# EDA- Hotel Booking analysis

**Objective**

We receive a dataset of hotel reservations.

Our major goal is to utilize EDA on the provided dataset to analyze general hotel booking patterns and the relationships between the various factors that influence hotel bookings.

**Dataset**

A dataset of hotel reservations is handed to us. A city hotel and a resort hotel's reservations are included in this dataset. The following characteristics are present.

* hotel: Name of hotel (City or Resort)
* is\_canceled: Whether the booking is canceled or not (0 for no canceled and 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
* country: Country of origin of customers (as mentioned by them)
* market\_segment: What segment via booking was made and for what purpose.
* distribution\_channel: Via which medium booking was made.
* is\_repeated\_guest: Whether the customer has made any booking before(0 for No and 1 for

Yes)

* 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 on waiting list.
* customer\_type: Type of customer (Transient, Group, etc.)
* adr: Average Daily rate.
* required\_car\_parking\_spaces: No. of car parking asked in booking
* total\_of\_special\_requests: total no. of special request.
* reservation\_status: Whether a customer has checked out or canceled, or not showed
* reservation\_status\_date: Date of making reservation status.
* There are 119390 rows in all.
* There are 32 columns overall.

**Removing Duplicate** Rows is the first step in data cleaning and feature engineering.

**Dealing with empty values**

Company and Agent columns' null values were changed to 0s.

Column country's null values were changed to "others" values.

The column mean was used to substitute null values in column children.

**Data Cleaning and Feature Engineering**

(1) **Removing Duplicate Rows** is the first step in data cleaning and feature engineering.

(2) **Dealing with empty values**

Company and Agent columns' null values were changed to 0s.

Column country's null values were changed to other values.

The column mean was used to substitute null values in column children.

3**) Changing the data types in the columns.**

Children, companies, and agents' data types were changed to int types.

Reservation status date's data type was changed to date type.

(4) **Eliminating anomalies**

In the adr column, one outlier was discovered. It was merely dropped.

(5) **Adding fresh columns**

By combining stays over the weekend and stays during the week, the new column total stay was created.

The total people column was created by combining adults, kids, and newborns.

**Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:**

* Bar Plot.
* Histogram.
* Scatter Plot.
* Pie Chart.
* Line Plot.
* Heatmap.
* Box Plot

**Exploratory data analysis**

**Univariate**

Q1) Which room type generates the most adr and is in the highest demand?

Q2) Which type of dish is the most popular among customers?

Q3) What is the average number of reservations for each hotel?

Q4) Which channel is used the most frequently to make hotel reservations?

Q5) Which months are the busiest?

Q6) What country are most of the visitors from?

Q7) What is the average length of stay at hotels?

**Analysis of univariate variables:**

1)The most popular room type is A, however rooms H, G, and C produce greater advertising. To maximize revenue, hotels should offer more A and H room types.

2) The most common form of dinner is BB (Bed and Breakfast).

3) The City Hotel is busier than the Resort Hotel since 60% of reservations are for the City Hotel and 40% are for the Resort Hotel.

4) Customers make reservations through a variety of channels, with TA/TO being the most popular option.

5) The busiest and most lucrative months for both hotels are July and August.

6) The majority of the visitors were from European nations, with the majority coming from Portugal.

7) The average stay is shorter than 4 days, and city hotels are typically preferred for short stays, but for long stays resort hotel is preferred.

**Bivariate**

Q1) Which hotel seems to make more revenue?

Q2) Which channel is mostly used for the early booking of hotels?

Q3) Which distribution channel brings better revenue-generating deals for hotels?

Q4) What is the trend of bookings within a month?

Q5) What is the most reserved room type in market segment

Q6)In which month most people arrive at both hotels?

**Analysis of Bivariate variables**

1) Overall adr of City hotel is slightly higher than Resort hotel and no. of bookings of City hotel is also higher than Resort hotel. Hence, City hotel is makes more revenue.  
2) TA/TO is mostly used for planning Hotel visits well ahead of time.

3) GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via TA/TO. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals.

4) Arrivals in hotels increases at weekends and the avg adr tends to go up as month ends.

5)Overall in market segments the reservation for type A room is higher and among the reservations online TA are more in bookings

6) August is the month where most no. of people visited the hotels

**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 adr of City hotel is slightly higher than Resort hotel.
* Most of the guests came from european countries, with most of guests coming from Portugal.
* Guests use different channels for making bookings out of which most preferred way is TA/TO.
* July- August are the busier and most profitable months for both of hotels.
* Within a month, adr gradually increases as month ends, with small sudden rise on weekends.
* More number of people in guests results in more number of special requests.
* Bookings made via complementary market segment and adults have on average high no. of special request.
* And many more conclusions.
* For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.

**Challenges**

* There was a lot of duplicate data.
* Data was present in wrong datatype format.
* Choosing appropriate visualization techniques to use was difficult.
* A lot of null values were there in the dataset.