

# EDA ON HOTEL BOOKING ANALYSIS

STAR SHAPIONS

RISHAV KUMAR, TEJAS PATIL, SYED FARUK, SHRADDHA GUJAR, ARUN  
PALLE.

## Introduction -

We were given a hotel booking dataset.csv file. By the help of that given dataset.csv file. We have to analyse the given data and perform EDA on a given dataset.csv file. Our objective is to reach a logical as well as useful result about the rise and fall in hotel booking and how elements affecting hotel booking associate with each other.

## Dataset -

This data set contains booking information for a city hotel and resort hotel. It includes information such as type of the hotel, how many time it is cancelled, lead time, arrival date & month, stays in weekend nights, stays in week nights, children, adults, babies, country, market segment, distribution channel, is it repeated guest, previous cancellation, previous bookings, reserved room type, agent, company, customer type, required car parking & reservation status etc.

Understanding the whole dataset with help of column names. Given below is clearly mentioned.

hotel : Name of hotels

is\_canceled : Indicating the booking was cancelled (1) or not cancelled (0)

lead\_time : Number of days that elapsed between the entering data of booking

arrival\_date\_year : Year of arrival date

arrival\_date\_month : Month of arrival date

arrival\_date\_week\_number : week number of year of arrival

arrival\_date\_day\_of\_month : Day of arrival date

stays\_in\_week\_nights : The number of weekend nights ( saturday and sunday ) the guest stayed in hotel

stays\_in\_week\_nights : Number of week days ( monday to friday )the guest stayed in hotel

adults : Number of adults stayed in hotel

children : Number of children stayed in hotel

babies : Number of babies stayed in hotel

meal : type of meal booked by customers

country : country of origin

market\_segments : 'TA' means travel agent and 'TO' means team operators

distribution\_channel : Booking distribution channel

is\_repeated\_guest : Repeated guest (1) or not repeated guest (0)

previous\_cancellations : Number of booking that were cancelled by customers

previous\_bookings\_not\_canceled : Number of bookings that were not cancelled by customers

reserved\_room\_type : Code is represented by room which is booked by customer

assigned\_room\_type : code is type of room assigned to the booking

booking\_changes : Number of charges made to the booking

deposit\_type : Indicates on the customer made a deposit to guarantee the booking

agent : ID for travel agency

company : Company ID entity that made booking or responsible for booking payment

days\_in\_waiting\_list : Number of days from booking to confirmation booking

customer\_type : booking assuming for four categories

adr : Average daily rate sum of all loading transactions dividing by total number of staying nights

required\_car\_parking\_spaces : Car parking space required by customer

total\_of\_special\_requests : Total special requests made by customer

reservation\_status : Reservation status, assuming in three categories

reservation\_status\_date : Date of the last status was set

## Data cleaning –

1. detecting duplicates.
2. Handled null values  
Null values column name company is dropped.
3. Null values in column name countries were replaced by 'others'.  
Null values in columns names agent and children were replaced by '0'.
4. Detecting outliers

## Questions performed in EDA -

1. Which hotel is most preferred by customers?
2. Which month visitors visit highly?
3. Which type of room is highly booked and preferred by customers?
4. Which year got the best sales?
5. Which hotels are mostly cancelled by the customers?
6. Which type of customers are highly visited on both hotels?
7. What is the percentage of repeated guests?
8. What is the percentage distribution of deposit type?

## Libraries and tools used in EDA -

1. Pandas
2. Numpy
3. Seaborn
4. Matplotlib

## Graphs & plots been used -

1. Count plot
2. Pair-plot
3. Heat map
4. Box plot

5. Dist plot
6. Pie chart

### Challenges faced –

1. Huge amount of data was present in the dataset.
2. Dealt with some missing values.
3. Huge amount of null values were present in the dataset.
4. Faced difficulties in understanding the data.

### Final outcome (Result) -

We learnt

1. Guests mostly preferred city hotels because city hotels have maximum bookings.
2. August is one of the months with a high amount of visitors.
3. Code 'A' rooms are most preferred by customers because code 'A' rooms are highly booked by customers.
4. In 3 years of data we got to know sales of 2016 are higher than 2015 & 2017.
5. City hotels are mostly cancelled by the customers after booking.
6. Transient types of customers are highly visited in both hotels.
7. 3.2% of customers are repeat guests.
8. 87.6% of data is distributed in deposit type.

### Conclusion –

We analyzed the entire hotel booking dataset. We covered a range of factors affecting the hotel booking. We estimated the cancellation rates, booking ratios according to months. most preferred hotels, most preferred type of room. Percentage of repeated guests. And the time when they have highest and lowest numbers of visitors by Exploratory data analysis. We also tried to predict whether a hotel is likely to receive a disproportionately high number of guests.