Hotel Booking Analysis

Sahil pardeshi, pravin bejjo and kirtesh verma
Data science trainee, almabetter nashik

Introduction:

The hotel industry is the section of the service industry that deals with guests accommodations or lodgings.By most definitions, the hotels industry refers not only to hotels,but also to many forms of overnight accommodations,including hostels,motels,inns and guest houses. The data contain “booking due to arrivals between the 1st of july 2015 and the 31st August of 2017’’.This data contain booking information for a city hotel and resort hotel,and i includes information such as when the booking was made,length of stay, number of adults,children,and/or babies and the number of available parking spaces among other things.

The hospitality industries are the key drivers of growth and development. There are many sectors of hospitality industries like accommodation,food and beverages,tourism,events,tourist attractions,and recreations.

1.problem statements.

Our main objective of this EDA project is to appreciate the features that play an important role in deciding the booking factors of cities, hotels and resort types. We prepared important questions like What is the ADR (average daily rate)of both the hotel types,was there any deposit before the booking,what type of booking is preferred by customer(online or offline),from which countries most guests come.

2.Methodology.

Exploratory data analysis(EDA)
EDA involves generating summary statistics for numericals data in the datasets and creating various graphical representations to understand the data better.
It is a process of investigating the datasets to discover patterns, and anomalies,and form a hypothesis based on understanding of the datasets.
In statistics, EDA is an approach of analysis sets to summarize their main characteristics often using statistical graphs and other visualization methods.

3.Data understanding.

The datasets contain the booking information of both the hotels. There are 119390 entries and 32 columns present in the datasets.The hotel bookings are seen mainly in 1july of 2015 and the 31st august 2017 with the customer effectively arriving and canceling booking.

Resort Hotel:
The resort hotel is the luxury facilities that is intended primarily for vacationaly and is usually located near special attractions. Such as in hill stations, beaches near oceans, pilgrimages, and other regions, etc.

City Hotel: The city hotel is located mainly in cities.it provides meals and various facilities to travelers or guests.

It is always located near the railways station, bus station, airports etcs.

Columns presents in the datasets

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Hotels:

H1 is the resort hotel

H2 is the city hotel

4.Data wrangling.

Data wrangling is the process of cleaning the datasets from their null values and duplicates values. The EDA performed on the clean datasets results in better visualizations of different features and the data interpretation is more accurate.

a) **Loading the datasets :**

The load method provides a techniques for filling a single datatable with data,retrieved from an IdataReader instance.this method provides same functionality but allow us to load multiple data sets
We are using google colab which allows the user to write and execute the arbitrary python codes through the browser and it's well suited for data analysis and ML(machine learning).
The hotel booking dataset is provided by Almabetter.

Libraries we used:

For loading and visualization we have used the following libraries like numpy, pandas,matplotlib, seaborn.

1)Numpy: numpy is the python library used for programming languages,adding support for large,multidimensional arrays and matrices with large collections.

2) **Pandas:**

It is a software library written for python programming,flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive.pandas allow us to access many of matplotlibs and numpy's methods with fewer codes.

3)Matplotlib:

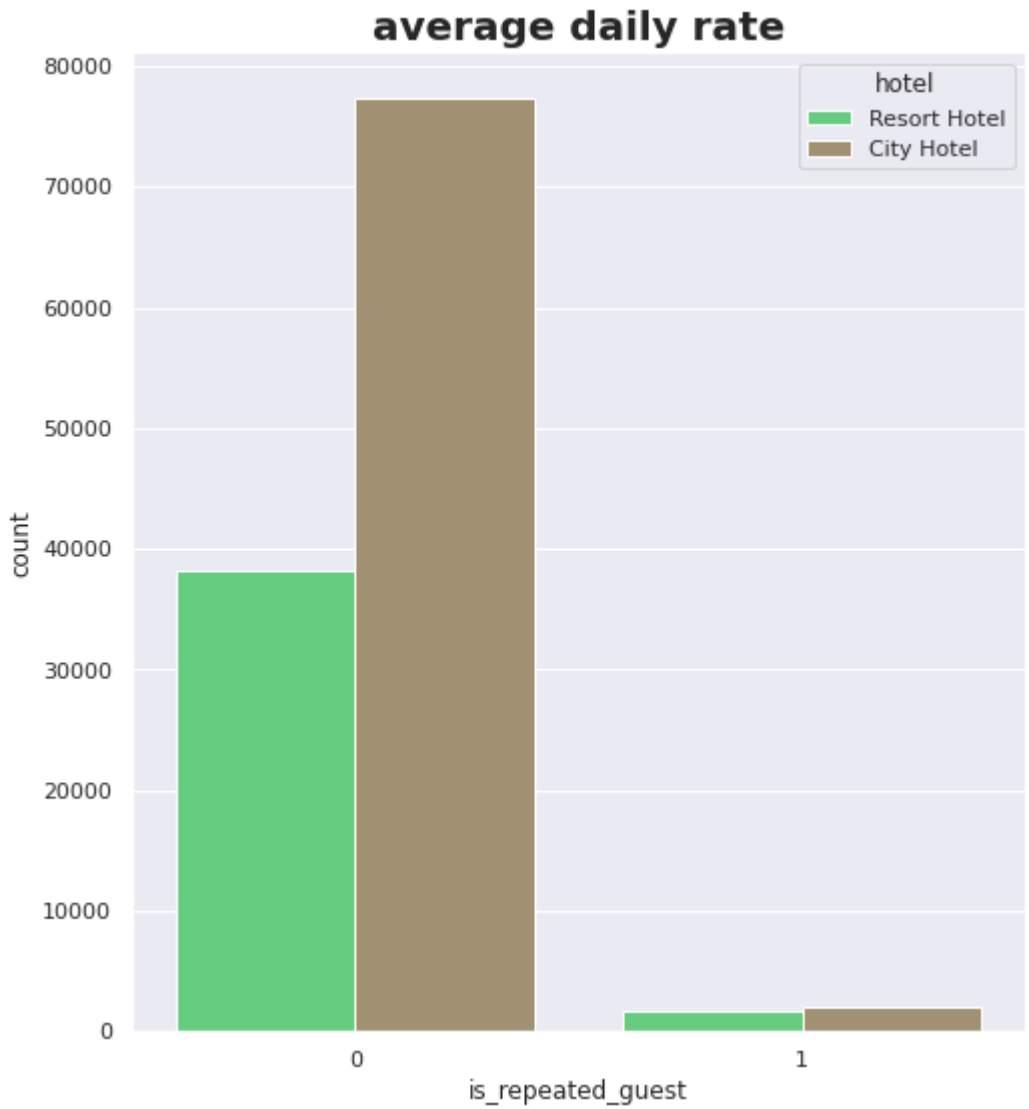
It is a pyplot collection of functions that make matplotlib work like MATLAB.eg.creates a figure,lines a plotting area.

4)Seaborn: it is an open source python library built on top of matplotlib. It is used for data visualization and EDA. seaborn works easily with dataframes and pandas librariesThe graphs can also be customize.it is used in ML

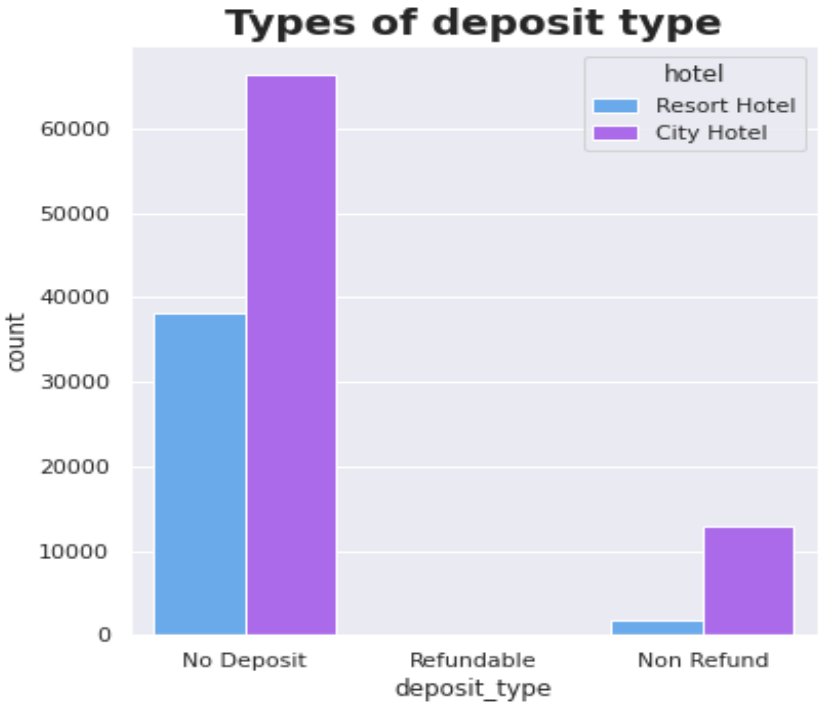
5. Data visualizations:

Once we clear data then our next step is to visualize the data for a clear understanding of different features in graphs .

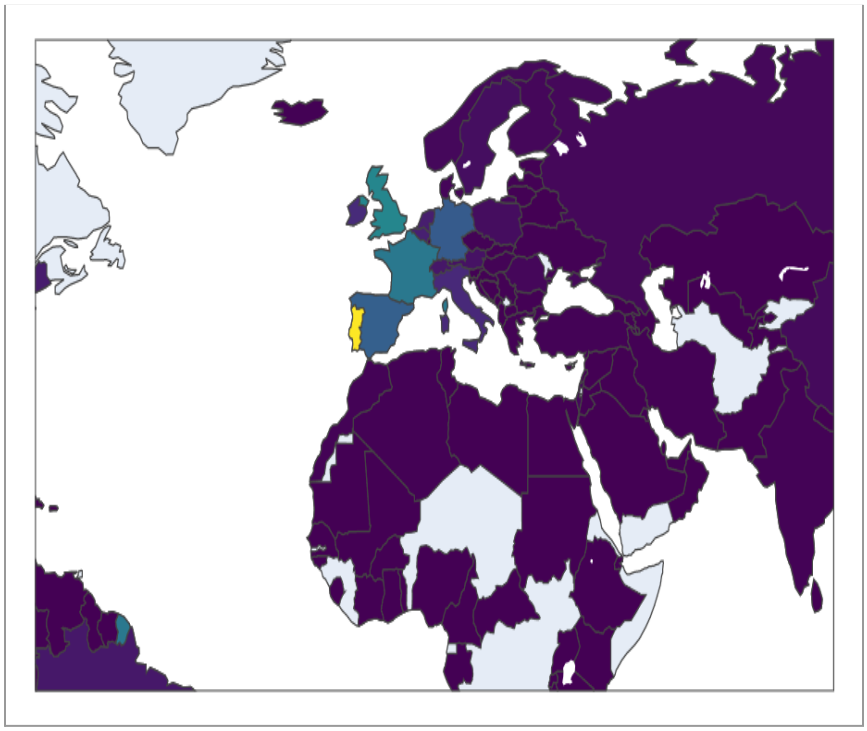
1.Average daily rates.



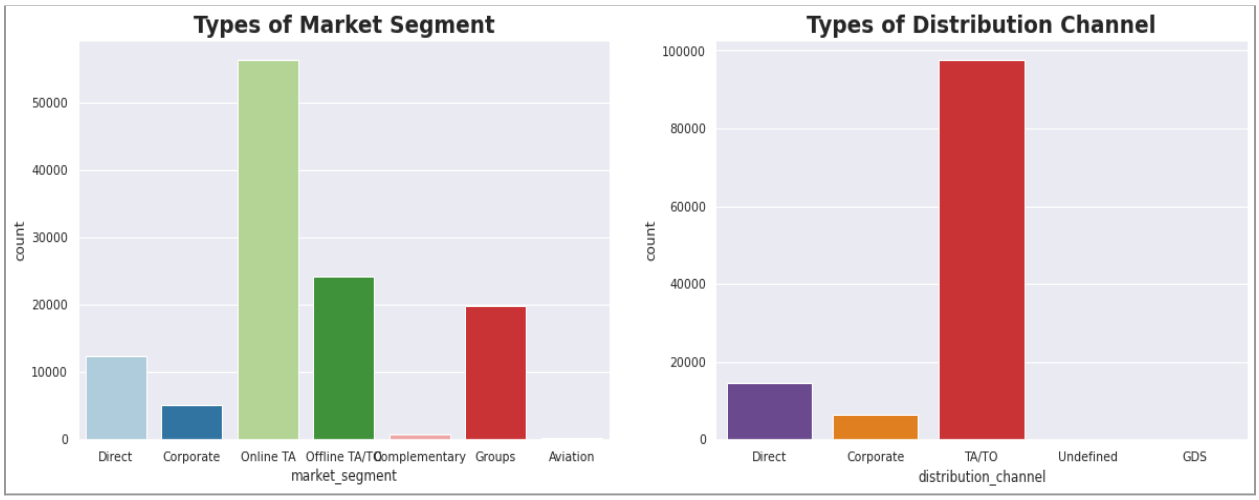
2)deposit type.



3) From which countries most guests come.



4)what type of booking is most preferred by customers.



6) Challenges Faced:

Null(nan) values:

Nan stands for Not A number is one of the common ways to represent the missing value in the data.In our interpretations.there are four columns in our datasets that contain the null values. They are agents, company, country and children which were filled with 0.Nan values in textual fields like company for instance filled with unknown values.Nan values remaining was some problematic.

Size:

When we deals Big datasets, it leads to problems like poor data quality,solving the wrong problems, inability to operate insights etc . There are 32 variables present in the given datasets and finding the most important factors and reference one was difficult.

7.conclusions:

After performing the EDA on the given datasets of hotel booking. We have understood the given factors that govern the booking**.**

We have taken the few conclusions points they are

a) **The Number of market segmentation is online or offline booking preferred by customers.**
b) **The Average daily rates are lower during winter and higher during summer season.**
c) **No prerequisites of deposit types lead to high cancellation rate.**
d) **There is no deposit before booking a hotel.**
e) **Most of the guests come from western European countries like Portugal,Spain,UK,france**
f) **Most famous meal is the BB type.**