**Hotel Booking Analysis**

**Ajay Singh, Ajeet Kumar,  
 Laxmi Priya, Shruti Sharma**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

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

**Keywords: hotel industry, booking cancellations, adr, trends, hotel bookings**

**1. Problem Statement**

* In this project we will analyze the data of the hotel booking dataset.
* This hotel dataset contain booking information for city and resort hotels with their corresponding variables such as canceled bookings, arrival data per annum, arrival data per month, arrival data per day , types of guests( children , adults , babies) and company etc.
* Hotel booking is a very big field and depends upon the different factors of such as it’s type of booking data , date ,year , month ,types of meal etc.
* Our main purpose behind this analysis is to fetch the important data factor to check in which time we can get hotel booking in minimum/maximum price , types of meal guests does prefer to have, ratio of babies , children , adults who do come in hotel.
* When do guests prefer to come in week.

**Data Description:**

* **hotel** : Hotel (Resort Hotel or City Hotel)
* **is\_canceled** : Value indicating if the booking was canceled (1) or not (0)
* **lead\_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
* **arrival\_date\_year** : Year of arrival date
* **arrival\_date\_month** : Month of arrival date
* **arrival\_date\_week\_number** : Week number of year for arrival date
* **arrival\_date\_day\_of\_month** : Day of arrival date
* **stays\_in\_weekend\_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* **stays\_in\_week\_nights** : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* **adults** : Number of adults
* **children** : Number of children
* **babies** : Number of babies
* **meal**: Type of meal booked. Categories are presented in standard hospitality meal packages:
  + Undefined/SC – no meal package
  + BB – Bed & Breakfast
  + HB – Half board (breakfast and one other meal – usually dinner)
  + FB – Full board (breakfast, lunch and dinner)
* **country**: Country of origin. Categories are represented in the ISO 3155–3:2013 format
* **market\_segment** : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
* **distribution\_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
* **is\_repeated\_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0)
* **previous\_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
* **previous\_bookings\_not\_canceled** : Number of previous bookings not cancelled by the customer prior to the current booking
* **reserved\_room\_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.
* **assigned\_room\_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
* **booking\_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
* **deposit\_type** : Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
  + No Deposit – no deposit was made
  + Non Refund \* a deposit was made in the value of the total stay cost
  + Refundable – a deposit was made with a value under the total cost of stay.
* **agent** : ID of the travel agency that made the booking
* **company** : ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
* **days\_in\_waiting\_list** : Number of days the booking was in the waiting list before it was confirmed to the customer
* **customer\_type** : Type of booking, assuming one of four categories:
  + Contract - when the booking has an allotment or other type of contract associated to it
  + Group – when the booking is associated to a group
  + Transient – when the booking is not part of a group or contract, and is not associated to other transient booking
  + Transient-party – when the booking is transient, but is associated to at least other transient booking
* **adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
* **required\_car\_parking\_spaces** : Number of car parking spaces required by the customer
* **total\_of\_special\_requests** : Number of special requests made by the customer (e.g. twin bed or high floor)
* **reservation\_status** : Reservation last status, assuming one of three categories:
  + Canceled – booking was canceled by the customer
  + Check-Out – customer has checked in but already departed
  + No-Show – customer did not check-in and did inform the hotel of the reason why
* **reservation\_status\_date** : Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel

**2. Introduction To EDA:**

**What is EDA?**

Simply defined, exploratory data analysis (EDA for short) is what data analysts do with large sets of data, looking for patterns and summarizing the data set's main characteristics beyond what they learn from modeling and hypothesis testing. EDA is a philosophy that allows data analysts to approach a database without assumptions. When a data analyst employs EDA, it’s like they’re asking the data to tell them what they don’t know.

[EDA](https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm) as an approach to data analysis, not a model that uses these techniques:

* Maximize insights into a dataset.
* Uncover underlying structures.
* Extract important variables.
* Detect outliers and anomalies.
* Test underlying assumptions.
* Develop parsimonious models.

EDA is typically used for these four goals:

* Exploring a single variable and looking at trends over time.
* Checking data for errors.
* Checking assumptions.
* Looking at relationships between variables.

EDA methods typically fall into graphical or non-graphical methods and univariate or multivariate methods. It relies heavily on visuals, which analysts use to look for patterns, outliers, trends and unexpected results.

### **Graphical vs. non-graphical EDA**

Graphical exploratory data analysis employs visual tools to display data, such as:

* **Box plots**: used to graphically depict data through their quartiles at five data points—lowest, first, median, third and maximum values; also sometimes called a whisker plot. Analysts use it to look at large sets of data. An example of this in practice is a utility that tracks water usage on a monthly basis.
* **Heatmap**: data visualization that uses colors to compare and contrast numbers in a set of data; also known as shading matrices. An example of this in practice would be traffic analyses, which look at heavy traffic patterns by time of day, day of the week and season.
* **Histograms**: bar chart that groups numbers together in a series of intervals, especially when there is an infinite variable, such as weights and measures. For example, it can be used to measure agricultural growth where units would be grouped in height ranges (100–150 cm vs. 100, 101, 102, etc.).
* **Line graphs**: one of the most basic types of charts that plots data points on a graph; has a wealth of uses in almost every field of study.
* **Pictograms**: replace numbers with images to visually explain data. They’re common in the design of infographics, as well as visuals that data scientists can use to explain complex findings to non-data-scientist professionals and the public.
* **Scattergrams or scatterplots**: typically used to display two variables in a set of data and then look for correlations among the data. For example, scientists might use it to evaluate the presence of two particular chemicals or gases in marine life in an effort to look for a relationship between the two variables.

Non-graphical exploratory data analysis involves data collection and reporting in non-visual or non-pictorial formats.

**3. Challenges Faced:**

1. Understanding the Data set.
2. Handling Null/Duplicate values.
3. Finding relationships between different variables.
4. Plotting various types of Graphs to get insights and conclusions

**4. Approached used to Solve the challenges:**

We had the data set containing more than 119390 entries and 32 variables such as ‘hotel type’, is\_canaceled, is\_repeated\_guests, children, adults ,babies, agent,distribution channel, market segment , adr and many more.

**Step 1**

The first step was understanding the columns and their meanings. After that copy the dataset, so our original dataset remains unchanged. Here, we have created another data frame with name data\_f and for our better understanding we changed the column name in it. check the datatype of each column in our dataset. Now we Convert the data type of columns ArrivingYear, ArrivingDate, ArrivingMonth, Canceled, RepeatGuest into ‘str’.

**Step 2**

In second step we tried to find out missing values in each row. So we found that there were four columns namely ‘Company’, ’Agent’, ’Children’, and 'Country' having missing values. So for company, agent and country columns, we filled the null values with 0 and for the Children column we filled with mean value of children. After this we checked for the duplicate rows in our data. We found 31994 duplicate rows so we dropped those rows. Later we dropped the rows where the addition of children, adults and babies was 0. That means we assumed, no booking was done or booking was done but no one checked-in. Thus finally we had 87229 rows with us.

**Step 3**

The third step after getting cleaned data, was to get various insights from the data by plotting count plots, pie charts, line charts, heat map(for finding correlation between variables), scatter plot ,box plots etc and draw some conclusions from it.

**Conclusions /Insights from EDA:**

* 'PRT'(Portugal )has the highest no of visitor (27355) while 'GBR'(Great Britain) has the second highest no of the visitor(10424) and 'FRA'(France)has the third highest no of the visitors(8823)
* August is the highest reservation month (11242) meanwhile the lowest reservation month is january(4685)
* BB is the most popular meal ordered by the visitor(67907)
* City Hotel have the Highest reservation by the visitors as compare to Resort Hotel
* The 72.4% booking were not canceled and 27.5% booking have been canceled
* Online bookings are more preferred. 70% bookings are made via online mode.
* The highest cancelation rate was found in City Hotel and Highest booking was also found in City Hotel.
* The bare minimum number of the guests were repeated
* Most of time the visitors were arrived in the year of 2016 and 2017 in the city hotel.
* The maximum of visitors like city Hotels.
* The prices for resort hotels were higher and fluctuate more than city hotels.
* City hotel: the ADR was more costly during the month of july,august, may, june; Resort Hotel: the ADR was more costly in during the month of july, august, june

References:

* [http://Mastersindatascience.org](http://mastersindatascience.org)
* rpubs by rstudio
* analytics vidhya