## Abstract:

In this project, we will try to predict the possibility of a booking for a hotel based on different factors and also try to predict if they need special requests based on different features. The data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. From it, we can understand the customer's' behavior and it might help us make better decisions.

***Keywords: data analytics, pricing, variables***

## Introduction:

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series or decision trees.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

## Problem Statement:

1. We will be analyzing some key metrics for hotel bookings like:

- The number of cancellations
- Number of bookings on weekday vs weekends
- Most preferred meal types
- Country wise bookings
- New customers acquired
- Customer lifetime value of the existing customers
- Type of rooms preferred by customers
- Booking types,
- Hotels available for booking
- The revenue of the hotels

2. We will be using various lenses to look through the data to analyze patterns associated with each segment such as:

- The type of hotel
- Day of week
- Type of customers
- Type of rooms

3. Finally, we will also try to predict the future bookings either based on time series analysis or decision trees.

Using the results from the above analysis, business can make key decisions regarding the customer experience they desire to deliver.

## Hotel:

- Resort hotel
- City hotel

Provided data set has different columns of variables crucial for hotel bookings. Few of them are as below:

Hotel: The category of Hotels. There are 2 Categories of the Hotel:-

## Room details:

reserved_room_type

- Type of room reserved stored in alphabet codes.

assigned_room_type

Type of room reserved stored in alphabet codes

## Bivariate Analysis:-

- Type of the Room **A** is in most demand by the Customers.
- Type of the Rooms D, E and F are some of the highest adr(average daily rate) generating rooms.
- Type of Room A is generating more profit to the Hotel.
- In the year 2016 most number of the Hotel Rooms were booked as in comparison of the year 2015 and 2017
- The TA/TO (Travel Agents/Travel Offices)were able to book the Highest number of the Hotel Rooms.

- Rooms booked by the Online Travel Agents/offices(Online TA/TO) are having Highest Number of Bookings in comparison of the offline Travel Agent/Travel Offices(Offline TA/TO).
- The Customers generally likes to came in the month of October and May

- **Hotel Wise Analysis:-**

- Nearly around 60% bookings are for the City hotel and 40% bookings are for the Resort hotel.
- The City hotel has significantly higher bookings , hence City Hotel is much busier than Resort Hotel.
- The City Hotel is generating more Revenue than the Resort type Hotel by looking at the trend of booking analysis
- Approximately 30 % of the City Hotel bookings and 25 % of the Resort hotel bookings got canceled.
- There are many Transient(staying in a place for a short period of time) Customers stayed In the Hotels in comparison of other guest categories.
- As per the Trends The Majority of the stays are less than for the 5 days. There are very few long stays at hotels but Resort Hotel is preferred for long stays.
- The Customers preferers to Stay at weekend nights

## What is the most common channel for making the booking of the hotels?

1. market_segment

Market segment distinction Provides source of information through which customer booked

- Term "TA" - "Travel Agent"
- Term "TO" - "Tour operators"
- Both "TA" and "TO" are considered the same kind of market segment.

2. distribution_channel
   - It is also called "marketing channel"
   - It is the Network through which customer booked

3. customer_type
   - 'Transient' - Simply individual guests requiring a short stay at the hotel
   - 'Contract' - Agreement between hotel authority and customer to require volume room bookings on contract basis.
   - 'Transient-Party' - Booking is Transient and associated with other transient booking
   - 'Group' - Multiple rooms are booked under single customer responsibility

## Which distribution channel has highest the cancellation percentage?

- While doing the Analysis we were able to figure out that TA/TO has highest booking cancellation percentage. Therefore, a booking via TA/TO is almost 28% likely to get cancelled.
- There can be n reasons to get the Hotel rooms to be cancelled, but while looking at the Trend many cancelation occurs during check in and check outs the majority reason can be that they are not getting the same rooms which they have booked.
- Thus by looking at the Trend the TA/TO are affecting the businesses more.

The Most number of the Cancellation are done by not paying any deposit

## In which year the Highest Number of Bookings were Done?

Information regarding the timeline such as year, week, month used to get the insights of the usage of hotels in particular periods which helps to find demand for hotels. arrival_date_year

- The Year 2016 most number of the Hotel Rooms were booked as in comparison of the year 2015 and 2017

## Which Months are the most busiest months for the hotels?

- The Customers generally likes to came in the month of October and May
- The Customers preferers to Stay at weekend nights
- There are few customers who prefers to make changes to the bookings

## Steps involved:

1. Understand the data.
2. Univariable study.
3. Multivariate study.
4. Basic cleaning.
5. Test assumptions

### Data Collection and Understanding

Data understanding focuses on the comprehension of the information available in the project. In this step we basically check on the kind of variables provided with the dataset, dtype of the columns, shape of the data frame.

### Null values Treatment

Our dataset contains numbers of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

Pandas **isnull()** and **notnull()** methods are used to check and manage NULL values in a data frame The percentage of null values in each variable is found using the following formula.

$$\text{Percentage} = \frac{\text{Number of null values}}{\text{Total number of values}}$$

### Encoding and dropping of categorical columns

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format

Hence, filling those null values with appropriate values i.e., filling the null values in the children column as "0" Filling the null values in the country with the country name which has maximum count in the data.

- **The Main center of discussion is the analysis of given hotel bookings data set from 2015-2017.**
- **The Team Creation has done the analysis of given data set**

## Conclusion:

- Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. Also the overall average daily rate and revenue of the City hotel is slightly higher than the Resort hotel.
- Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.
- Customers used different channels for making bookings out of which most preferred way is TA/TO.
- July- August are the most busier and profitable months for the both of the hotels.

- Nearly 28% of bookings via TA/TO are cancelled.
- In the Year 2016 most Guest/Customers came to the Hotels.
- Customers preferred to stay in weekends rather than weekdays.
- Booking made by the TA/TO are Resulting the more repetition of the Customers.
- The Second most booking Channel is the Direct booking done by the Customers at the Time of arrival.
- City Hotels are more liked and booked by the Customers.
- hotel or from somewhere else that's why they do not need parking space.

- The data for 2015 and 2017 is for different months. Even though we have converted them to same base line of weekly numbers, there are chances that some weeks perform differently as compared to other weeks
- The definition of new customers is not very well described. A new customer this year will be existing next year, or they can be existing customer from the 2nd booking. A deeper analysis in required based on definition
- The weekday vs weekend analysis can be further drilled for the type of bookings
- The classification model uses few variables. We can tune the model with new variables and adjusting the cost of misclassification. Additionally, we need to split the data into train and validation to avoid overfitting

- The forecasting aspect can be further drilled to analyze the residuals and split the model across years and various factors such as customer type or hotel type

**References-**


1. https://pandas.pydata.org/
2. https://stackoverflow.com/Analytics
3. https://www.wikipedia.org/