**Hotel bookings data analysis**

* **Lalit Ahirrao**
* **AniketGajmal**
* **Rushikesh Pawar**
* **Prasad Ghegade**
* **Samarth Gangurde**

**Abstract :-**

Online hotel reservations are a popular method for booking [hotel](https://en.wikipedia.org/wiki/Hotel) rooms. Travelers can book rooms on a computer by using [online security](https://en.wikipedia.org/wiki/Transport_Layer_Security) to protect their privacy and financial information and by using several online [travel agents](https://en.wikipedia.org/wiki/Travel_agent) to compare prices and facilities at different hotels.Prior to the [Internet](https://en.wikipedia.org/wiki/Internet), travelers could write, telephone the hotel directly, or use a travel agent to make a reservation. Nowadays, online travel agents have pictures of hotels and rooms, information on prices and deals, and even information on local [resorts](https://en.wikipedia.org/wiki/Resorts). Many also allow [reviews](https://en.wikipedia.org/wiki/Review) of the traveler to be recorded with the online travel agent.Online hotel reservations are also helpful for making last minute travel arrangements. Hotels may drop the price of a room if some rooms are still available. There are several [websites](https://en.wikipedia.org/wiki/Websites) that specialize in searches for deals on rooms.

### **Reason behind the project :-**

Every year, more than 140 million bookings were made on the internet and many hotel bookings were booked through top-visited travel websites like Booking.com, Expedia.com, Hotels.com, etc. but Booking cancellations have a substantial impact on demand management decisions in the hospitality industry to avoid cancellations of bookings.To find out Methods they prefer for bookings,Months of the traffic, Days they will stay, meal they will prefer To know statistical measure of these segments we will deep dive into the Exploratory Data Analysis

### **Project Description :-**

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions! This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data. Explore and analyze the data to discover important factors that govern the bookings

## **Dataset Resources:-**

This project 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

* reasons for booking cancellations across various parameters
* best time to book hotel
* peak season

# **Data Description:-**

**Hotel**:Hotel (H1 = Resort Hotel or H2 = 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

**rrival\_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**:*BB – Bed & Breakfast*HB – only two meals including breakfast meal\* FB – breakfast, lunch, and dinner text

**country**:Country from they belong

**market\_segment**:*TA: Travel agents*TO: Tour operators

**is\_repeated\_guest**:is guest repeated

**previous\_cancellations**:Have they canceled previous bookings

**previous\_bookings\_not\_canceled**:

**reserved\_room\_type**:Type of room they reserved

**assigned\_room\_type**:Type of room they assigned

**booking\_changes**:booking changed or not

**deposit\_type**:deposit type they preferred

**agent**:was there any agent between for hotel recommendation

**company**:

**days\_in\_waiting\_list**:

**customer\_type**:

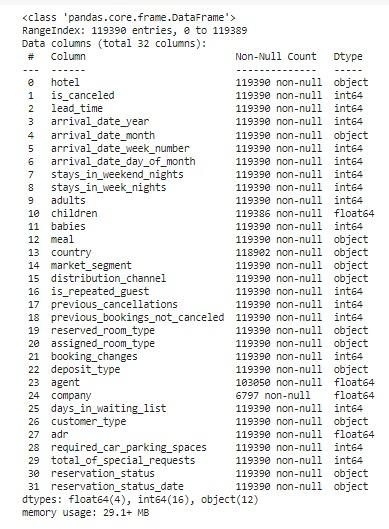
**adr**:Average daily rate

**required\_car\_parking\_spaces**:Spaces required for car parking

**total\_of\_special\_requests**:total number of special requests

**reservation\_status**:Status of reservation is it canceled,waiting,confirmed

**reservation\_status\_date**:date of reservation



## **PREPROCESSING OF DATASET:**

Preprocessing of Data can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance.as name suggest we prefer the preprocessing of data before training the model.Data preprocessing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set

**DATA CLEANING:**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

* · Missing Values
* · Incorrect Format
* · Incorrect Headers
* · Anomalies/Outliers

**DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to

further improve the efficiency and accuracy of data mining.

**DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

**MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

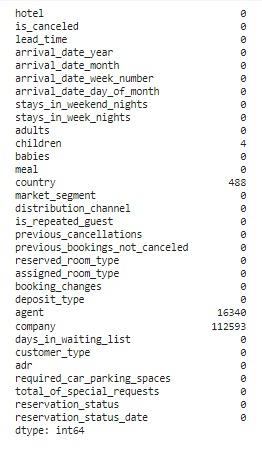
If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.

2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.



**DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the isnull() and notnull() functions from the pandas library to determine null values

**EXPLORATORY DATA ANALYSIS:**

EDA means trying to understand the given data much better, so that we can make some sense out of it. Using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

**UNIVARIATE ANALYSIS:-**

If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis looks at one feature at a time. When we analyze a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

Univariate analysis is the simplest form of analyzing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

**BIVARIATE ANALYSIS:-**

If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

**Numeric Analysis:-**

Analyzing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyze it in three different ways.

· Scatter Plot

· Pair Plot

· Correlation Matrix

**Categorical Analysis:-**

Analyzing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyze those mainly using mean, median, and box plots.

**MULTIVARIATE ANALYSIS:-**

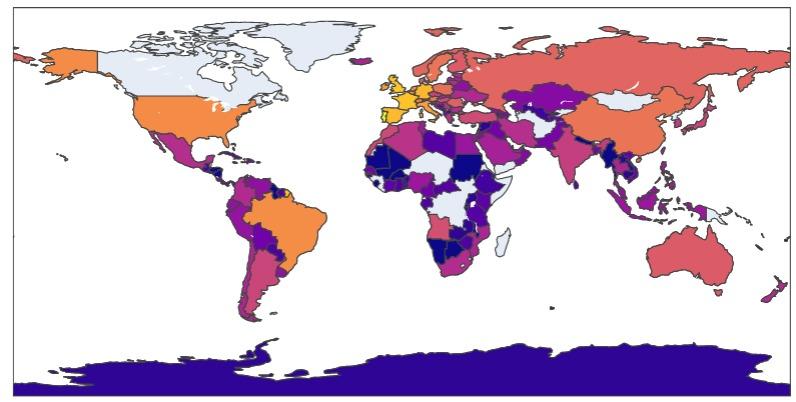
Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behavior more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

**Choropleth Map**

A Choropleth Map is a map composed of colored polygons. It is used to represent spatial variations of a quantity.

Making choropleth maps requires two main types of input:Geometry information:This can either be a supplied GeoJSON file where each feature has either an id field or some identifying value in properties; orone of the built-in geometries within plotly: US states and world countries (see below)A list of values indexed by feature identifier.

****

**REFERENCES:**

* [Predicting a Hotel Booking Demand | by Dimas Adnan | Towards Data Science](https://towardsdatascience.com/predicting-a-hotel-booking-demand-7608a7dbf5a4)
* [Exploratory Data Analysis (EDA) in Python | by Atanu Dan | Medium](https://medium.com/@atanudan/exploratory-data-analysis-eda-in-python-893f963cc0c0)
* [Choropleth maps with Python (plotly.com)](https://plotly.com/python/choropleth-maps/)