**Hotel Booking Analysis**

**Abhash Jain, Karan Rawat, Mounika Dontula**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Data analytics in hotel and resort brands is a growing strategy implemented to support business decisions designed to generate revenue or save costs. This study contains the real-world data record of hotel bookings of a city and a resort hotel containing details like bookings, cancellations, guest details etc. from 2015 to 2017. Main aim of the project is to understand and visualize dataset from hotel and customer point of view i.e. reasons for booking cancellations across various parameters, best time to book hotel and peak season and also gives suggestions to reduce these cancellations and increase revenue of hotels.

***Keywords: Python, Data cleaning, Pandas, Matplotlib, Data visualization etc.***

**Problem Statement**

Hotel and resort operators spend a considerable amount on marketing costs to attract the most valuable customers in increasingly competitive markets. As marketing strategies become more complex, companies seek a method to identify their most valuable customers and evaluate current marketing systems. Data analytics is a trending business strategy across many industries.

Objective is performing EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

**Introduction**

The discussion of data analytics would not be complete without an understanding of what is the data that is used in analytics. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. EDA focuses more narrowly on checking assumptions required for model fitting and hypothesis testing. It also checks while handling missing values and making transformations of variables as needed filling the counts with.

EDA build a robust understanding of the data, issues associated with either the info or process. It’s a scientific approach to get the story of the data.

An interpreted, object- oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you’ll decide the way to handle missing values for EDA.

**Types of exploratory data analysis:**

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

## **Dataset used for EDA**

For the simplicity of the article, we will use a single dataset.We will use the hotel booking for this. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features:

* 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.

**Methodology:**

## Business understanding

Analytics in the hotelier world today is important, and nowadays this business cannot be run with some sensible and smart use of data.

Here I demonstrate how to use data to analyse three business important concepts in the fields of revenue management and marketing.

The analysis tries to answer three questions

1. How strong is the seasonality in these hotels?
2. Up to what point ADR, length of stay, and lead time for groups reservation differ from individual/transient ones? are rooms harder to sell in nearby dates of a group's stay?
3. Can we predict a cancellation, just with the information available at the moment this reservation has been made?

## Data understanding

The original data has been extracted from the booking changelog, one day prior to arrival date, to avoid leakages. There are some variables that had to be extracted from, other DB systems. The data presented consists of booking records from two hotels: one resort hotel and one city hotel.

In total there are 119,390 records and 32 features, with all of these features presenting almost (or none) null values, except for the variable company, agent, children etc.

### **Installations**

Below is a list of the modules used in the analysis is shown

* pandas
* NumPy
* matplotlib
* seaborn
* plotly

## **Handling Missing Values**

You all must be wondering why a dataset will contain any missing value. It can occur when no information is provided for one or more items or for a whole unit. For Example, suppose different users being surveyed may choose not to share their income, some users may choose not to share the address in this way many datasets went missing. Missing Data is a very big problem in real-life scenarios. Missing Data can also refer to as NA (Not Available) values in pandas.

(1) Removing Duplicate rows

* All duplicate rows were dropped.

### (2) Handling null values

* Null values in Columns Company and agent were replaced by 0.
* Null values in column country were replaced by 'NA'.
* Null values in column children were replaced by 0.

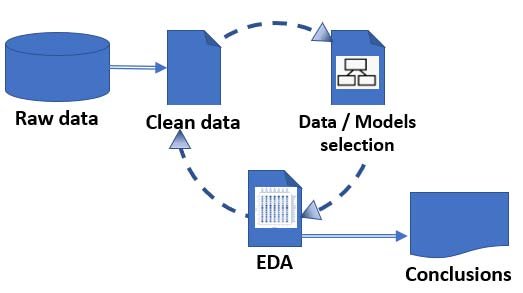
### (3) Converting columns to appropriate data types

* Changed data type of children, company, agent to int type.
* Changed data type of reservation\_status\_date to date type.

(4) Checking for outliers

* The mean and median difference is quite large for most of the features.
* We are removing those outliers.

## Exploratory Data Analysis



## **Data visualization**

Data Visualization is the process of analysing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data. There are various types of visualizations –

* **Univariate analysis:** This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
* **Bi-Variate analysis:** This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
* **Multi-Variate analysis:** When the data involves three or more variables, it is categorized under multivariate.

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

* Bar Plot.
* Pie Chart.
* Line Plot.
* Box Plot
* Count plot
* Heat map

# **Bar Plot:**

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

# **Pie chart:**

A **Pie Chart** is a circular statistical plot that can display only one series of data. The area of the chart is the total percentage of the given data. The area of slices of the pie represents the percentage of the parts of the data. The slices of pie are called wedges. The area of the wedge is determined by the length of the arc of the wedge. The area of a wedge represents the relative percentage of that part with respect to whole data. Pie charts are commonly used in business presentations like sales, operations, survey results, resources, etc as they provide a quick summary.

**Subplots:**

Subplots mean **groups of axes that can exist in a single matplotlib figure**. subplots() function in the matplotlib library, helps in creating multiple layouts of subplots. It provides control over all the individual plots that are created.

**Countplot:**

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for [barplot()](https://seaborn.pydata.org/generated/seaborn.barplot.html#seaborn.barplot), so you can compare counts across nested variables.

**Line plot:**

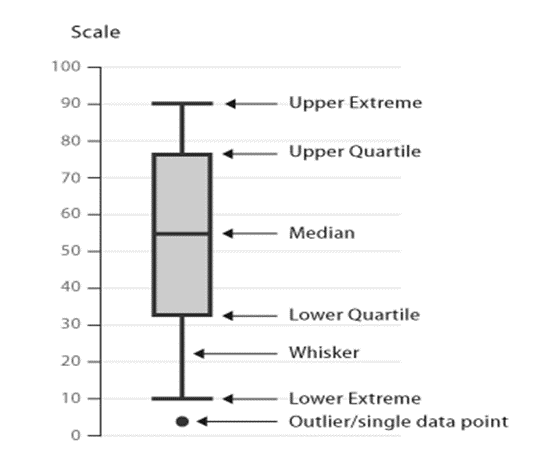
line plot () Draw a line plot with the possibility of several semantic groupings. The relationship between x and y can be shown for different subsets of the data using the hue, size, and style parameters. These parameters control what visual semantics are used to identify the different subsets.

**Heat map:**

A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually, the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used.

**Box and Whisker Plot (or Box Plot):**

It is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.



**We will be tackling this problem statement in five stages:**

1. Booking wise analysis.

* How many bookings cancelled each year?
* What is the booking difference between weekends and week days night?
* From which market segment bookings done the most?
* What is the booking percentage difference between city hotel and resort hotel?
* What are the bookings percentage each year?

1. Guest wise analysis.

* How many guests repeated each year?

1. Type of visitors.

* Which is the most booked accommodation type?
* From which country visitors comes the most?

1. Month wise analysis.

* Which is the most occupied month for hotels?
* What is the average daily rate for each month per hotel type?

1. Room wise analysis.

* Which room type have the most demand?
* How many rooms wrongly assigned to with respect to booked room type by each hotel?
* Which room type generates highest ADR?

**Summary**

BOOKING WISE ANALYSIS

1.   The total bookings got canceled 44199 i.e. 37% of total booking. Most of the booking cancelled for City Hotel during the year of 2016 and 2017 that is 61% of total booking canceled.

2.  The total bookings got canceled 44199 i.e. 37% of total booking. Most of the booking cancelled for City Hotel during the year of 2016 and 2017 that is 61% of total booking canceled

3. More than 60% of the population booked the City hotel

4. More bookings were made in 2016, compared to the previous year. But the bookings decreased by almost 14% the next year.

5. As we can see from analysis most bookings were made by online TA market segment. And the least bookings were made by aviation and complementary market segment.

GUEST WISE ANALYSIS

1. In resort hotel 7334 room assigned wrong i.e 25.4% of total reserved room type in resort hotel. In city hotel 6661 room assigned wrong i.e 14.5% of total reserved room type in city hotel.

2. The 280 guests were repeated in the year 2015 , 1619 guests were repeated in the year 2016  and 1306 guests were repeated in the year 2017.

TYPE OF VISITORS

1. Majority of visitors travel with family.The visitors with babies prefer mostly the resort hotel

2. The majority of the bookings are from Portugal. The countries UK,France,Spain and Germany also holds a great portion in bookings.The approx 70% comes from these 5 countries.

MONTH WISE ANALYSIS

1. In the month of August ,maximum number of bookings were made. Then the second number of majority of bookings were made in the month of July. In the month January there are least number of bookings.

2. The adr for City Hotel is highest for the months May and August, the adr for Resort Hotel is highest for the August month and the adr for City Hotel is more expensive than Resort Hotel for each month.

ROOM WISE ANALYSIS

1. The room type 'A' have the most demand and the adr is also high.

2. In resort hotel 7334 room assigned wrong i.e 25.4% of total reserved room type in resort hotel. In city hotel 6661 room assigned wrong i.e 14.5% of total reserved room type in city hotel.

3. The room type 'A' have the highest adr.

**Conclusion**

In summary, EDA is an essential step in many types of research but is of particular use when analysing hotel booking records. The tools described in this study should allow the researcher to better understand the features of a dataset and also to generate novel hypotheses.

**Take Home Messages**

1. Always start by exploring a dataset with an open mind for discovery.

2. EDA allows to better apprehend the features and possible issues of a dataset.

3. EDA is a key step in generating research hypothesis.

## Challenges

(1) Data was present in wrong data type format.

(2) Choosing appropriate visualization techniques to use was difficult.

(3) A lot of null values were there in the dataset.