* **Introduction: -**

[The key to success in the hotel industry is to find ways to exceed guests’ expectations and provide delightful perks.](https://twitter.com/share?text=The+key+to+success+in+the+hotel+industry+is+to+find+ways+to+exceed+guests%27+expectations+and+host+delightful+events.&via=socialtables&related=socialtables&url=https://www.socialtables.com/blog/hotel-sales/hotel-industry-success/)

Datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled.

We are provided here with a compact hotel bookings dataset. Our main objective is to use our knowledge of Python and perform EDA on the given dataset to draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

* **Purpose of study: -**

We will be exploring this dataset to discover important factors that govern the bookings, which contain booking information for a city hotel and a resort hotel. We will analyze important aspects of bookings which will help to give us insights to run a profitable hotel business. Like: -

• The time of year to book a hotel room?

• Optimal length of stay to get the best daily rate?

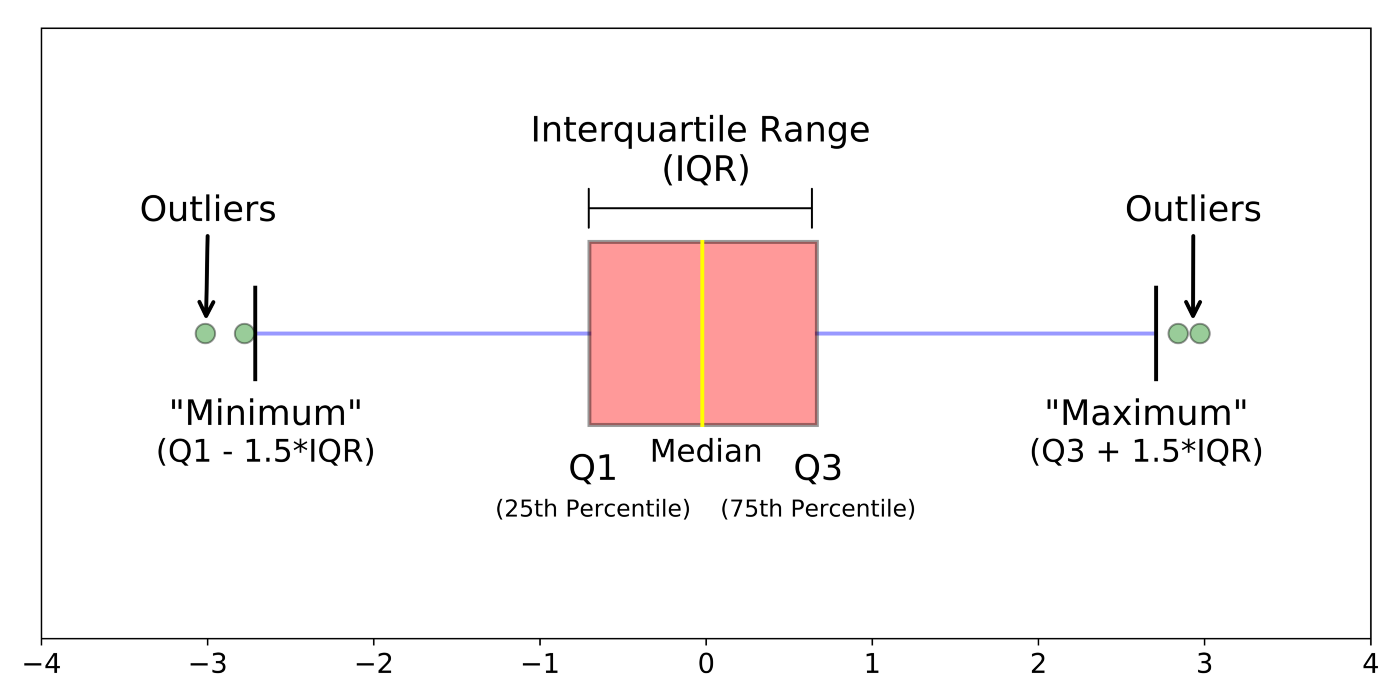
• To predict whether or not a hotel was likely to receive a disproportionately high number of special requests.

* **Attributes for the study: -**

We are given a hotel bookings dataset. This dataset contains booking information of 2 hotel types-a city hotel and a resort hotel. It has the below information: -

* **hotel**: hotel types
  + - City hotel
    - Resort type
* **is\_canceled:**
  + - 0 for no canceled
    - 1 for canceled
* **lead\_time**: time (in days) between booking transaction and actual arrival.
* **arrival\_date\_year**: Year of arrival
* **arrival\_date\_month:** month of arrival
* **arrival\_date\_week\_number**: week number of arrival date.
* **arrival\_date\_day\_of\_month:** Day of month of arrival date
* **stays\_in\_weekend\_nights**: No. of weekend nights spent in a hotel
* **stays\_in\_week\_nights:** No. of weeknights spent in a hotel
* **adults**: No. of adults in single booking record.
* **children:** No. of children in single booking record.
* **babies:** No. of babies in single booking record.
* **meal:** Type of meal chosen
  + - **BB**:- bed and breakfast
    - **HB**:-Half board (Breakfast and dinner)
    - **FB:-** Full Board (All meals included)
    - **SC:-** Self catering (No meals Included)
* **country:** Countries where the hotels are located
* **market\_segment:** market segment for booking
  + - Aviation
    - Complimentary
    - Corporate
    - Direct
    - Groups
    - Online (TA)
    - Offline (TA/TO)
* **distribution\_channel:** Via which medium booking was made.
  + - Corporate
    - Direct
    - GDS: - Global Distribution System
    - TA/TO: - Travel Agent/Operator
* **is\_repeated\_guest:**
  + - 0 for new customer
    - 1 for repeated customer
* **previous\_cancellations:** No. of previous canceled bookings.
* **previous\_bookings\_not\_canceled:** No. of previous non-canceled bookings.
* **reserved\_room\_type:** Room type reserved by a customer.
* **assigned\_room\_type:** Room type assigned to the customer.
* **booking\_changes:** No. of booking changes done by customers
* **deposit\_type:** Type of deposit at the time of making a booking
  + - No deposit
    - Refundable
    - No refund
* **agent:** Id of agent for booking
* **company**: Id of the company making a booking
* **days\_in\_waiting\_list:** No. of days in waiting to book
* **customer\_type:** Type of customer
  + - Contract: - bookings done by the contract
    - Group: - Group booking
    - Transient: - Customer staying for shorter period
    - Transient-Party: - Group of customers staying for a shorter period
* **adr:** Average Daily rate.
* **required\_car\_parking\_spaces:** No. of car parking preferred by customers at the time of booking
* **total\_of\_special\_requests:** total no. of special request.
* **reservation\_status:**
  + - checked out
    - canceled
    - not showed
* **reservation\_status\_date:** Date of making reservation status.
* Total number of rows in data: 119390
* Total number of columns: 32
* **Data Cleaning and Data Manipulation**
* **Handling null values**
* **Company Id and Agent Id: -** these columns have null values of 93% and 15% respectively. Thus, these columns are dropped
* **Country:** - this has null values less than .5% thus the null values are filled with the mode value.
* **Children and babies:** - there are only 4 null values thus the null value is filled with mean
* **Data Manipulation:** Creating columns derived by calculating the old data
  + **Kids=** Children +babies
  + **Stay=** stays\_in\_weekend\_nights+ stays\_in\_week\_nights
  + **Guest=** Adults+kids
  + **Revenue=** stay of non-cancelled guests \* ADR
* **Handling Outliers: -** Outliers are observations that lies an abnormal distance from other values in a random sample from a population. These outliers affect the result of the study. So, to clear this we have used the Interquartile Method.

  The interquartile range tells the spread of the middle half of your distribution. Quartiles segment any distribution that's ordered from low to high into four equal parts. The interquartile range (IQR) contains the second and third quartiles, or the middle half of your data set.



## **Exploratory Data Analysis**

We have used three variates to study the data:

* **Univariate Analysis:** In Univariate Analysis, we choose a single feature from the data and try to determine what the output or the target value is, i.e., one feature/variable at a time.
* **Bivariate Analysis:** Here we try to analyze two features instead of one, and determine the classification of output. It is a statistical technique applied to a pair of variables (features/ attributes) of data to determine the empirical relationship between them.
* **Multivariate Analysis:** Here wedeal with complex set of data with more than two feature and variables. There are two techniques: Dependence techniques, which look at cause and-effect relationships between variables, and interdependence techniques, which explore the structure of a dataset.

## **Data Visualization: -**

A picture speaks a thousand words. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

We have used Matplotlib and seaborn libraries to visually represent our study. The graphs used in the study are: -

* Box Plot
* Histogram.
* Pie Chart
* Bar Plot.
* Line Plot.
* Scatter Plot.
* Twin Axes
* Heatmap.
* Geomapping

## **Challenges: -**

* Big dataset is quite difficult with lots of missing values and wrong datatype format which complicates our study.
* Firstly, understanding Outliers and learning the method to handle them was a challenge. Secondly to write a code to get this result.
* To code in such a way to visualize the graphs as per rows and columns with fixed figure size to retain as per the subplots.

## **Conclusions: -**

Our study, would help guests in choosing the right hotel, right duration and right time for an optimal stay. Also, in prospective of hotel management it helps in focus of market segment, distribution channel, customer type and many more to get in most revenue.

* In our data 64% of the share is of city hotel types. Which says City hotel is more preferred than resort hotel.
* ADR is higher for the city hotels whereas after cancellations and due to longer stays revenue in city hotels is higher.
* 25% of the booking in resort and 37% of the bookings on the city hotels gets cancelled.
* Very small share of customers is repeated
* Usually, customers with a group size of 2 are more prone to cancel.
* Kids don’t play a major role in cancellations
* The booking is highest in the year 2016 for both types of hotels. ADR and revenue are on a constant rise for all 3 years for the city hotel whereas thought the booking is high for the resort type hotel in 2016 the revenue dips as the ADR dips.
* The peak time for city hotels is between the month of May to August whereas for resort type hotel the peak is in July and August. The revenue and ADR show results similar to this.
* Most preferred meal type is BB. The ADR is high for meal types of HB followed by BB in city hotels whereas in resort type hotel its HB and FB.
* Online TA are a major share for booking but revenue after all cancellation is by Offline TA
* The major revenue generated is by the Transient type customers
* Refundable deposit type has higher ADR but bookings with no deposit have high chance of cancellation.
* Car parking don’t play a major role for customer booking.
* Some customers tend to plan ahead and book with a reasonable lead time in city hotels. We also see that more the lead time more the chances of cancellation and also more requests for booking changes
* City hotels are in higher demand thus the wait time for it is high. Also, its seen that the ADR is lesser if booking is done ahead of time.
* Resort hotels get a greater number of special requests.
* Room type A and D are more in demand. But the changes of assigned room type does not affect the cancellations
* City hotels are preferred for brief stays of around 2 to 3 days whereas the resort type usually is preferred for either a day or a week. Also, the ADR is higher with shorter stay but it reduces as the stay extends
* The most remote islands like United States Minor Outlying Islands have pretty expensive rooms.
* Portugal has around 28% of the hotel share which makes it the country with the most hotels in our data with highest revenue earned.

**Thank You!!!!!**