**Hotel Booking Analysis**

**Manish Kumar**

**Data Science Trainee**

**Alma Better**

**Abstract:**

Data analytics in the hotel industry is key to marketing strategy, building customer loyalty, and enhancing productivity. It enables hotels to personalize experiences for their guests, introduce better hotel pricing strategies, and expand their customer base. Here are some ways in which data analytics makes a positive impact on the hotel industry.

* Customer Data Analysis & Market Segmentation.
* Real-Time Data and Hotel Pricing Strategies.
* Managing Hotel Booking Channel
* Inventory Management
* Demand Forecasting.

**Library used: Numpy, Pandas, DataFrame, Data Visualizations**

**1.Problem Statement**

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order 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 hotel booking dataset can help you explore those questions!

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.

**Explore and analyze the data to discover important factors that govern the bookings.**

**2.Introduction**

Data analytics in the hotel industry is key of marketing strategy, building customer loyalty, and enhancing productivity. It enables hotels to personalize experiences for their guests, introduce better hotel pricing strategies, and expand their customer base. Here are some ways in which data analytics makes a positive impact on the hotel industry.

* Customer Data Analysis & Market Segmentation
* Real-Time Data and Hotel Pricing Strategies
* Managing Hotel Booking Channels
* Inventory Management
* Demand Forecasting

In our project we have to do analysis of different parameters which are as follows:-

**Hotels:** There are only two types of hotels Resort hotel and City hotel.

**Market Segments:** We have eight unique market segments from where customers are coming i.e. ‘Direct’, ‘Corporate’, ‘Online TA’, ‘Offline TA/TO’, ’Complementary’, ‘Groups’, ‘Undefined’, ’Aviation’.

**Meal:** These are the four types of meal in the given data:

* SC: self-catering (no meals are included).
* BB: bed and breakfast.
* HB: half board.
* FB: full board.

**Country:** We have data of 177 countries coded in short form.

**Lead Time:** At a hotel, the time taken between when a customer makes a reservation and their actual arrival is called the Lead Time.

**Cancellations:** It contains only two values 1 or 0.

* 1: booking has cancelled
* 0: booking is currently active.

**Stays in weekend nights:** We are getting two unique values for column ‘stays in weekend nights’ i.e. 1 and 2 , that means someone booked for one weekend night (Saturday or Sunday) and some booked for two weekend nights (Saturday and Sunday).

**Stays in week nights:** We are getting five unique values for column ‘stays in weekend nights’ i.e. 1 to 5, that means someone booked for at least one week night to maximum five week nights.

**Adr:** Average daily rate for individual order.

**Car Parking spaces:** Required car parking spaces, this column contains number of car parking spaces required as per customer demand.

**3.Hotel Booking Data Analysis**

**Handling NaN values:**

Data in real world are rarely clean and homogeneous. Data can either be missing during data extraction or collection.

Missing values need to be handled because they reduce the quality for any of our performance metric. It can also lead to wrong prediction or classification and can also cause a high bias for any given model being used.

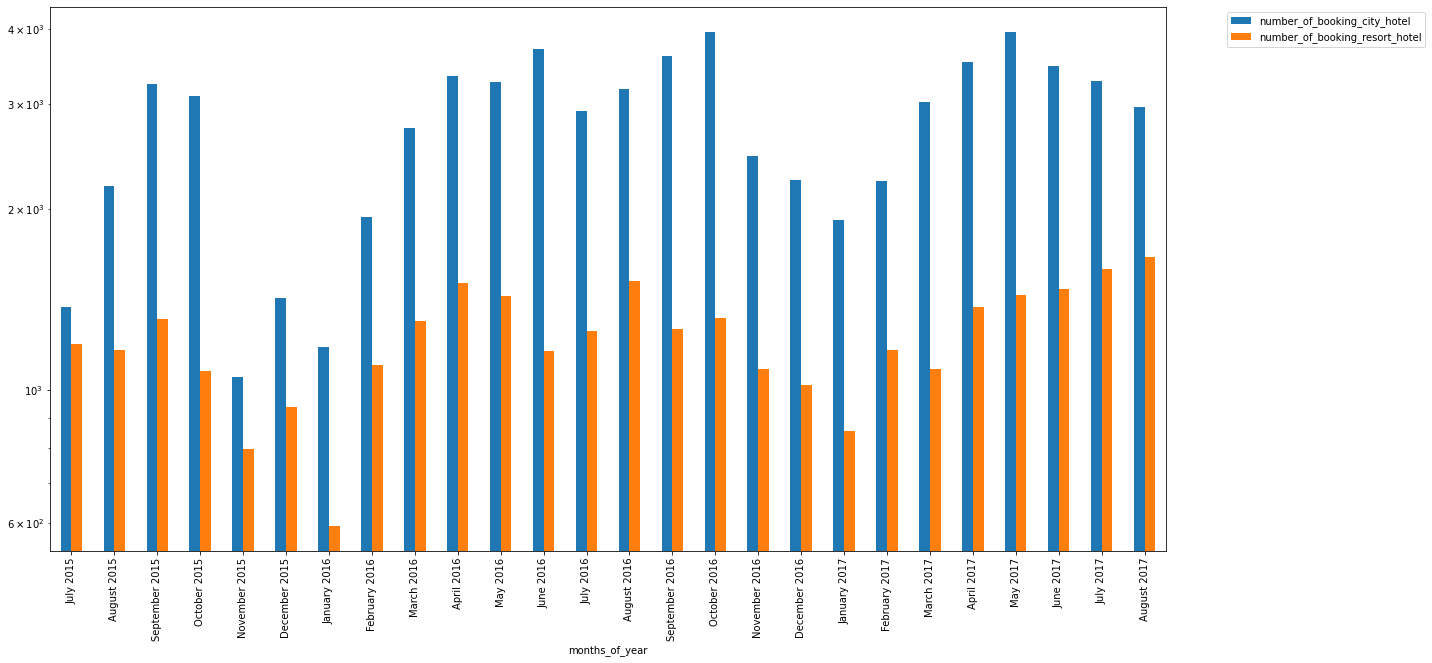
There are several options for handling missing values each with its own PROS and CONS. However, the choice of what should be done is largely dependent on the nature of our data and the missing values. Below is a summary highlight of several options we have for handling missing values.

* Drop missing values.
* Fill missing values with test statistic.
* Predict missing value with a machine learning algorithm.
* Here we can observe that four columns have some null values, let’s calculate the total percentage of null values:
* Total no. of rows/data = 119390.
* In column **‘children’** we have four (119390-119386=4) null values.
* In column **‘country’** we have 488(119390-118902 = 488) null values.
* In column **‘agent’** we have 16340 (119390-103050 =16340) null values.
* In column **‘company’** we have 112593 (119390-6767 = 112593) null values.
* Here we can observe that in ‘company’ column 94.3% values are null, hence we can not perform any analysis on ‘country’ data. Hence it will better if we drop this column.
* Yes! Now we can observe above that every column has equal number of non-null values that means all NaN values are now eliminated.
* Now, We have three columns ‘arrival\_date\_year’, ‘arrival\_date\_month’, and ‘arrival\_date\_day\_of\_month’ which gives the information of check-in date.

**4.BOOKING ANALYSIS**

**Monthly booking analysis:**

Under this block, we are going to trace the number of bookings for every month each year for both hotels.



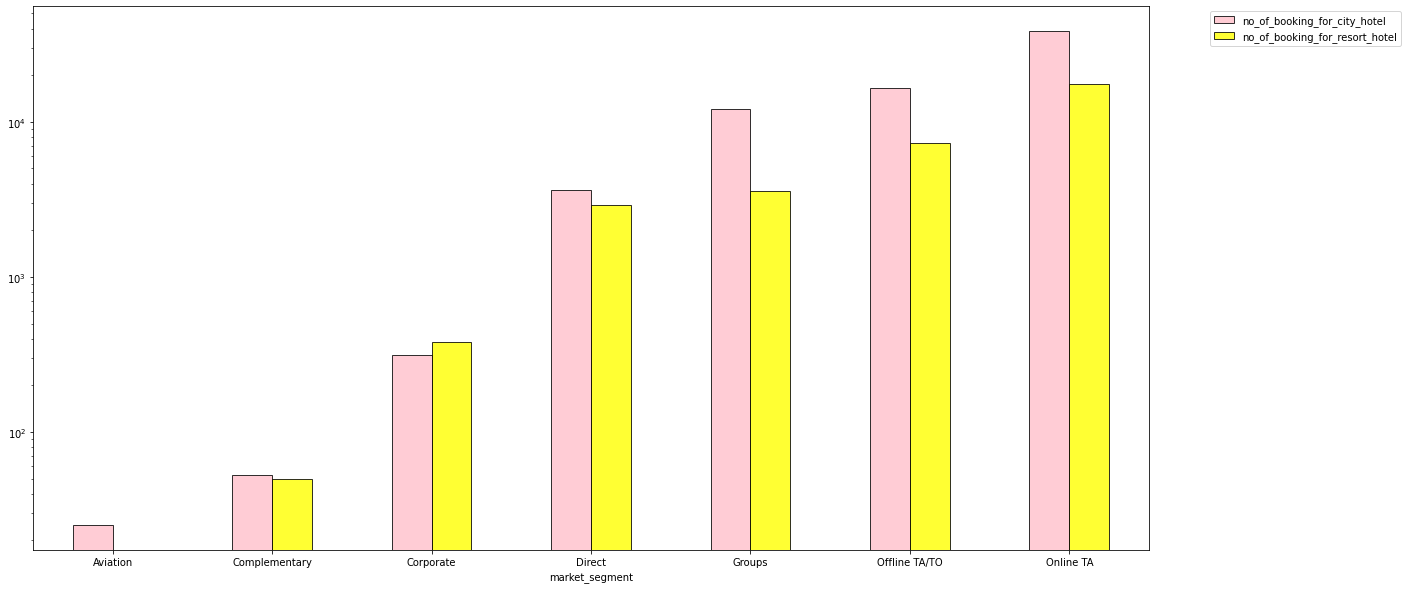
After tracing the data, we observe that for both the hotels, booking start increasing from spring to summer season and then decrease till winter season.

However, comparing the magnitude of booking we see that city hotel has always scored higher inspite of the trend of booking being same.

**Analysis of number of bookings through each market segment:**

We have seven distinct market segment in our data. The term ‘TA’ means ‘Travel Agents’ and ‘TO’ means ‘Tour Operators’.

Under this segment we are going to analysis traffic from each market segment for both the hotels.



We found that TA are the major contributors of booking whether it is offline or online. However online TA booking are more than offline.

Aah… Internet is the lifesaver.

In general, we saw that city hotels have more bookings than resort hotels, but that is not the case in Corporate market segment.

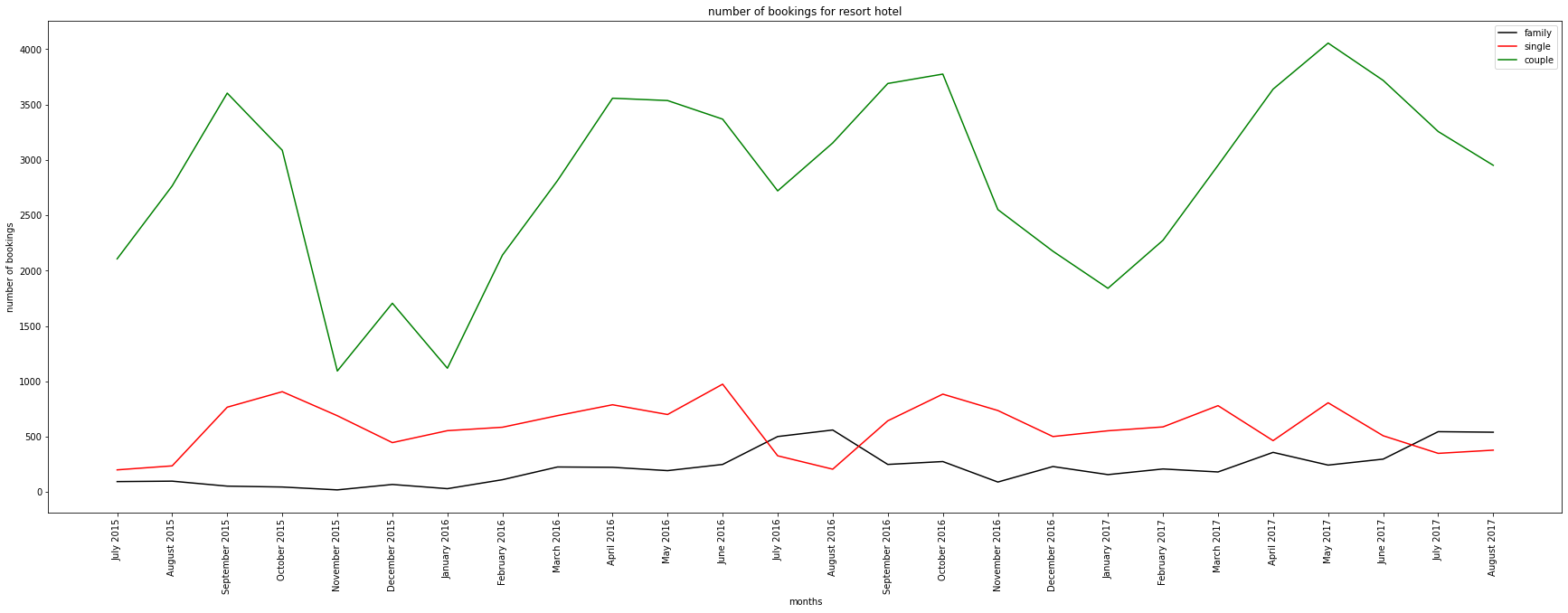
From aviation segment, there is no booking for resort hotel which is obvious.

**Average number of booking of couples, families and single person of couples hotel wise:**

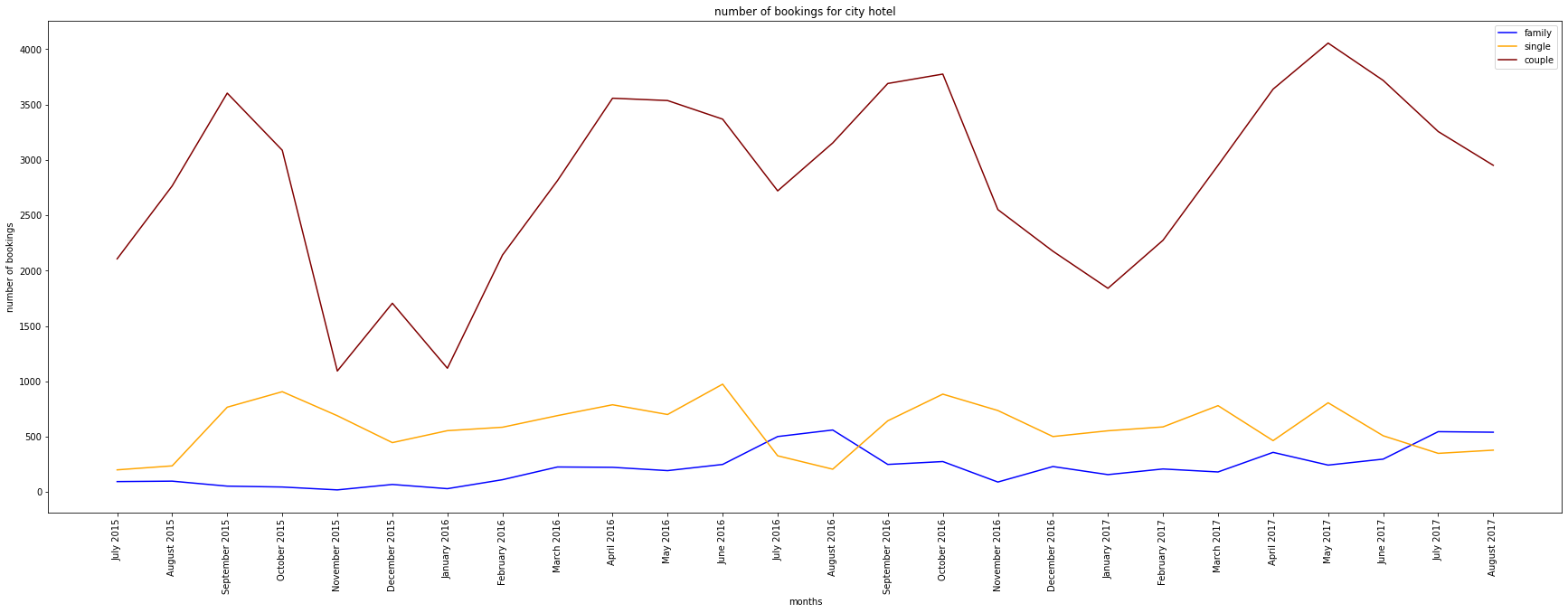
We have seen that customers can be broadly categorized as:-

* Single
* Couple
* Family

Average number of booking of couples, families and single for resort hotel:



Average number of booking of couples, families and single for city hotel:



It is clearly visible that trend of booking is same for both the hotel. It seems like “couples” are running the hotel business.

So, for any hotel, couples should be the prime target.

Talking about singles and family, there is no clear trend of booking. Their booking level is more or less the same throughout the year.

**Lead Time Analysis:**

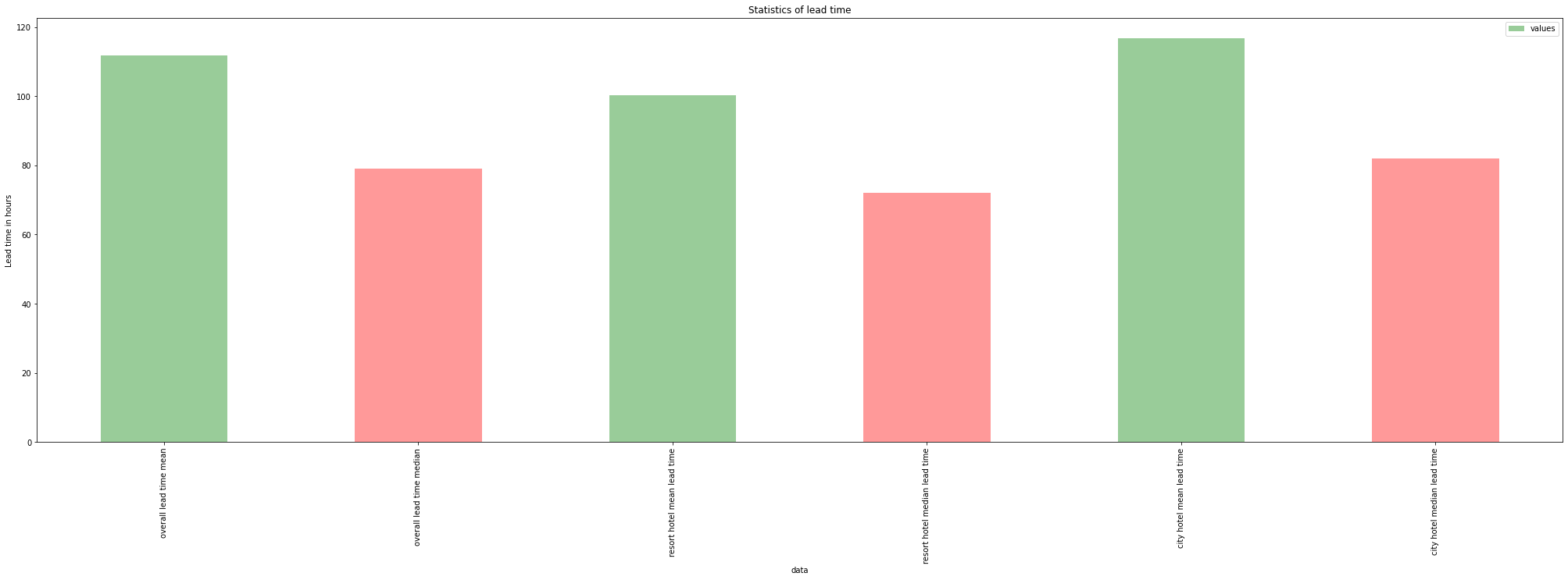
At a hotel, the time taken between when a customer makes a reservation and their actual arrival is called the **Lead Time.** Sometimes this time period has to be restricted.

Integrating a Lead Time Restriction into a hotel’s booking system is something all hotel Reservation Teams and/or Revenue Management departments should consider.

**Why?**

Because it can help with the management of room availability overall, ensuring that profitability is always maximized. This usually applies to an offer, which is available up to a certain number of days prior to arrival. Can be combined with other restrictions. For example a 15% discount, which is non-refundable, up till 30 days prior to arrival. Such promotion are set-up not to coincide with the regular booking window, in order not to down trade ADR on your normal demand.

When we have a normally distributed sample, we can legitimately use both the mean or the median as our measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal.



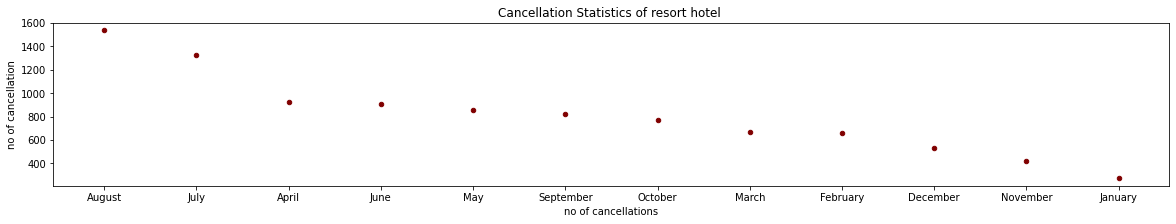
However, when our data is skewed, for example, as with the right-skewed data, We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data.

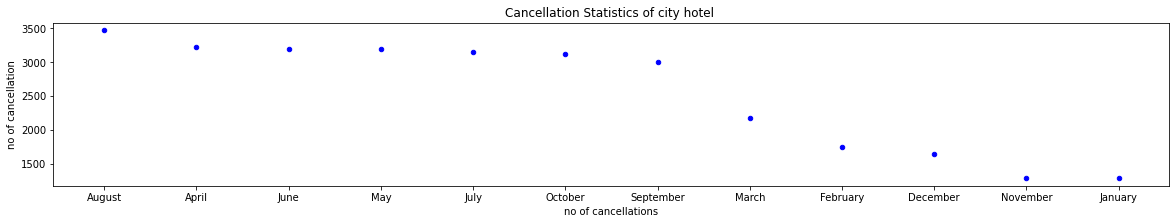
The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean.

**Cancelled booking analysis:**

Booking cancellations are undoubtly one of the biggest headache for hotel managers.

This can arise many problems such as lowering the price of rooms so it can resell easily or loss of revenue when they cannot resell the room.





What we can do to reduce it let’s see include a cancellation policy in your hotel reservation policy and make it visible to your guests. Make sure that your customers are held responsible in case of a no-show. This is another effective way to attract guests and reduce cancellations at your hotel. You can confirm the booking with your guest and then offer them a discount once done. You can throw in a clause here stating that if the guest pays upfront, they get a fatter discount.

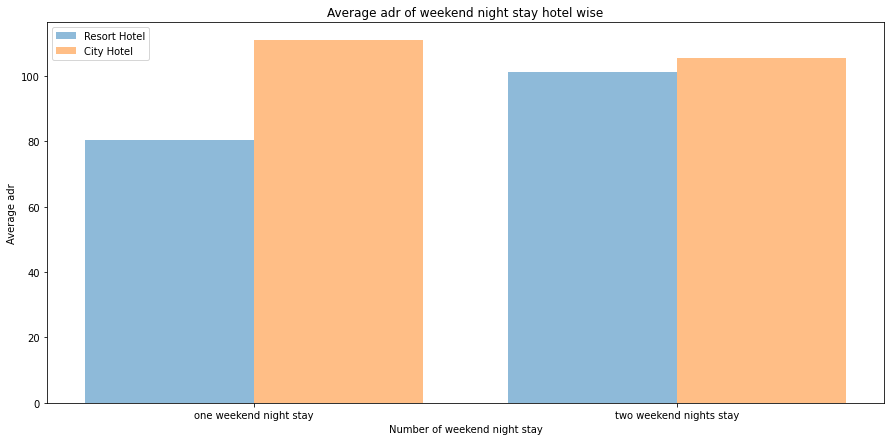
**Analysis of ‘adr’ (average daily rate) v/s weekend night stay, week night stay and composite night stay:**

We know that in general hotel demands are higher in weekend than week days, hence this effects the booking price and also effects the revenue for hotels, So it is very necessary to analyse the same.

In this part of the project we want to observe how ‘adr’ got effected by selecting weekend nights, week nights or both.

I divided the this analysis in four parts:

1. How ‘adr’ is changing with increase in number of weekend nights for those orders which has only bookings of weekend nights.
2. How ‘adr’ is changing with increase in number of week nights for those orders which has only bookings of week nights.
3. How ‘adr’ is changing with increase in length of stay.
4. How ‘adr’ is varying with time (month wise)



**Meal preference analysis:**

Analyzing customer meal preference data allows us to understand the customer preferences and purposes. Moreover, analyzing the customer data can help us to forecast the demand and pattern of customer behaviour with higher accuracy.

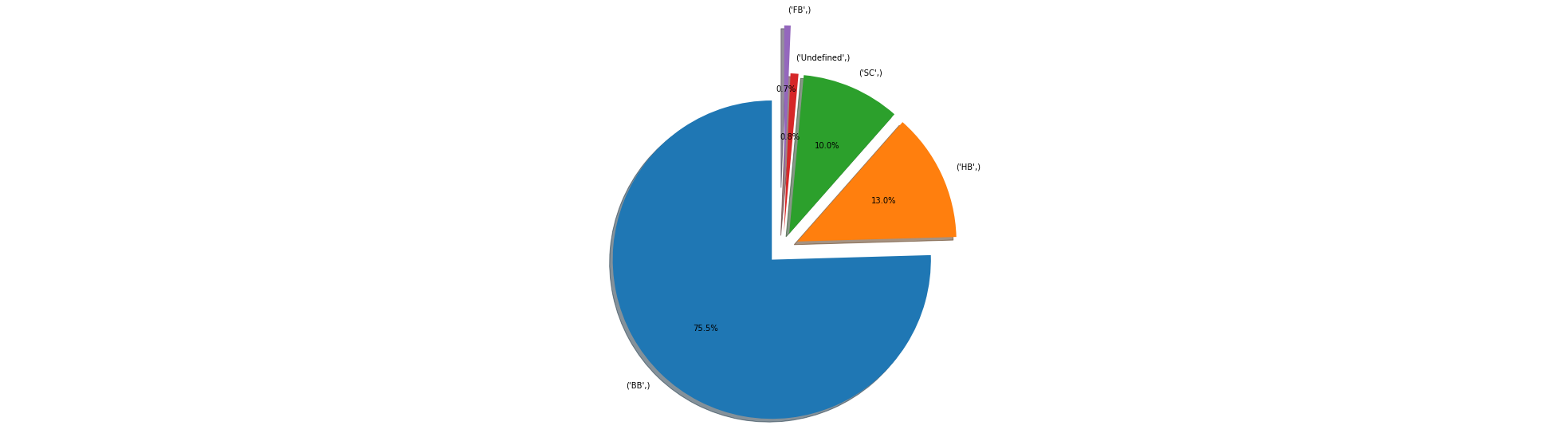
As we already know that:-

SC means self-catering (no meals are included)

BB means bed and breakfast, in which breakfast is served in room itself.

HB means half board, in which breakfast and dinner are included.

FB means full board, in which breakfast, lunch and dinner are included.



**Observation and Suggestions:**

It is observed that most guests preferred BB (bed & breakfast).

BB is best for those who aren’t morning people and don’t want to worry about preparing breakfast for themselves or having to choose somewhere to have breakfast served. That’s why customers prefer BB, In some hotels BB is included in room rate.

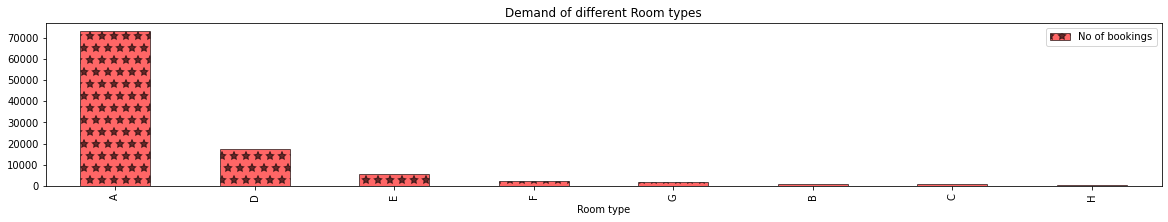
FB is the least preferred meal type. This meal type suits to those people who don’t want to move much and stay at the hotel during the whole day. In addition to this high price rate for FB meal may be the reason for its least preference.

**Hotel management can work upon “rate of the meal” to lure customers towards FB and HB meal.**

**Analysis of Room type:**

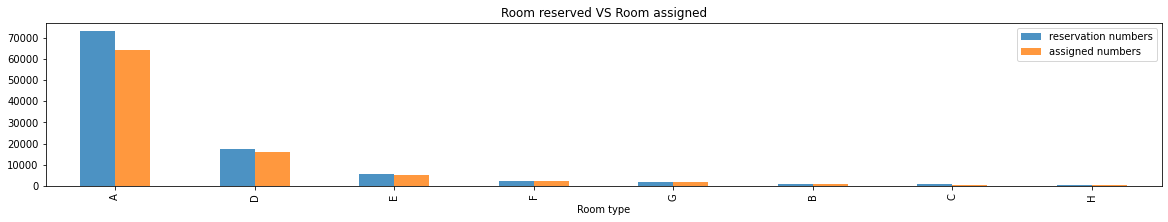
For any hotel management, it is very important to know that “what is the demand of the room and how much demand they are able to fulfill?”

This is a very critical analysis as this is a concern of “Customer Satisfaction”



**Observation:**

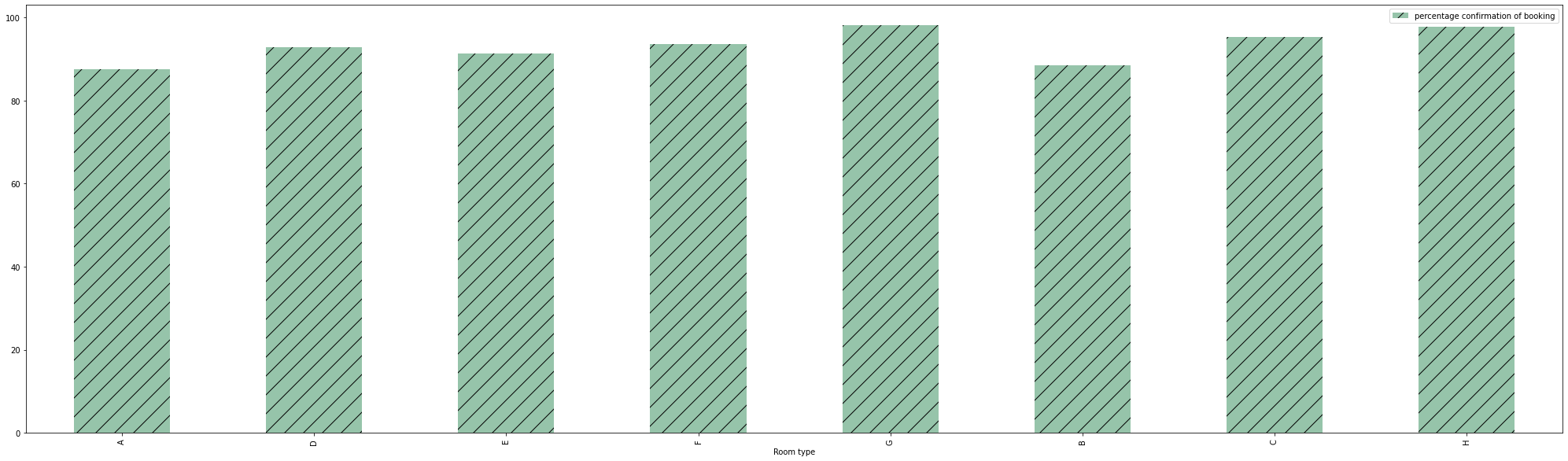
Demand of **type-A** room is highest followed by  **type-D** room.



**Observation:**

However demand of type-A room is highest but the demand is not fulfilled properly as we can see a visible gap in demanded numbers and assigned numbers.

So, Hotel management should look into it seriously.



**Observation:**

From customer perspective, if they book type-G or H type of room, there are high chances they will be assigned room of their choice.

**5.CONCLUSIONS:**

* ‘company’ column 94.3% values are null, hence we cannot perform any analysis on ‘country’ data hence it will be better if we drop this column.
* Comparing the magnitude of booking we see that city hotel has always scored higher inspite of the trend of booking being same.
* We found that TA are the major contributors of booking whether it is offline or online. However, online TA bookings are more than offline.
* Talking about singles and family, there is no clear trend of booking. Their booking level is more or less the same throughout the year.
* Here is a significant difference in mean and median which means distribution of lead time is skewed. Thus, giving more emphasis on median is relevant.
* Number of cancelled bookings is in the month of August.
* BB more preferred over other meal preferences.
* People from 112 countries visit both hotels, whereas people from 53 countries visit only city hotel and people from 9 countries visit only resort hotel.
* PRT-Portugal has the highest contribution both the hotels. However, for the second place, each hotel has different contributor.
* Maximum customers don’t require any parking spaces. However, there are some instances where customers require one parking space.
* Demand of type A room is highest followed by type D room.

**References:**

1. Numpy, Pandas, Matplotlib & Seaborn documentation.
2. Almabetter recorded classes.
3. Articles on towards data science.