Hotel Booking Analysis

Sarath Soman
Hariharapanda Deepak
Shyam Gadekar
Shrikant Kute

Data science trainees,
Alma Better, Bangalore

Abstract:

The hotel industry is a volatile market which changes seasonally and depends on various factors. To understand the business context of the hotel industry we need to understand the effect of different parameters affecting the hotel performance like when the booking was made, length of stay, number of persons staying, etc. For the period of 2015 July to 2017 August.

Firstly, we explored and inspected the provided dataset. Secondly cleaning of the dataset was done which included finding duplicate rows, null values and outlier in the dataset. Then we did Exploratory Data Analysis (EDA) on the dataset to get correlation between each attribute in the dataset. With the help of data visualization, we targeted different attributes of the dataset to represent the categorical and numerical data in graphical representations to understand it more clearly. Analysis of all the obtained data was done to find out the possible key factors behind hotel booking and recommendations were made.

Keywords: data cleaning, data inspection, data visualization

1. Problem Statement:

Understanding the effect of different parameters affecting the hotel performance like when the booking was made, length of stay, number of persons staying, etc. For the period of 2015 July to 2017 August.

Performing Uni-variate, Hotel Wise, and Distribution Channel wise, Lead and waiting time Analysis and booking cancellation analysis for the data.

Find out the key factors driving the hotel booking trends and make decisions based on them.

2. Dataset information:

For understanding the data provided by the company we can look at each variable and try to understand their meaning and relevance to this problem.

Attributes in detail:

hotel: type of hotels
is_canceled: cancelled or not
lead_time: no. of days before actual arrival in the hotel
arrival_date_year: year of booking
arrival_date_month: month of booking
arrival_date_week_number: week number of the year in which booking
arrival_date_day_of_month: arrival month date
stays_in_weekend_nights: no. of weekend's guest stayed
stays_in_week_nights: no. of weekday's guest stayed
adults: no. of Adults staying in the hotel
children: no. of children staying in hotel
babies: no. of babies staying in the hotel
meal:
BB- Bed & Breakfast
HB- only two meals including breakfast meal
FB- breakfast, lunch & dinner
SC- self-catering
Country: Country from which booking is made.
market_segment:
TA: Travel agents

TO: Tour operators
Direct: Direct booking
Corporate: Corporate booking
GDS: Global Distribution System
is_repeated_guest: is a guest repeated or not.
previous_cancellations: cancellation in past
previous_bookings_not_canceled: not cancelled in the past.
reserved_room_type: which type of room is reserved.
assigned_room_type: which type of room is assigned against reservation
booking_changes: no. of changes made in booking
deposit_type:
No Deposit: Bookings made without deposit amount
Non Refund: Deposit will be made and no money is refunded against cancellation
Refundable: Deposit will be made and money will be refunded against cancellation completely or after collecting cancellation fees
agent: Agent number
company: Company number
days_in_waiting_list: number of days taken for confirmation of hotel booking.
customer_type: Transient: People who make individual bookings
Group: People booking at contracted rates for large groups or events
Transient party: Transient booking related to other transient booking
Contract: Booking where a contract is associated with it.
adr: Average Daily Rate
required_car_parking_spaces: is car parking space required/
total_of_special_requests: no. of special requests made.
reservation_status: status of reservation (canceled, check-out or No-Show)
reservation_status_date: Date at which reservation status is updated

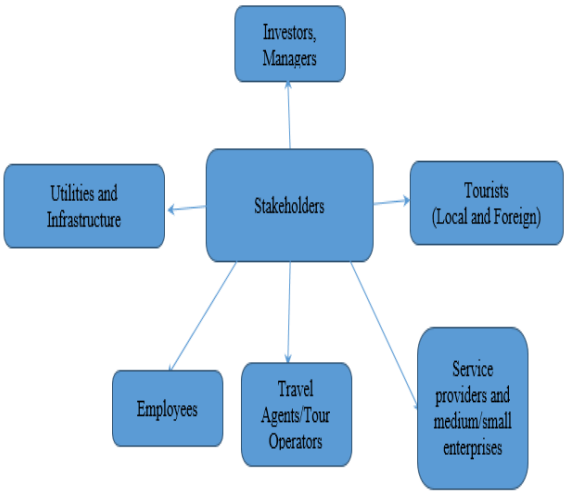
The provided dataset for Hotel Booking has 119390 rows and 32 columns. Each having a data type of object, integer, & float.

2. Introduction:

The hotel industry is a volatile market which changes seasonally and depends on various factors. To understand the business context of the hotel industry we need to understand the different stakeholders in the business.

Enterprise overview:

To understand the purpose of the project
it is necessary to know about the business.
We need to understand data obtained
and then convert it into an analytical problem definition



3. Exploratory Data Analysis (EDA):

Exploratory data analysis is a method by which we can understand the data set, we can create visual information from the data. We can understand the statistical part of the data like mean, mode, etc. With EDA we can find nulls, missing values, duplicate values in the dataset, and outliers. We can find correlations between features in the dataset. In EDA, we can perform different data visualization methods on the data and analyse it. In our dataset, we used 3 steps: data inspection, data cleaning, and data visualization.

Data inspection:

To perform data analysis, we first need to import the Hotel booking csv file into the Pandas data frame. We used 4 libraries in Pandas dataset for data management, NumPy for numerical computation, matplotlib and seaborn for data visualization.

In data inspection we checked shape of data set which is 119390*32, and all the data types of hotel booking dataset holds. In hotel booking dataset object, float, int, data types are present. Checking the information dataset gives a summary of data including data types, null values count and memory usage.

Data cleaning:

Since raw data sometimes includes nulls, missing values, duplicate values, and outliers in the data set, data cleaning is a very important process in EDA.

The data after loading has a shape of (119390, 32), which means the data has 119390 rows and 32 columns. Among the 119390 rows, 31994 rows are found to be duplicated by using the method duplicate () on the data frame. The duplicate rows are removed from the data frame using the method drop_duplicates () on the data frame. The new data frame formed after dropping the duplicates has a shape of (87396, 32).
The null values present in the new data frame are found using the method isna(). The null values of different columns are shown below.

Column Name	Null Values
company	82137
agent	12193
country	452
children	4

The Columns Company, agent, and children are of numerical data of type int and float. So, the data is filled using the number "zero" using fillna (0). Considering the value of '0' for company and agent represents "others" and for children the count of null values are negligible and the children are assumed to be zero.
The column country having null values are filled with others using the fillna ()

Data visualization:

Data visualization in graphical form is done using matplotlib and the seaborn library. Graphical data helps to visualize data clearly and correlate each attribute with each other.

Understanding the most Preferred Hotel:

There are two types of hotels which are analysed they are Resort Hotel and City Hotel.

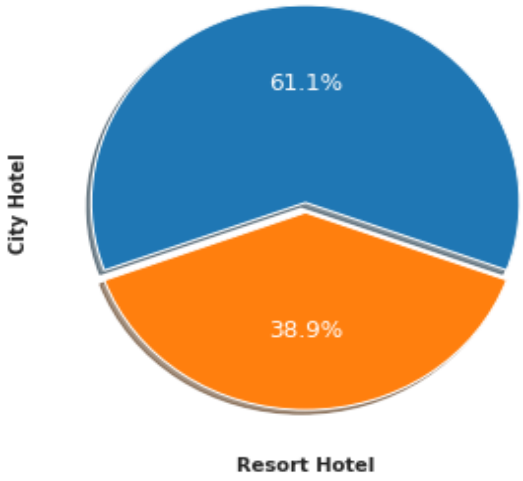
A City hotel is a commercial establishment where bonafide travellers rent the rooms for temporary, overnight lodging with guest facilities.

A Resort hotel is a multi-amenity commercial establishment that provide vacationer or a tourist to obtain services like lodging, food, entertainment and shopping

From the above plot we can observe that City hotel has the largest number of bookings

Hotel Booking Percentage using bar chart and Pie Chart:

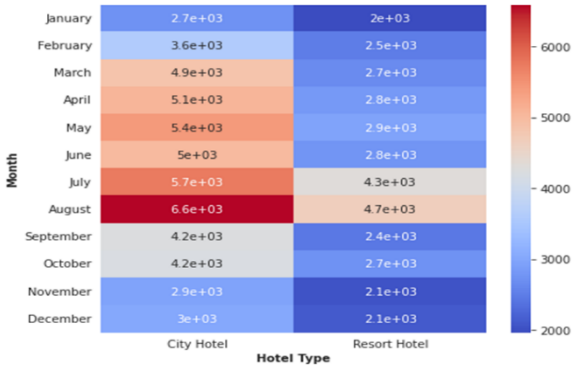
A detailed analysis of which is the most preferred hotel is studied. A bar graph is plotted to visualize the number of bookings of each hotel and a pie chart is plotted to visualize the percentage share of bookings of each type of hotels.



City Hotel is preferred by most of the customers and it contributes to 61.1% of the total booking made.

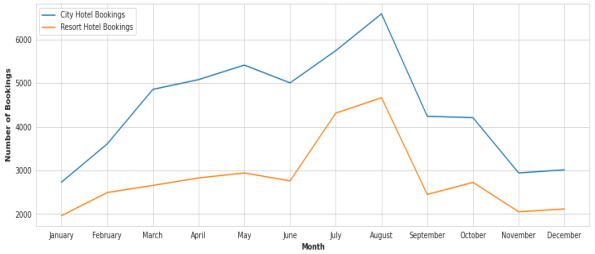
Monthly Booking Analysis of Hotels using Heat Map and line graph:

The trend of bookings in the city hotel and resort hotel is performed. We use a Heat Map to understand the trend to the hotels. A heat map is a visualization technique that plots the magnitude of colour based on the intensity of the numerical value.



From the heat map formed between the number of bookings and months. We can draw the following conclusions:

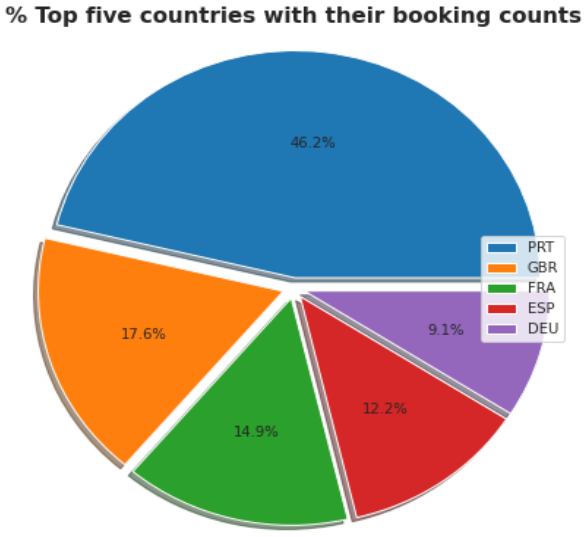
- The August month has the most number of bookings.
- City Hotel has maximum bookings during the period March to August.
- Resort hotels are facing low bookings throughout the year except in August.

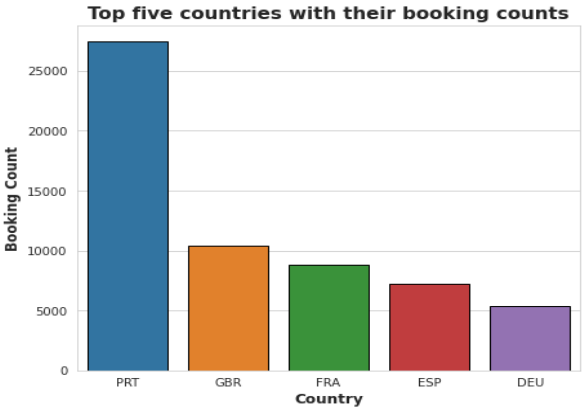


Bookings surged from February to August, although there was dip in June, and began to decline after September

Top 5 Countries with Booking Count using pie and bar chart:

To understand the most important countries to focus on to increase the business a detailed analysis of top 5 countries making booking is studied.

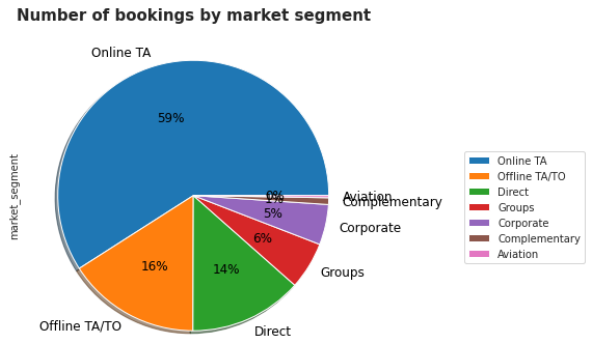




From the above two plots we can say that Portugal(PRT) has the maximum bookings

followed by Great Britain(GBR), France(FRA), Spain(ESP) and Germany(DEU).

Understanding Market Segment using a pie chart:



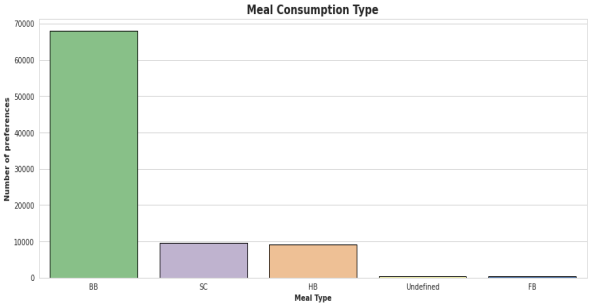
From the above pie chart we can say that most bookings are made by online Travel Agency, followed by Offline TA/TO. Direct Booking is also lower but almost equal to offline TA/TO.

Weekdays Vs Weekend Analysis:

As we can see from the bar chart, guests prefer to stay weeknights the most at both hotels than weekend nights.
Guests prefer to stay in City Hotel on weekends the most.

Meal Type Analysis:

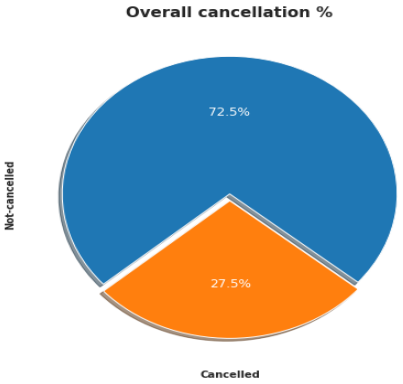
Every year many travellers compare various board packages when booking their trip. Hotels offer various board packages like Bed and Breakfast (BB), Half Board (HB), Full Board (FB) and Self-catering (SC) as options to serve the guests. Food preparation is one of the most important duties of the hotel as it adds value to the guest's experience. Often guests with good service in food tend to revisit the hotel. So, we have made a detailed analysis of meal type by creating a bar graph with number of preferences and meal types.



We have observed that the maximum number of guests prefer Bread and Breakfast.

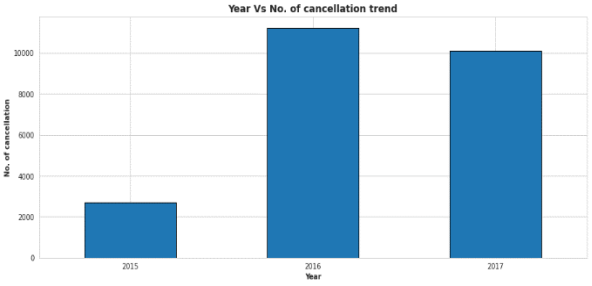
Cancellation Analysis:

Overall Analysis:
Cancellation of bookings is one of the major issues faced by the hotels. So a detailed analysis of cancellations is made. We have found the percentages of cancellations with the help of a pie chart.



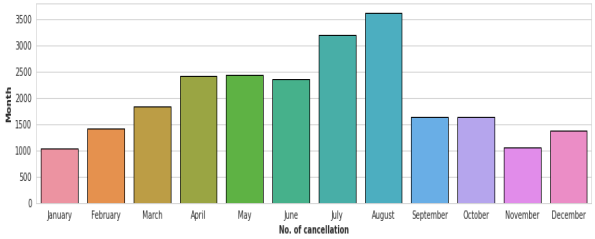
We can find that the total cancellation percentage is 27.5 percent.

Year wise Analysis:
Year wise analysis of cancellation is done to find which year has most cancellations



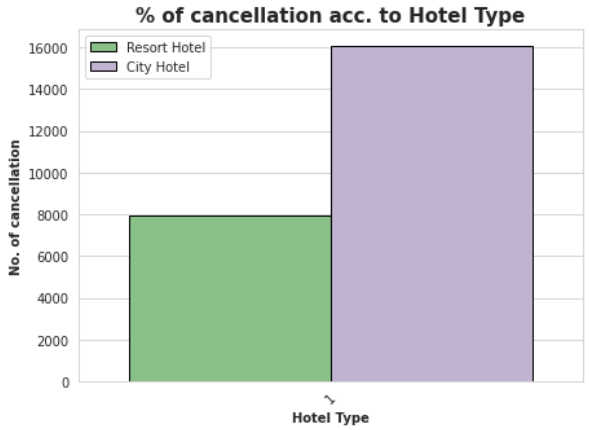
Year 2017 have only data up to June month and we can see from the bar chart that cancellation of booking is almost equal to year 2016 so it can be concluded that 2017 has the highest cancellation

Month Wise Analysis:



A bar graph is plotted between number of cancellations and months and it is found that month August has the highest cancellations and January, November has the least number of cancellations

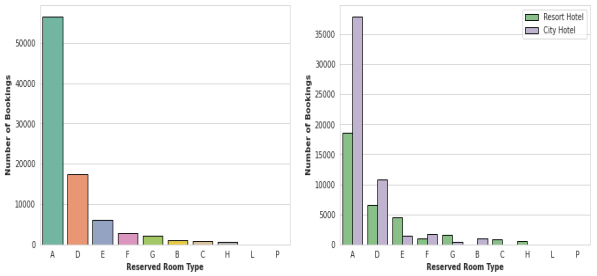
Hotel Wise Analysis:



Hotel Wise Analysis of cancellation is made and it is found that City hotel is facing high cancellations.

Room Type analysis

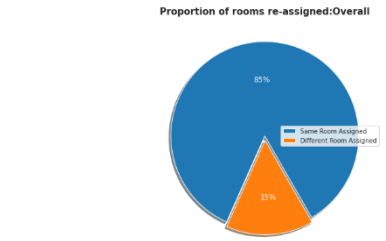
We have performed an analysis of the most prepared room type room wise. A hotel wise analysis of the most preferred hotel is done and bar graphs are plotted

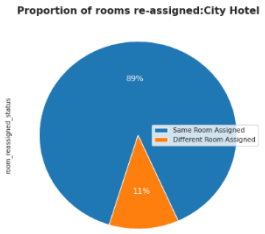


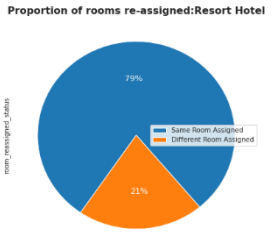
From the above two graphs we can say that Room A is the most preferred hotel followed by room type D.
City hotels have the highest bookings in Type A and Type D rooms.

Room re-assigned analysis

Hotel Wise Analysis:

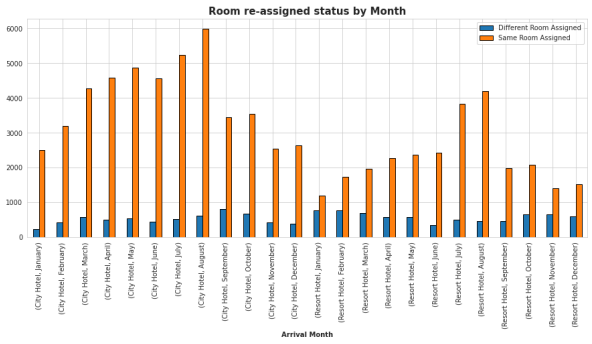






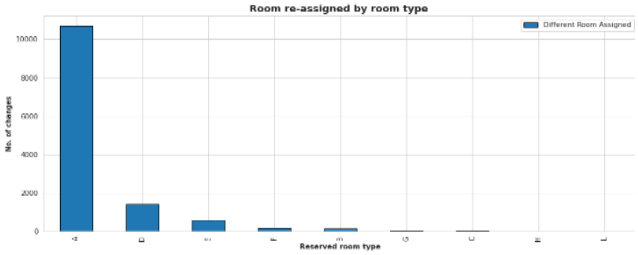
- Over all ~15% room change was observed throughout the year
- City Hotel face 11% room change and Resort Hotel face 21% room change throughout the 3 years

Month wise analysis



Highest number of changes in room were made in the month of the September for City Hotel and January for Resort Hotel

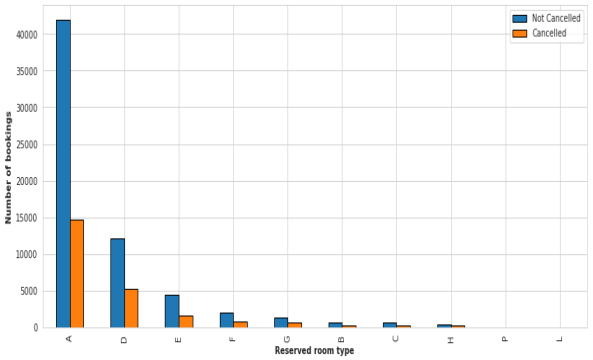
Room type analysis:



Type A rooms are most susceptible to room shifting, followed by D, E, F, B, and C, and insignificant in H and L.

Room type cancellation analysis:

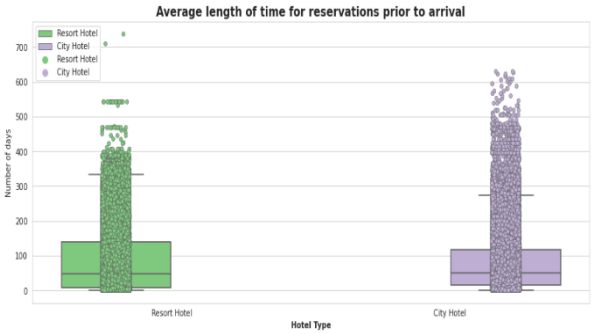
The hotel industry is a challenging landscape as the guest's expectations are greater than ever before. To provide a great customer experience we should follow a customer-centric strategy.
Often we experience large waiting times in confirmation of bookings which leads to cancellation of rooms. So we have made an analysis of the most preferred rooms by finding the most reserved room.
Sometimes due to high cancellation rates alternative room types are assigned to the customer. These rooms are prone to cancellation so an analysis is made on assigned room type and cancellation. A bar chart is drawn between the assigned room type and booking is cancelled or not.



Type A rooms are most susceptible to room shifting, followed by D, E, F, B, and C, and insignificant in H and L.

Room types of A and D are most preferred by the customers are having low percent of cancellation rates

Hotel wise analysis of Lead Time using box and scatter plot:



Hotel Type	count	mean	min	25%	50%	75%	max
City Hotel	53424.0	77.684112	0.0	14.0	50.0	118.0	629.0
Resort Hotel	33968.0	83.371938	0.0	8.0	47.0	138.0	737.0

It is clearly visible from the plot that most of the guests prefers to book the hotel 2-3 months before arrival

Country wise analysis of Lead Time and Waiting Time using a heat map with background gradient:

To tackle the problems of high cancellations due to high waiting time a deep country wise study of lead time and waiting time is performed for the countries which have bookings greater than 1000. It is important for a hotel to understand the lead time because the hotel needs to get prepared to satisfy the guests at the time of arrival. Longer lead times gives the hotel management time to prepare themselves. Longer waiting times on the

other hand increases the chances of cancellation. So, we have created a heat map to identify highest lead times and waiting times.

	lead_time								waiting_time							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
country																
BEL	2081.000000	94.290245	80.007766	0.000000	28.000000	75.000000	149.000000	396.000000	2081.000000	0.119173	4.283490	0.000000	0.000000	0.000000	0.000000	185.000000
BRA	1995.000000	80.864662	79.065161	0.000000	19.000000	55.000000	127.000000	354.000000	1995.000000	0.204010	4.771928	0.000000	0.000000	0.000000	0.000000	167.000000
CHE	1570.000000	87.754777	76.339889	0.000000	24.000000	68.000000	135.000000	457.000000	1570.000000	0.087261	2.661369	0.000000	0.000000	0.000000	0.000000	86.000000
CN	1093.000000	106.650503	88.937360	0.000000	33.000000	89.000000	166.000000	465.000000	1093.000000	0.143641	3.601288	0.000000	0.000000	0.000000	0.000000	199.000000
DEU	5387.000000	105.089103	89.521446	0.000000	31.000000	83.000000	165.000000	457.000000	5387.000000	1.051977	12.672491	0.000000	0.000000	0.000000	0.000000	224.000000
ESP	7252.000000	52.196773	61.621102	0.000000	8.000000	30.000000	73.000000	367.000000	7252.000000	0.146890	3.556934	0.000000	0.000000	0.000000	0.000000	287.000000
FRA	8837.000000	74.135866	71.760677	0.000000	17.000000	51.000000	111.000000	478.000000	8837.000000	0.846667	12.670392	0.000000	0.000000	0.000000	0.000000	379.000000
GBR	10433.000000	117.419055	100.141759	0.000000	33.000000	93.000000	180.000000	709.000000	10433.000000	0.377264	6.535174	0.000000	0.000000	0.000000	0.000000	150.000000
IRL	3016.000000	114.276857	87.299131	0.000000	41.000000	98.000000	168.000000	485.000000	3016.000000	0.043103	1.573874	0.000000	0.000000	0.000000	0.000000	61.000000
ITA	3066.000000	83.231246	75.470956	0.000000	19.000000	64.000000	128.000000	348.000000	3066.000000	0.961840	9.256337	0.000000	0.000000	0.000000	0.000000	174.000000
NLD	1911.000000	79.920984	75.768942	0.000000	16.000000	56.000000	128.500000	365.000000	1911.000000	0.358974	6.701947	0.000000	0.000000	0.000000	0.000000	185.000000
PRT	27449.000000	85.110277	87.644767	0.000000	3.000000	25.000000	98.000000	737.000000	27449.000000	1.279864	12.941051	0.000000	0.000000	0.000000	0.000000	391.000000
USA	1875.000000	68.748800	74.041411	0.000000	12.000000	44.000000	103.000000	542.000000	1875.000000	0.388800	5.686501	0.000000	0.000000	0.000000	0.000000	147.000000

<figure size 7200x7200 with 0 Axes>

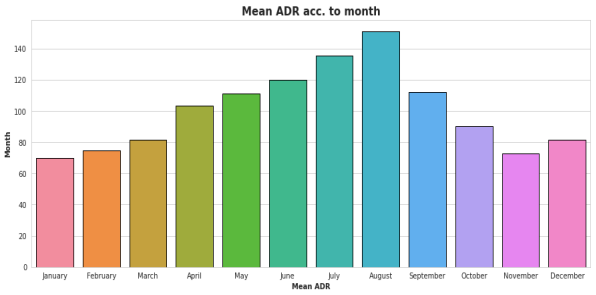
From the above heat map we can come to the following conclusions:

1. Portugal (PRT) has the maximum number of bookings and has low mean lead time and highest mean waiting time. This indicates that the chances of bookings not confirming is high. As Portugal has the largest share of bookings there is a need to introduce an advertising strategy or give discounts well in advance before the peak time of bookings.
2. Germany (DEU) has high lead time and high waiting time. This means there is a large demand for rooms and the hotels are under-priced. Increasing the room rent will maximize the hotel revenue.
3. Great Britain (GBR), Ireland (IRL), Switzerland (CHE) and Belgium (BEL) have high mean lead time and medium amount of waiting time. This is the ideal requirement for getting high profits so the people from these countries should be advertised to increase the number of bookings.

Revenue Analysis Using ADR:

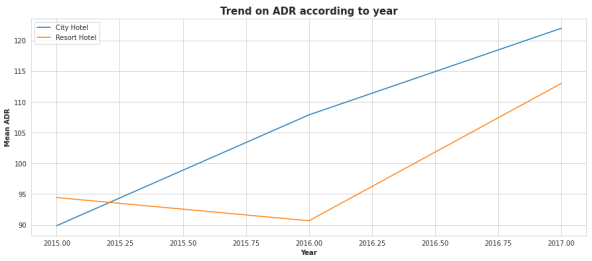
ADR is defined as the ratio of Total revenue to the total rooms occupied. It is one of the Key Performance Indicators for analysing the revenue of the hotel business. ADR is a useful tool in fixation of hotel room prices to maximize revenue.
We have made a study of ADR to find the monthly trend for July 2015 to August 2017. We have plotted a bar graph with ADR with month.

Month Wise Analysis



The ADR follows an increasing trend from January to August and decreasing trend from August to December.
Months from May to October have high ADR values. This indicates that maximum income is generated during these months. Increasing the prices for these months may be considered to increase the revenue.

Month Wise Analysis



From the line plot we can see that City Hotel had linear increase in ADR throughout the year and but for Resort Hotel ADR was falling till pre 2016 and post 2016 it started to see a raise.

Conclusion:

The majority of the hotels reserved are city hotels. City hotel receives approximately 60% of bookings, while Resort hotel receives 40% of bookings; thus, City hotel is busier than Resort hotel. **City hotels will undoubtedly require the most targeted funding**.

Bookings increased from March to August, with August having the highest number of bookings. The busiest month for both hotels is August, followed by July and May. The months of November, December, and January have the lowest bookings. **Both hotels should**

be prepared for peak season bookings from March to August, and hotel room preparation (room preparation, maintenance, staffing, etc.) should begin in November, December and January to handle the rush of bookings.

In both hotels, guests preferred type A rooms, followed by types D, E, F, G rooms. A 15% change in the room was observed overall. Room changes were most common in type A rooms, followed by D,E,F,B,C, and negligible in H, L.**So, it is clear that Type A are the most preferred followed by Type D rooms and most susceptible to room changes. Therefore, both hotels should focus on increasing the number of Type A rooms by replacing other types of rooms that are being booked by a smaller number of guests. .This will result in an increase in ROI for both hotels.**

Guests prefer to stay the weeknights the most at both hotels than weekend nights.

The majority of the guests were from Europe. Guests from Portugal was the highest, followed by the United Kingdom (GBR), France (FRA), Spain (ESP), and Germany (DEU).**This may be due to the ease in visa process for this region**.

Bed and breakfast (BB) is the most ordered meal, followed by SC (no meal), HB (half board), undefined, and FB (full board**). This may be because guests prefer to have breakfast in the morning before leaving the hotel for sightseeing and eat outside the hotel while sightseeing. Hotels should mostly concentrate on breakfast quality and quantity for ROI.**

A total of 27.5% of cancellations have been observed. The most bookings were cancelled in 2017, with the fewest in 2015. August had the highest number of cancellations, while January had the lowest. The City Hotel receives more cancellations than the Resort Hotel. **To minimize cancellations, hotels should track the price of their stay on other channels, including OTAs (online travel agencies) with which the hotel is not a partner. This is to ensure that their customers are not in a hurry to rebook their hotel rooms on other channels at lower prices**.

The majority of bookings are made through online travel agencies, with offline TA/TO coming in second. Direct booking is also less, but it is nearly equal to offline TA/TO. **This must be because guests want to avoid all preparations/procedures prior to travel. To do this, they hire a travel agent to make all necessary travel arrangements and hotel reservations at the best price. Hotel management should promote online TA and give special offers and packages for increasing their hotel bookings.**

The ADR follows an increasing trend from January to August and decreasing trend from September to December. **ADR is a useful tool for pricing hotel rooms to maximize revenue... ADR values are high from March to August. This means that your maximum income will accrue during this month. An increase in price during this month can be considered in strategy for increasing the income.**

High Lead time & Low Waiting time is ideal condition for the hotel

Lead Time	Waiting Time	Demand	Price
High	High	High	Low
Low	Low	Low	High
Low	High	Room getting cancelled due to non-confirmed booking	
High	**Low**	**Ideal**	**Ideal**

- **High Lead time & Low Waiting time is ideal condition for the hotel**

Maximum number of customers are from the transient category which is near about 75.1 %.**Transients are usually walk-in or direct booking guests, or groups of guests who only stay 8-10 days, so their visit has a specific**

purpose. This is the largest market segment, so hotels should offer discounts, special services for this customer and can spend more money on advertising to get this customer segment.

Reference:

1. GitHub
2. Kaggle
3. Medium