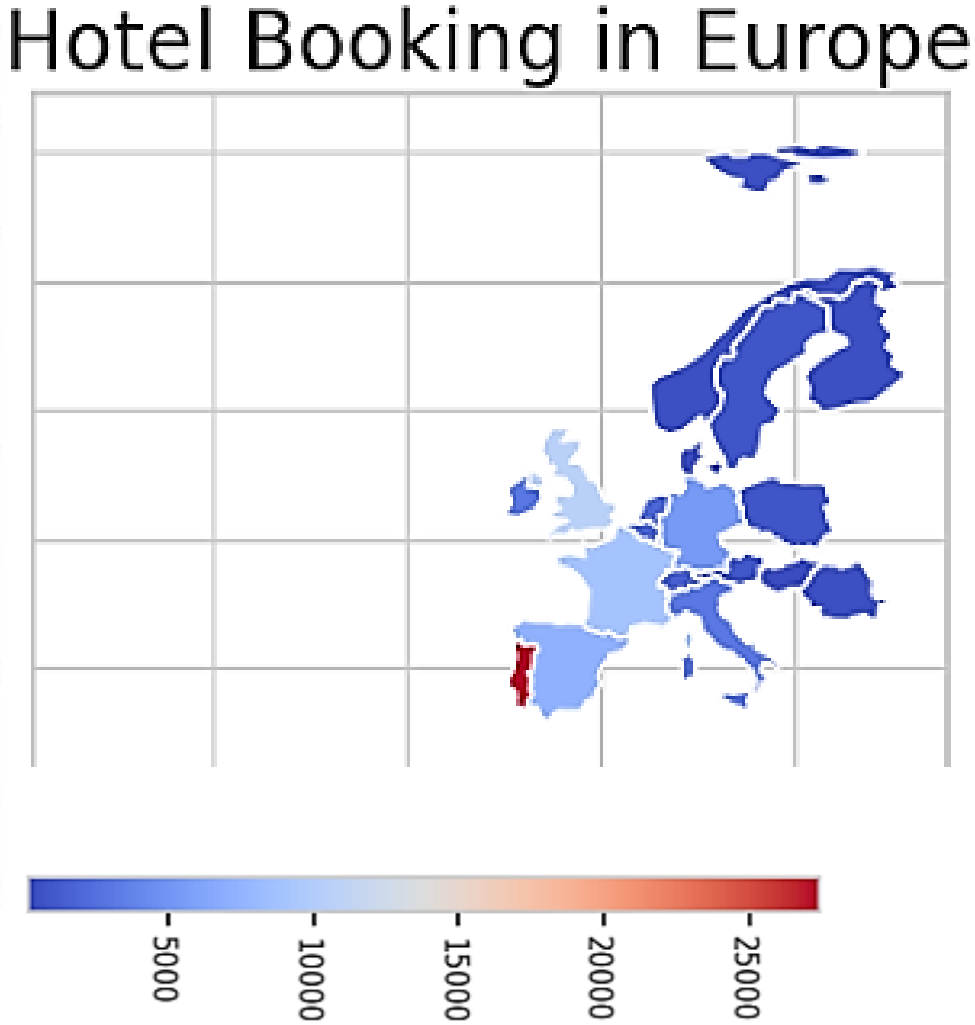
Hotel Booking Analysis for Europe



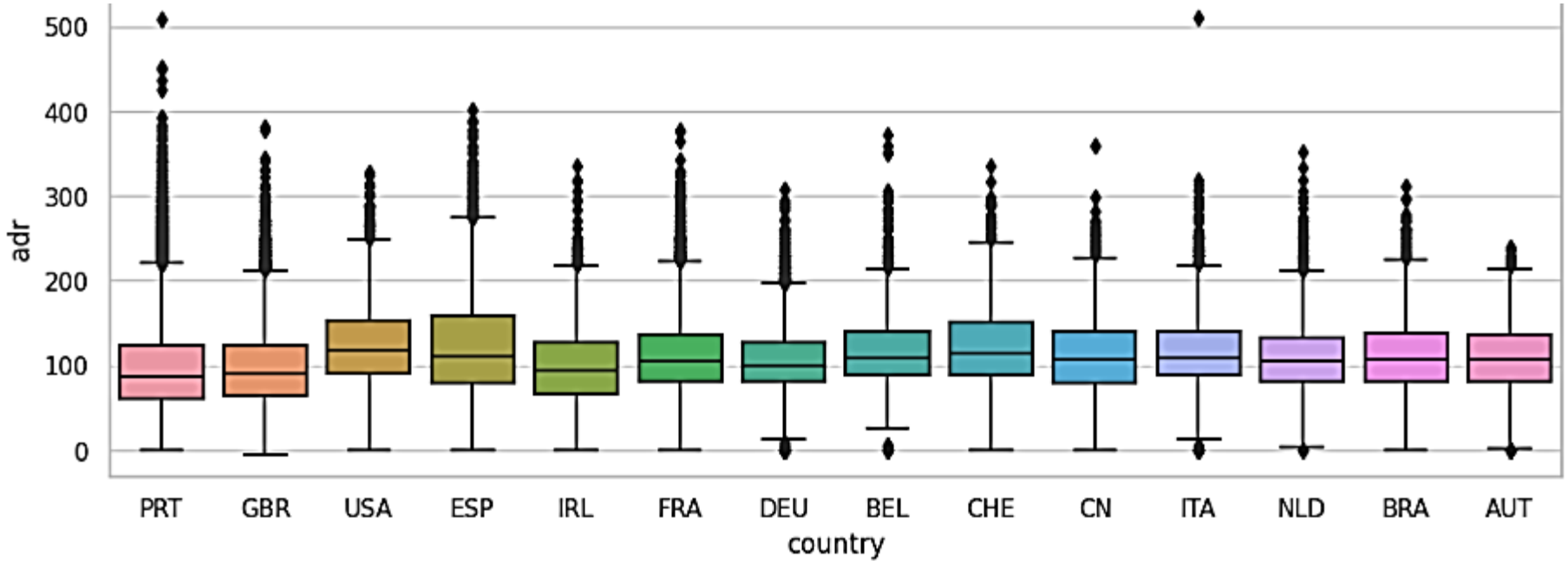
- The above bar chart gives us the number of bookings in european countries. It can be seen most bookings are from Portugal, United Kingdom and France which is same for whole data.

Hotel Bookings on Map



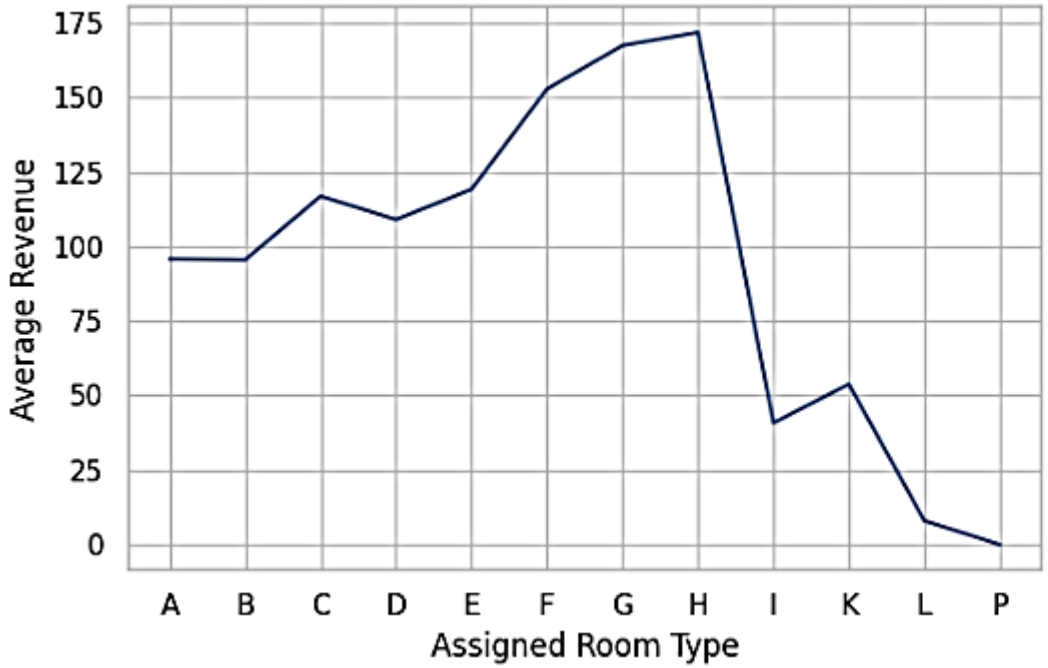


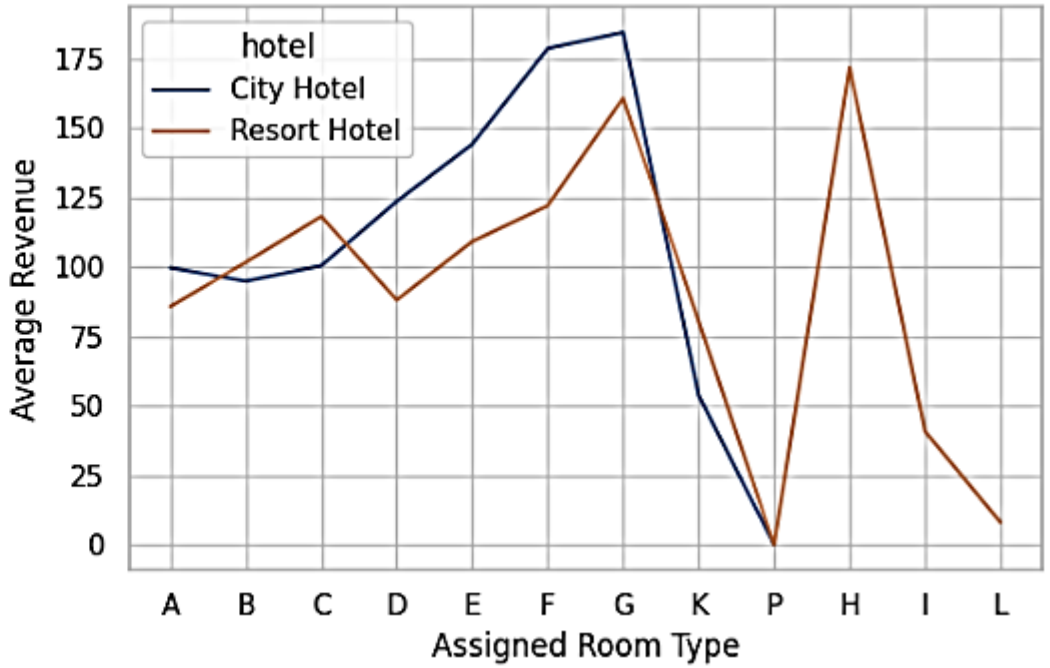
Country wise Analysis of ADR



- The country wise mean ADR is around 100 for countries with total ADR> 100000 and excluding all the outliners having ADR >600.
- It can also be seen after using the appropriate filters still their are outliners.
- Switzerland(CHE),Belgium(BEL) and Spain(ESP) has highest Average ADR
- Most above countries have average ADR around 100

Average ADR Analysis for assigned room type

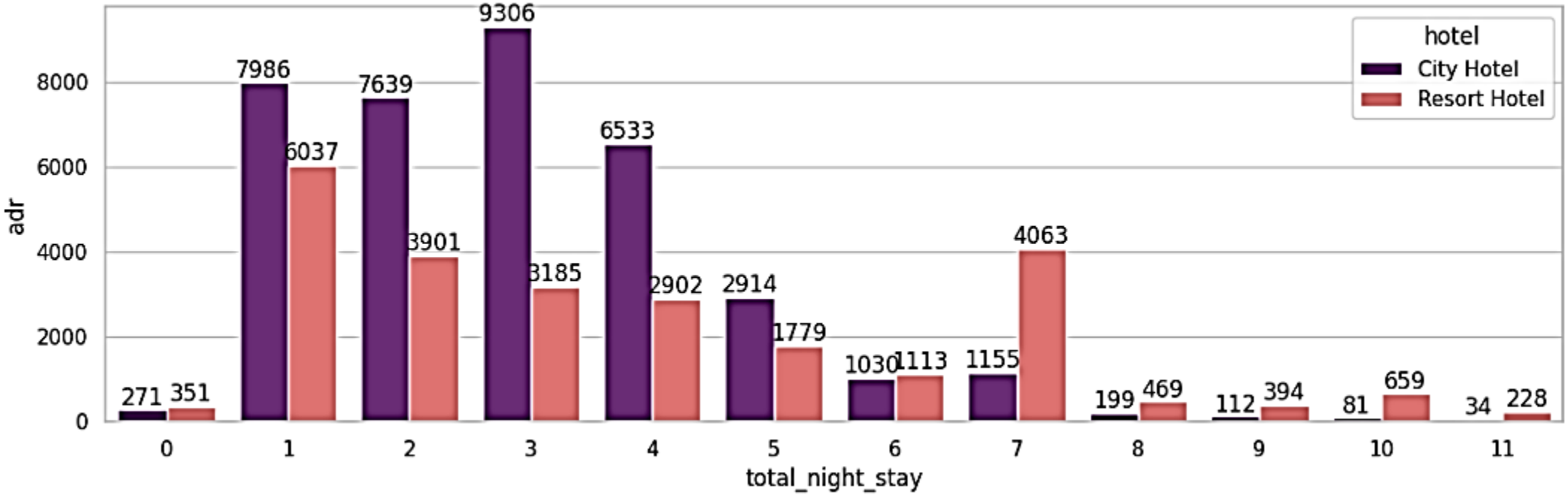




- Overall room type H followed by G yield most average ADR, Meanwhile P and L contributes toward least.
- Comparing same for the room type. In City Hotel the maximum is for G type followed by F and for Resort Hotel the maximum is for H type followed by G. In both the cases the minimum average ADR is for P type rooms.
- It is to be noted their is no H, I, L type rooms for Resort Hotel.

AlmaBetter

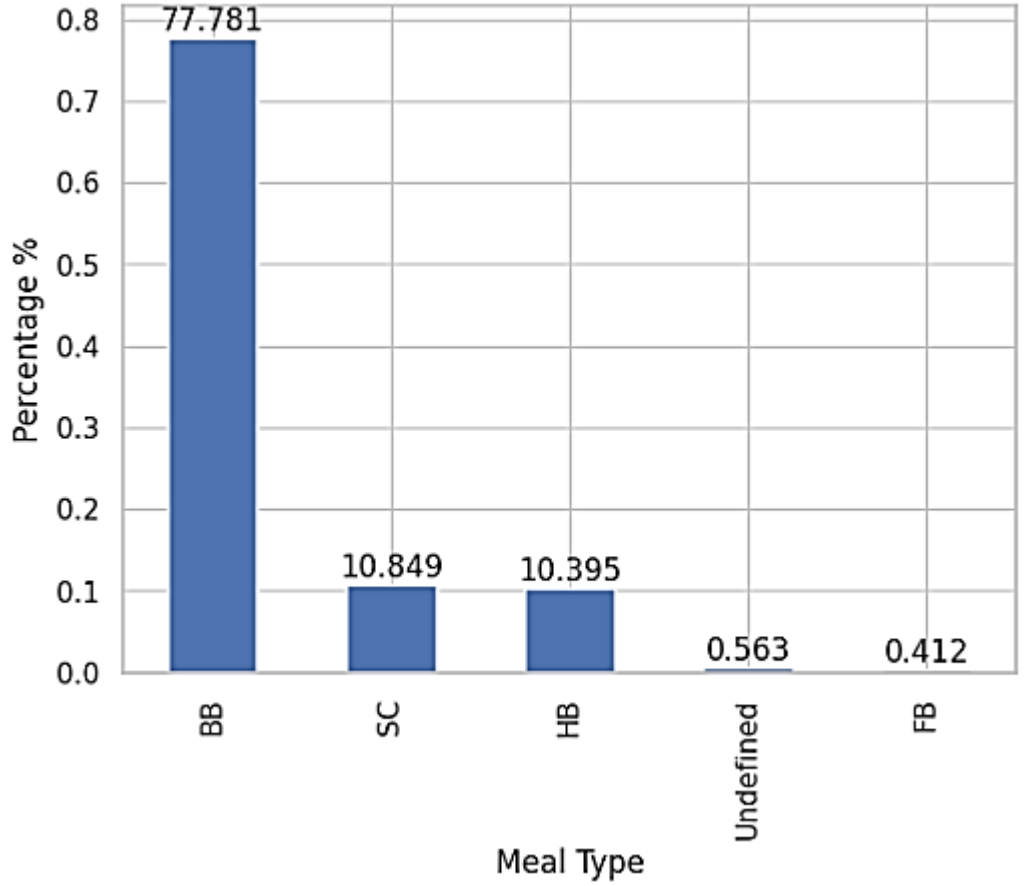
ADR Analysis on the basis of Total Night Stay



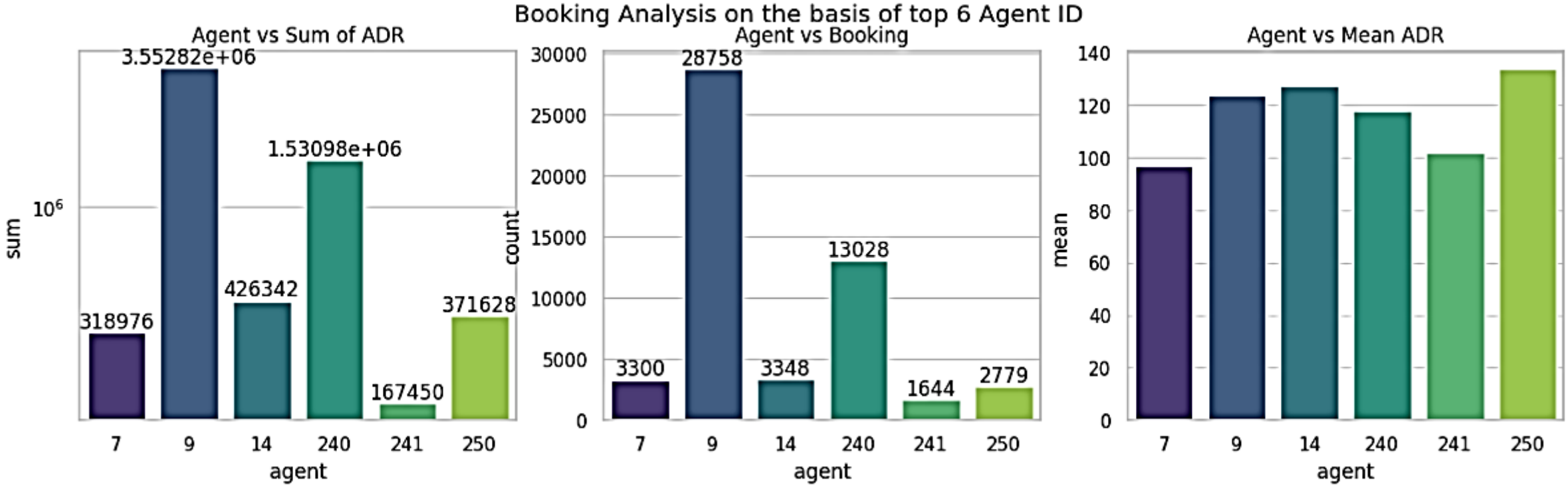
- For city hotel three night stay followed by one gives us most revenue.
- In the case of resort hotel one night followed by seven and two gives most revenue.
- This can also be inferred that with longer night stay the customers usually get better deal hence the ADR decreases with longer stay.

Customer Meal choice Analysis

- Out of the meals, BB (Bed & Breakfast) is the most ordered meal which is around 77.8%, followed by HB(Half Board), SC(no meal package), Undefined and FB (Full Board).



Agent wise Analysis for ADR and Booking counts



- The 6 best performing agents IDs are 9, 240, 14, 250, 7, 241 respectively for sum of revenue generated.
- The 6 best performing agents IDs are 9, 240, 14, 7, 250, 241 respectively for total number of bookings that are shown in pie chart 2.
- The 6 best performing agents IDs are 250, 14, 9, 240, 241, respectively for mean ADR in the time period of 3 years.

Inferences and Conclusion

We've drawn many inferences from the above analysis.

- City hotels(61.13%) receive more guests throughout the year compared to Resort hotels(38.87%).
- The average ADR for city hotel is more than resort hotel which is around 11.
- The booking has increased during the three year period, with maximum in 2016 and a little decrease in 2017. Also, The maximum booking are in thr month of July and August the reason behind it can be holiday season in various countries, weather good for vacction and summer in colleges and schools. Moreover the least are in the month of January, November and December.
- With longer night stay the customers usually get better deal hence the ADR decreases with longer stay.
- Online TA booking are around 60% followed by offline TA/TO (15.89%) and then Direct booking (13.5%). Promotions and offers should be used to increase the direct booking.
- Out of all the hotel booking around 80% is done by TA/TO (Travel Agent/ Tour Operator) hence we should be focused on this and work toward increasing Direct Distribution channel (14.86%).

Inferences and Conclusion (contd..)

- Out of all the booking total 27.49% are being canceled hence seting Non-refundable Rates, Collect deposits, and implement more rigid cancellation policies can tackle this.
- Most number of guests are couples and only families prefer both city as well as resort hotels almost equally.
- In all only 3.91% are repeated guests, Low number of repeated guests. Hotels need to target repeated guests since retaining a customer is way less expensive than gaining a new one.
- It appears that a disproportionately high number of bookings are from Portugal, probably because the hotel is located in Portugal itself. The second country is the United Kingdom which is approx. 75% behind.
- Room type H followed by G yield most average ADR, Meanwhile P and L contributes toward least.
- Out of the meals, BB (Bed & Breakfast) is the most ordered meal which is around 77.8% hence hotels should put special attention to this.
- The six best performing agents are with IDs as 9, 240, 14, 250, 7, 241.

AlmaBetter

Thank You