**Hotel booking exploratory data analysis**

**Amit Adhaye, Kanhaiya Kurmi,**

**Vaibhav Jaiswal**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

The hotel booking is one of the generalized activities by all and not researched a lot. Thus, the data recording and analysis can play an important role in optimizing the service and thus improve the overall efficiency in business.

Our analysis can help understand what could be the causative parameters to focus upon for better customer satisfaction. The plots shown for price per guest, month, year, country and lead time wide trends, various trends of market segment analysis, demands for food, parking space etc were fruitful in answering the 8 questions identified in problem statements.

***Keywords: Data wrangling, Data cleaning, exploratory analysis, price per guest, data visualization.***

**1.Problem Statement**

In day to day life plenty of events are happening near to us. These events generate lot of data. Recording the data and analysing the highlighted events proved vital in generating the main characteristics using statistical graphs and other data visualization tools like python matplotlib, seaborn libraries.

Exploratory Data Analysis (EDA) on Hotel Bookings data which is basically present in the form of comma separated value(csv) file. there are many fields that contains different values regarding Resort Hotel Bookings and City Hotel Bookings.

In this capstone Project we are going to analyze a dataset comprising the demand of hotel bookings. The data set contains booking information for a city hotel and a resort hotel, and includes information such as market segments, cancellation, rate of stay, food ordered, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

In this EDA we will perform analysis on following Questions :

1. What is the bookings ration between Resort and City hotels?
2. How many tickets were canceled during booking and after booking?
3. Which month is the most occupied and which is the least occupied?
4. What is % booking for each year?
5. Which market country is contributing max booking?
6. What meal item mostly preferred by customer?
7. what is the price per guest in each booking?
8. Which market segments should be targeted for promotion?

**2. Introduction**

In this project, we will discuss exploratory data analysis and data visualization of the hotel booking data set. In this project, we need to find the average fees of the hotel and explore the data. The dataset provided were including information regarding following fields:

1**. hotel:** type of hotels  
**2.** **is\_canceled:** canceled or not  
**3. lead\_time:** no. of days before actual arrival in the hotel  
**4. arrival\_date\_year:** year of booking  
**5. arrival\_date\_month:** month of booking  
**6. arrival\_date\_week\_number:** week number of the year in which  
booking  
**7.arrival\_date\_day\_of\_month:** arrival month date  
**8. stays\_in\_weekend\_nights:** no. of weekends guest stayed **9. stays\_in\_week\_nights:** no. of weekdays guest stayed **10. adults :** no of adult guest  
**11. Children :** no of children guest  
**12. babies**   
**13. meal:** BB – Bed & Breakfast  
HB – only two meals including breakfast meal,FB – breakfast, lunch, and dinner  
**14. country**   
**15. market\_segment:** TA: Travel agents  
TO: Tour operators  
**16. distribution\_channel**   
**17. is\_repeated\_guest**   
**18. previous\_cancellations:** cancellation in past  
**19. previous\_bookings\_not\_canceled:** not cancelled in past  
**20. reserved\_room\_type**   
**21. assigned\_room\_type**   
**22. booking\_changes**   
**23. deposit\_type**   
**24. agent**   
**25. company**   
**26. days\_in\_waiting\_list**   
**27. customer\_type**   
**28. adr:** average daily rate **29. required\_car\_parking\_spaces**   
**30. total\_of\_special\_requests**   
**31. reservation\_status**   
**32. reservation\_status\_date**

**Contents of the Project:**

Installing and importing the required libraries

Data preparation and cleaning

Exploratory analysis and visualization of columns

Explaining the need and outcome of analysis

## **3. Data exploration and cleaning**

The stepwise process in data exploration and cleaning is:

Copy the content of original dataframe into new dataframe , so that we keep original data set safe to protect from unauthorized changes in dataset

We start first descriptive analysis on given Hotel Bookings dataset, to know brief about dataset and content present in the dataframe.

Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a Hotel Bookings dataframe and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Dealing with missing values : check if our dataset contains any null values

We have five columns with missing values: we just replace with some value

Children column contains the count of children, so we will replace all the missing values with the rounded mean value.

And our country column contains country codes representing different countries. It is a categorical feature so I will also replace it with the mode value. The mode value is the value that appears more than any other value. So, in this case, I am replacing it with the country that appears the most often.

**4. Exploratory Data Analysis**

This section deals with answering the questions framed in section 1 and building some relationship among various columns, which directly identifies the outliers for hotel booking demand analysis.

1. What is the bookings ratio between Resort and City hotels?

To calculate the ratio of Resort and City hotels we have created a method to get bookings ratio. method will take column name as first parameter and limit as another parameter.

The function mentioned in notebook takes series or a data frame column as input and return two arrays p and q.

Where:p is Unique values and q is percentage value of each unique value.

There are two unique values in hotel coulmn 'City Hotel' and 'Resort Hotel' percentage booking of city hotel is 66.4075 and Resort hotel is 33.595. More than 60% of the population booked the ‘City hotel’.

2. How many tickets were canceled during booking and after booking?

‘is\_cancelled’ column have two unique values: 1 if booking got cancelled, else 0.

Now let’s plot this result. I will write another function to plot the diagram. The good thing about writing function is that we can reuse the code again and again

This function takes two arrays, x, and y and displays the required diagram. The default plot type is a bar plot, but it can also plot the line plot. Optional arguments can be given to display title and labels.

Bookings got canceled 37% of the time. While booking guest did checked-in (did not cancel the booking) almost 63% of the time.

The hotel booking is generalised activity which involves the cancellation, stay in weed days and weekends. It involves different waiting times for different hotels. Thus, the bar chart below proposes the average number of cancellations, stays in weekdays and weekends, and waiting time in days for two categories of hotels.

It appears that City Hotel has more reservations than Resort Hotel and I believe that is also the reason why it has more cancellations as well. The stays in weeknights and weakened nights are more in resort hotel than city hotel. The waiting time in city hotel is more than resort hotel.

The data in hand consist of bookings data for all months in year 2015,2016 and 2017. So, it is worth noting the percentage mix of month wise bookings. This chart may help to identify the most preferred month for booking.

3. Which month is the most occupied and which is the least occupied?

## It is observed from plots that August month has more booking request as compare to other. January month has least booked.

4.What is Hotel Booking % for each year?

The data in hand consist of bookings data for year 2015,2016 and 2017. So, it is worth noting the trend of year wise bookings. This trend may help to forecast the future demand. The year wise values and trend is plotted in notebook.

## Year 2016 has maximum booking where year 2015 has minimum hotel bookings

5. Which country is contributing max booking?

According to data, We can see that 'Portugal' and 'United Kingdom' have more bookings as compare to other countries in plots shown in notebook.

We will have to target these countries to increase the bookings for both hotel types.

6. What item mostly preferred by customer?

The bookings at the hotel comprises of stay at nights, thus needs meal. The following pie chart presents the percentage of each food type preferred in each booking.

It is observed from plots that bed and breakfast (BB) is mostly preferred item (57730 bookings).

**7.** Which market segments should be targeted for promotion?

Market segmentation analysis is very essential for any business. One can not only identify the performing segment but also non-performing segment, thus upgrade it. Online TA is identified as largest contributor amongst all.

8. what is the price per guest in each booking?

ADR stands for average daily rate, and it's used to measure the average revenue that a hotel receives for each occupied guest room per day. This feature is important in the upgradation of business plan in near future.

The equation to calculate price per guest is

Total Guests stayed = no of (adults + children + babies)

Price per guest = ADR / total guest stayed

It is very interesting to know the revenue earned by various hotel categories for years under considerations. The box plot below gives inside about the above mentioned relationship.

The box plot shown in notebook presents the effect of type of room and meal type on the final price per guest. Thus, guest can plan the booking optimally considering the individual costs.

It will be interesting to see how much the customers coming from various channels pay daily. It can be observed that TA/TO distribution channel is the best performer.

**5. 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.

Mostly guests stay for less than 5 days in hotel and for longer stays Resort hotel is preferred.

Most of the guests came from european countries, with most no. of guest coming from Portugal.

Guests use different channels for making bookings out of which most preferred way is TA/TO.

For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.

Almost 30% of bookings via TA/TO are cancelled.

Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment does lowers the adr.

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.

Couples are the most common guests for hotels; hence hotels can plan services according to couples needs to increase revenue.

In this hotel booking EDA, our team has identified the various outliers in different category of fields in the data such as year wise, month wise bookings, stay, cancellation and lead time comparisons between two hotel categories, market share of each entity w.r.t cancelled and booked, meal preferences by guest, the price per guest and its relationship to type of room and meal type etc.,. This EDA will play important role in identifying various key features in hotel booking demand analysis

**References-**

1. kaggle
2. GeeksforGeeks

3. Stackoverflow