**Abstract**

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? And what if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?.

Understanding the business data may be a key aspect in every industry. During this project we are going to perform some exploratory data analysis techniques using descriptive statistics and graphical tools to better understand the data and the best time to book hotel rooms and optimal time to stay in a hotel.

**About Dataset**

We are given a Hotel booking analysis dataset having 32 columns with 119390 records of room bookings over the time from the different places of customers.

Columns description:

1. Hotel : Hotel(Resort Hotel or City Hotel)​
2. Is canceled : Value indicating if the booking was canceled (1) or not (0)​
3. Lead time : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date​
4. Arrival date year : Year of arrival date​
5. Arrival date month : Month of arrival date​
6. Arrival date week number : Week number of year for arrival date​
7. Arrival date day of month : Day of arrival date​
8. Stays in weekend nights : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to remain at the hotel​
9. Stays in week-nights : Number of week-nights (Monday to Friday) the guest stayed or booked to remain at the hotel​
10. Adults : Number of adults​
11. Children : Number of children​
12. Babies : Number of babies​
13. . Meal : sort of meal booked. Categories are presented in standard hospitality meal packages​
14. Country : Country of origin.`​
15. . Market segment : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”​
16. channel : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”​
17. Is repeated guest : Value indicating if the booking name was from a repeated guest (1) or not (0)​
18. Previous cancellations : Number of previous bookings that were canceled by the customer before the current booking​
19. Previous bookings not canceled : Number of previous bookings not canceled by the customer before the current booking.​
20. Reserved room type : Code of room type reserved. Code is presented rather than designation for anonymity reasons.​
21. Assigned room type : Code for the sort of room assigned to the booking.​
22. Booking changes : Number of changes/amendments made to the booking from the instant the booking was entered on the PMS until the moment of check-in or cancellation​
23. Deposit type : Indication on if the customer made a deposit to ensure the booking.​
24. Agent : ID of the agency that made the booking​
25. Company : ID of the company/entity that made the booking or liable for paying the booking.​
26. Days in roll : Number of days the booking was in the waiting list before it was confirmed to the customer​
27. Customer type : sort of booking, assuming one among four categories​
28. Adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the entire number of staying nights​
29. . Required car parking spaces : Number of car parking spaces required by the customer​
30. . Total of special requests : Number of special requests made by the customer (e.g., bed or high floor)​
31. Reservation status : Reservation last status, assuming one among three categories​
32. Canceled – booking was canceled by the customer​
33. Check-Out – customer has checked in but already departed​
34. No-Show – customer didn't check-in and did inform the hotel of the reason why​
35. Reservation status date : Date at which the last status was set. This variable are often used in conjunction with the Reservation Status to understand when the booking was canceled or when the customer checked-out of the hotel.

**Problem Statement**

Hotel booking analysis dataset having 32 columns with 119390 records of room bookings over the time from the different places of customers.

This data set contains booking information for a city hotel and a resort hotel and includes information like when the booking was made, length of stay, the amount of adults, children, and/or babies, and therefore the number of available parking spaces, among other things. All personally identifying information has been faraway from the data.​

In this project we will explore the data set and analyze the data to discover important factors that govern the bookings.

**Introduction**

The hotel industry may be a very volatile industry and the bookings depend on various factors such as type of hotels, seasonality, days of week and lots of more factors. This makes analyzing the patterns available within the past data more important to help the better. Using the given historical data, hotels can perform various campaigns to extend the business.

1. We will try to answer few question mentioned below:
2. Which hotel type is most preferred by customers?​
3. Which type of customers has more bookings?​
4. Which type of rooms are most preferred rooms?​
5. What is the cancellation of bookings with respect to Distribution Channel?​
6. What are the explanations for cancellation of bookings?​
7. Which agent made more bookings?​
8. What is the percentage of repeated guests?​
9. What is the percentage distribution of required car parking spaces?​
10. What is the percentage of booking changes made by the customer?​
11. Which channel is made more bookings?​
12. Which channel is used for early bookings?​
13. Which channel making good revenue generating with respect to hotel?​
14. What is the adr across the market segments?​
15. What is the number of bookings by month wise?​
16. What is the adr across the months?​

**Objectives**

We are provided with a hotel booking analysis dataset.

Our main objective is to perform Exploratory Data Analysis on a given dataset and acquire useful data insights about the general trends in hotel room bookings and what are the factors governing the hotel bookings and how they are governing by interacting with each other factors.

**Used Libraries and Packages**

**Pandas:** Pandas is an open source package that is mostly used for manipulation and data analysis. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays. It represents the data in tabular format structure.

**Matplotlib:** Matplotlib is a plotting library for the python programming language and its numerical mathematics extension NumPy. It provides the techniques to plot the data into different graphical representations to understand the data in an efficient manner.

**Seaborn:** Seaborn library is also a plotting library built on top of Matplotlib library. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Used Techniques**

**Bar Chart:** A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

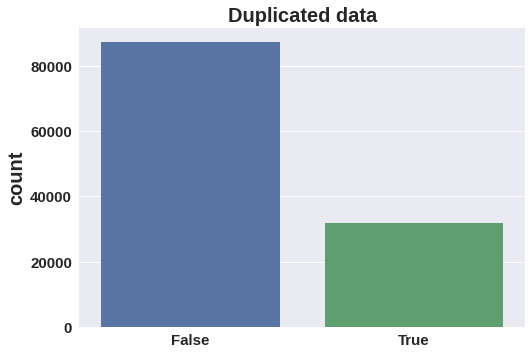
**Line Chart:** A line chart or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields

**Pie Chart:** A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

**Box Plot:** In descriptive statistics, a box plot or box plot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

**Scatter Plot:** A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

**Data cleaning**

Figure-1 

1. **Duplicated Data:** We can see in the above diagram that the dataset contains 31994 duplicated data. So in order to get proper analysis of data I have deleted the duplicated data.
2. **Null values:**
   1. Company – 82137: Fill null values with "0" by assuming company is others category.​
   2. Agent – 12193: Fill null values with "0" by assuming company is others category.​
   3. Country – 452: Fill null values with "Others" by assuming while collecting data user selected the option "Other".​
   4. Children – 4: Filled children column with "0" by assuming no children​
3. **Remove rows:** Found 166 records where children, adults and babies are having value 0. Which means there is no bookings made.​

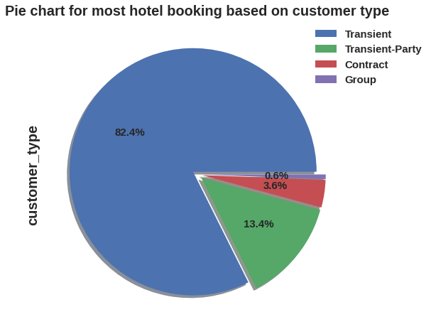
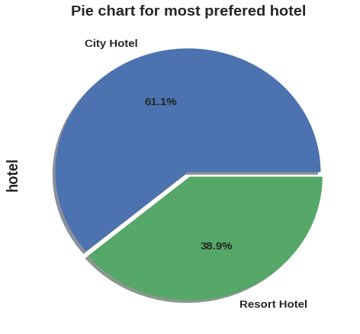
So I have decided to remove those 166 records to avoid inconsistency in data exploration.

1. After the data cleaning, the final dataset shape is 32 columns with 87230 records are there

**Feature added:**

1. Total Members: Total number of guests in every booking. By adding (Adults, Children and Babies)​
2. Total Stays: Total number of days guests stayed in the hotel. By adding (Stays in week-nights and Stays in weekend-nights)
3. After data cleaning feature addition, the dataset shape is 34 columns with 87230 records.

**Exploratory Data Analysis**

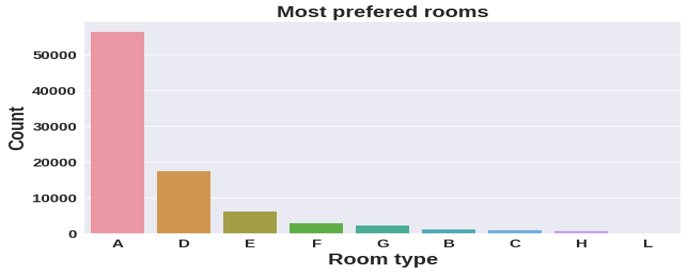


**Figure - 2 Diagram - 3**

1. From Diagram -2 we can see City hotel is the most preferred hotel with 61.3% where Resort hotel is less preferred with 38.9%

Hence, we can say the city hotel is the busiest hotel.

1. And from Diagram - 3 we can observe that Customer type Transient has more bookings with 82.4%.​Transient-party guests has 13.4%​. Contract type guests has 3.6%.​ Group type has less with 0.6%.
2. Transient: When the booking is not a part of group or contract, and it is not associated with other transient booking​
3. Transient-party: When booking is transient, and associated with at least one other transient booking​
4. Contract: When the booking has allotment or other type of contract associated to it.​
5. Group: When the booking is associated to group​



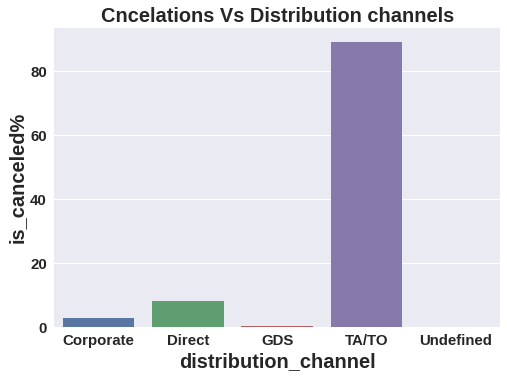
**Diagram - 4**

From Diagram - 4, Most of the guests are preferring the rooms “A”(Code of room type), So, code “A” type rooms can be increased to increase the bookings



**Diagram - 5**

By looking at the above result we can see the agent with ID Number: 9 has done more bookings 28721 and agent with ID number - 240 is in second highest in room bookings.

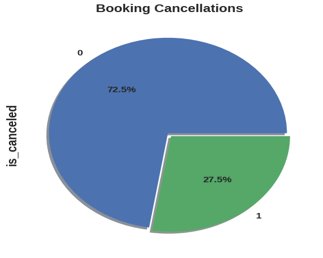
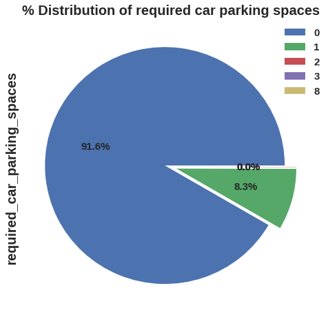


**Diagram - 6**

By looking at the above graph we can see the distribution channel "TA/TO" has more cancellations with 83.9%. Direct has 8.3%, Corporate has 2.69%, GDS channel has 0.15% and Others has 0.02%.

To reduce cancellations organization can take feedback from the guest to know what factor is causing the more cancellations.

1. Corporate- These are corporate hotel booking companies which makes bookings possible.​
2. GDS-A GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking engines and allows for automated transactions.​
3. Direct- means that bookings are directly made with the respective hotels​
4. TA/TO- means that bookings are made through travel agents or travel operators.​
5. Undefined- Bookings are undefined. maybe customers made their bookings on arrival.

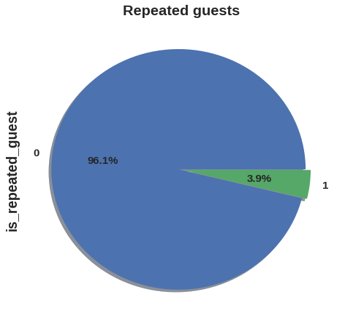
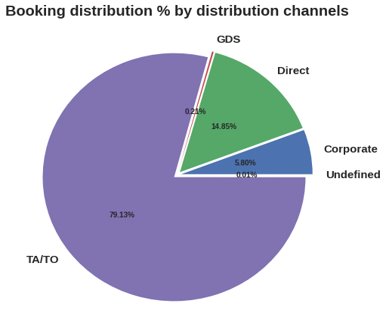
 

**Diagram - 7 Diagram - 8**

From Diagram - 7, we can see in the above result that 27.5% bookings have been canceled. To reduce cancellations organization can take feedback from the guest to know what factor is causing the more cancellations.

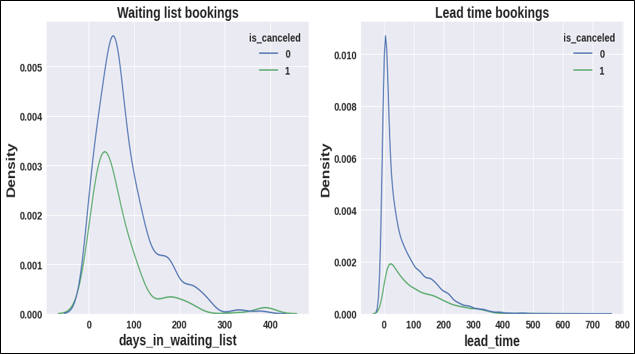
* 0 = Not canceled​
* 1 = Canceled

From Diagram - 8, 91.6% guests did not require the parking space where only 8.3% of guests required only 1 parking space.

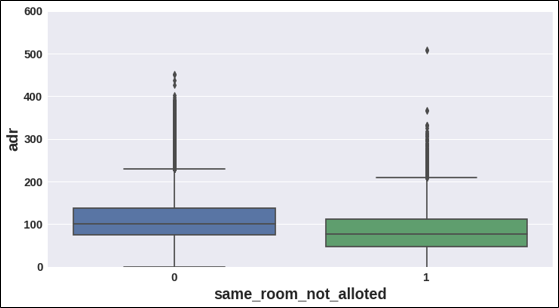
**Diagram - 9 Diagram - 10**

* From Diagram - 9, There is only few guest are repeated which is 3.9%.​ And 96.1% are not repeated·
  + 0 = Not repeated​ and 1 = Repeated​
* Most of the bookings are made by the distribution channel TA/TO with 79.13% where Direct-14.85%, Corporate-5.80, GDS-0.21% and Undefined-0.0.1%.



**Diagram - 11**

* By looking at the waiting list plot we can see less than 150 days the waiting list has been canceled and there are not canceled bookings also more in less than 150 days waiting list. Hence, we can say days on the waiting list are not much affecting the cancelation.​
* By looking at the lead time plot we can observe both the lines are similar with respect to the days on the waiting list. Hence lead time also not much affect the cancelation of bookings

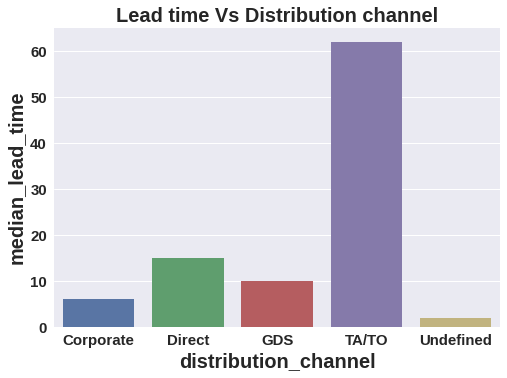
****

**Diagram - 12**

0: is same room allotted​

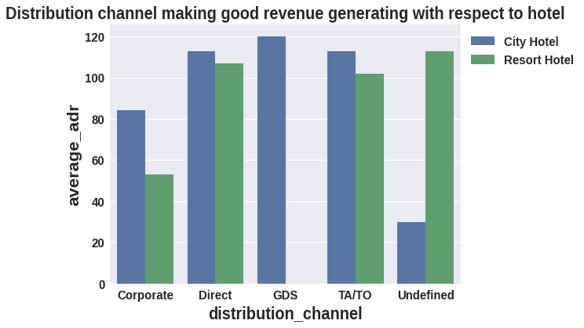
1: is Same room is not allotted​

* We can see in the plot the same room allotted and not allotted are almost similar. Hence, not getting the same room is affecting the daily "adr". Guests who are not getting the same room are paying the less adr compared to same room allotted.​



**Diagram - 11**

From Diagram - 11, Distribution channel TA/TO has the highest lead time. Hence, we can say TA/TO has the highest earlier bookings where Corporate has less lead time.



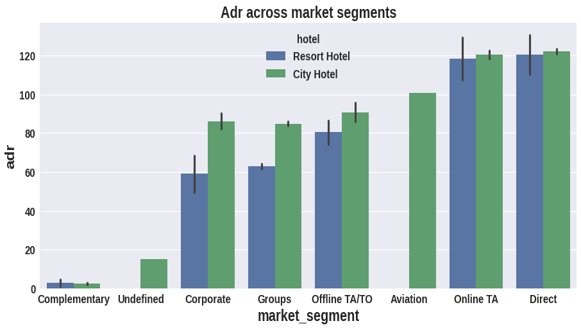
**Diagram - 12**

From Diagram - 12, Distribution channels 'Direct' and 'TA/TO' are contributing the most in both types of hotels. GDS distribution channel should focus on increasing the bookings of 'City Hotel'.



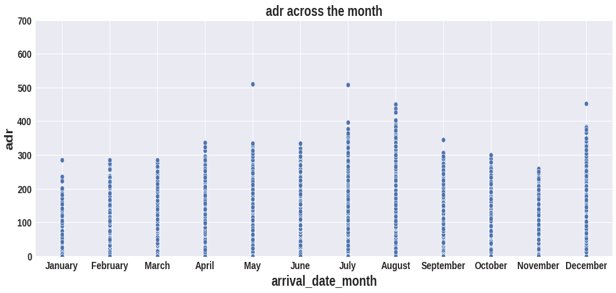
**Diagram - 13**

From Diagram - 13, The percentage of "0" changes made in the bookings was more than 82%. And the percentage of single bookings was 10%, the second highest.



**Diagram - 14**

From Diagram - 14, 'Direct' and 'Online TA' are contributing the most in both types of hotels. Aviation segment should focus on increasing the bookings of 'City Hotel'.



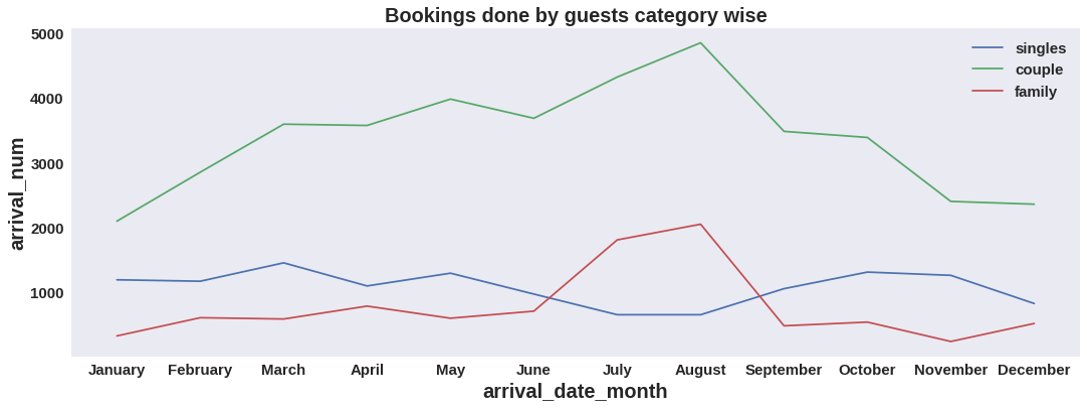
**Diagram - 15**

Avg adr rises from beginning of year up to middle of year and reaches peak in August and then lowers to the end of year. But hotels also do make some good deals with high adr at the end of year also.



**Diagram - 16**

Bookings have been increasing till the mid of year and we can see in August bookings went to highest. Hence, most of the people are planning trips in August.



**Diagram - 17**

* Most of the bookings are done by couples(although we are not sure that they are couples as data doesn't talk about that)?
* It is clear from the graph that there is a sudden surge in the arrival number of couples and families in the months of July and August. So better plans can be planned accordingly at that time for these types of customers.

**Challenges faced**

* Creating more columns based on existing data.
* Difficulty in selection of columns to get insights from data.​
* Selecting the visualization.

**Conclusion**

* City hotels are the most preferred hotels. Thus, City hotels are busiest than Resorts.​
* Transient customer type is more percentage of booking which is 82.4% and Group type is low which is 0.6%.​
* Most of the guests are preferring the rooms "A"(Code of room type). So, Code "A" type rooms can be increased to increase the bookings.​
* 27.5% bookings has been cancelled.​
* Distribution channel "TA/TO" has more cancellations with 89.13%.​
* Less than 150 days waiting list has been canceled and there are no canceled bookings also more in less than 150 days waiting list. Hence, we can say days in waiting list is not much affecting the cancelation.​
* Lead time also not much affecting the cancelation of bookings.​
* The same room allotted and not allotted are almost similar. Hence, not getting the same room is affecting the daily "adr". Guests who are not getting the same room are paying the less adr compared to same room allotted.​
* Agent with ID Number: 9 has done more bookings.​
* There is only few guest are repeated which is 3.9%.​
* 91.6% guests did not require the parking space where only 8.3% of guests required only 1 parking space.
* Distribution channel "TA/TO" has done more bookings and used for early bookings.​
* Distribution channels 'Direct' and 'TA/TO' are contributing the most in both types of hotels. GDS distribution channel should focus on increasing the bookings of 'City Hotel'.​
* Market segment "Direct" has the high adr in both Resort and City hotels where "Complementary" has less. Aviation segment can focus on City hotel.​
* Bookings have been increasing till the mid of year and we can see in August bookings went to highest. Hence, Most of the people are planning trips in august month.​
* Avg adr rises from beginning of year up to middle of year and reaches peak in August and then lowers to the end of year. But hotels do make some good deals with high adr at end of year also.​
* Mostly bookings are done by couples(although we are not sure that they are couple as data doesn't talk about that).​
* It is clear from the graph that there is a sudden surge in the arrival number of couples and families in the months of July and August. So better plans can be planned accordingly at that time for these types of customers.