HOTEL BOOKING ANALYSIS

Kirtesh Verma, Pravin Bejjo and

Sahil Pardeshi

Data Science Trainees,

Almabetter, Nasik

- ## Introduction

Have you ever wondered the trends for hotel bookings? How long a person stays? How often people cancel? What the busiest months are? In this analysis I explore a large dataset to examine these questions.

The data contains “booking due to arrive between the 1st of July of 2015 and the 31st of august 2017”. This dataset contains information on records for client stays at hotels. More specifically, it 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. For the purpose of this post, I only focused on some of the important variables to examine.

I. Problem Statement

The main objective of this project is to explore and visualize the dataset from hotel booking data using basic exploratory data analysis techniques. This will be done by finding out the information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of repeated guests, among other things.

❖ Understanding the data:

The dataset has around 119390 observations in it with 32 columns and it is mix between categorical and numerical values

```
<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
```

Columns used in the analysis:

- **Hotel**
  - Resort hotel

  - City hotel
- **is_canceled**
  - 1: Cancelled
  - 0: Not cancelled
- **lead_time**
  - No of days that elapsed between entering date of booking into property management system and arrival date
- **arrival_date_year**
  - Year of arrival date (2015-2017)
- **arrival_date_month**
  - Month of arrival date (Jan - Dec)

- **arrival_date_week_numberr**
  - Week number of year for arrival date (1-53)
- **arrival_date_day_of_month**
  - Day of arrival date
- **stays_in_weekend_nights**
  - No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel
- **stays_in_week_nights**
  - No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel
- **Adults**
- **Children**
- **Babies**
- **meal**
  - Type of meal booked. 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**
- **market_segment** (a group of people who share one or more common characteristics, lumped together for marketing purposes)
  - TA: Travel agents
  - TO: Tour operators
- **distribution_channel** (A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)
  - TA: Travel agents
  - TO: Tour operators
- **is_repeated_guest** (value indicating if the booking name was from repeated guest)
  - 1: Yes
  - 0: No

- **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.
- **day_in_waiting_list**
  - Number of days the booking was in the waiting list before it was confirmed to the customer.

- **customer_type:**
  - 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)**
- **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**
    - Cancelled – booking was cancelled 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.

II. Steps involved

1) Loading Data

For this project, we are using Google colab notebook IDE with a python programming language to write our script.

To get the data, we are using hotel booking data that shared by almabetter for this project.

Before loading the data into IDE, first we need to import various external libraries/packages that are needed for visualization and analysis.

a. **Load python libraries**
   - **Pandas and numpy library are used for data analysis**
   - **Matplotlib, seaborn and plotly library used for data visualization.**

b. **Load dataset**

   To load the dataset, we first need to mount google drive. Next we use pandas library and function to read the CSV file pd.read_csv(file_path)

2) **Cleaning dataset**

   The next step is cleaning up the data, oftentimes the data that we load have various faults, such as duplicates, missing values, incomplete data, etc. by doing clean up, the data quality will have better quality to be used for further analysis.

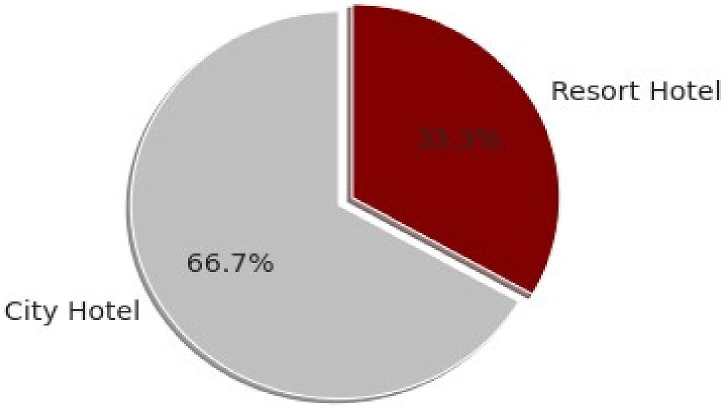
   a. **Removing duplicates if any**
   b. **Dropping null observations**

3) **Analyzing and Visualizing the Data**

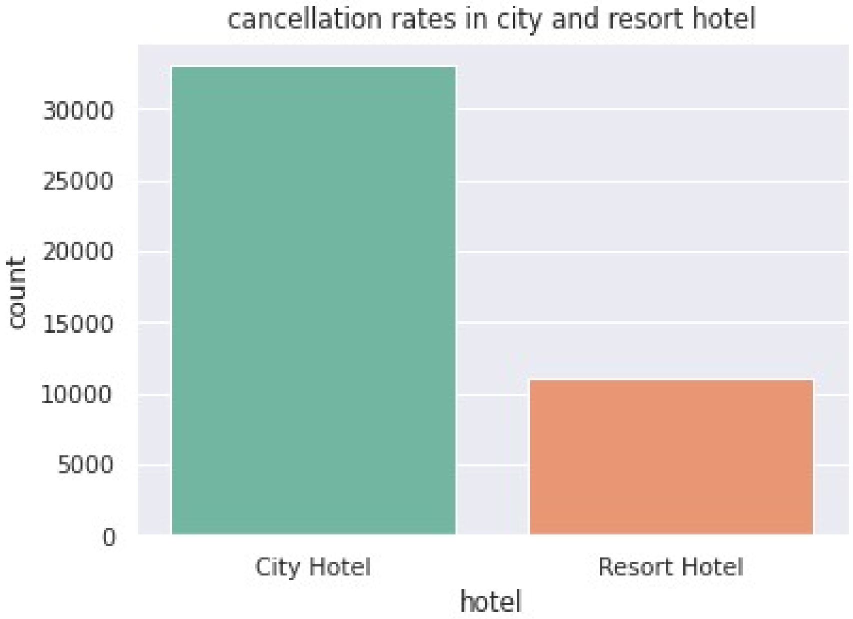
   After we clean up the data, the next step is the exploring the data by visualizing and analyzing the values of the features, explaining the process and the result.

   Some of the highlights from the analysis are as follows:

a. **City vs Resort Hotel:**

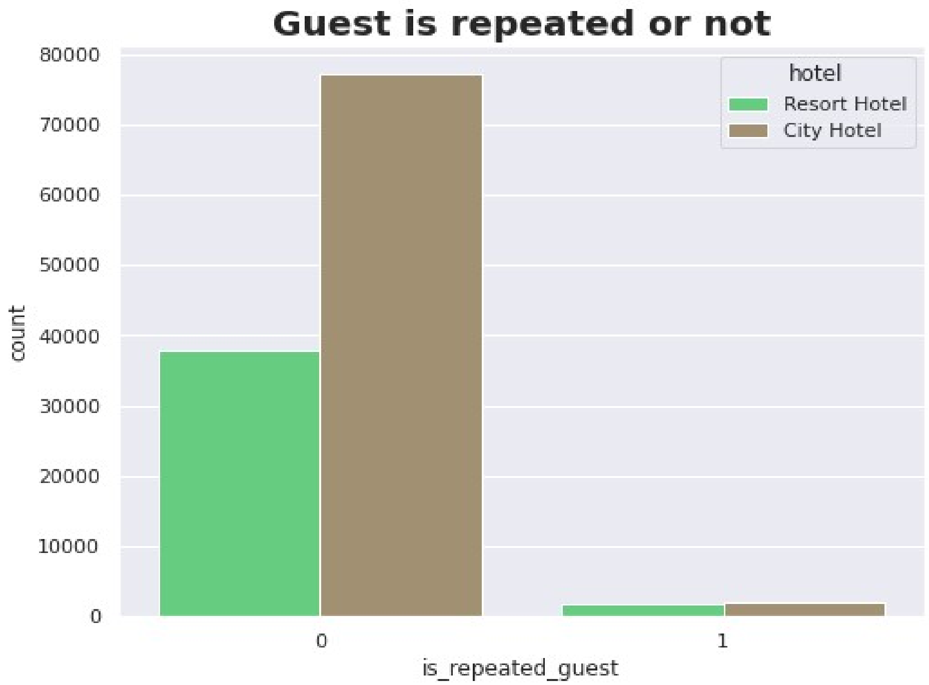


b. Cancellation Rate:

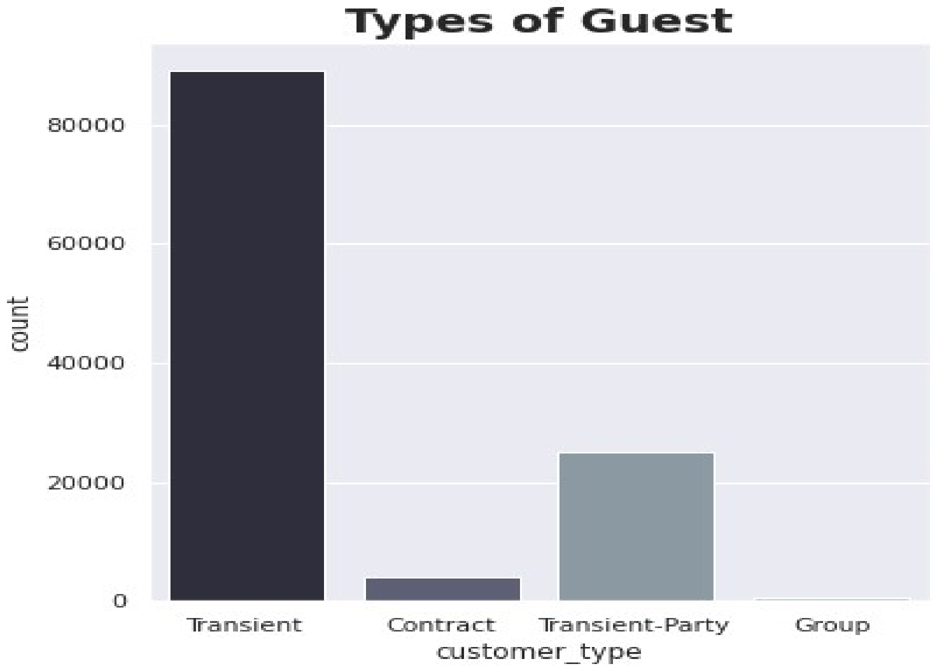


c. Type of guest:

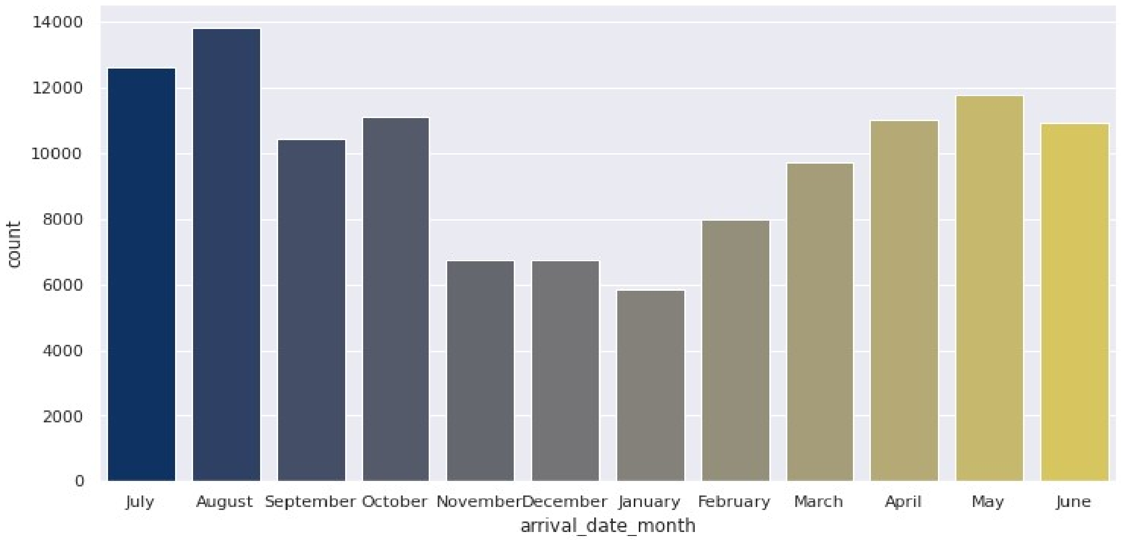
- Repeated or Not



- Transient or other type



d. Busiest Months:



❖ Challenges Faced:

- **Size:** Dealing with big dataset is always a problem. The bigger the dataset the more are the variables and finding out the most relevant one becomes difficult. The hotel booking dataset has 32 features and finding the most relevant one was difficult.

- **Null (NaN) values:** Data visualization done on data with null values is never clean. For a more accurate interpretation of the features and their relationship with each other we have to deal with null values. The given dataset has four columns which have null values. They are agent, company, country and children which were filled with 0. Removing the null value was a bit tricky.

❖ Conclusion:

After performing the EDA on the given dataset of hotel booking, we understood the important factors that govern the hotel booking. This will help the decision makers to plan accordingly for improving the business. The insights that we drew are as follows:

- Majority of people preferred city hotel over resort hotel. City hotel is cheaper than resort hotel so most of the people are like to go in city hotel rather than go to resort hotel.
- City hotel have the highest cancellation rates as compared to resort hotel, this can be verified by the fact that city hotels have higher booking rate than resort hotel.
- Transient, contract, transient-party, and group are the types of guests. Majority of the booking are the transient. This means that the booking is not part of a group or contract. With the ease of booking directly from the website, most people tend to skip the middleman to ensure quick response from their booking.
- Increasing in booking on august being is highest. Summer ends around august; it seems that summer period is a peak period for hotel booking.
- 2016 showed the highest rate of hotel bookings. (data from 2015-2017)

- Majority of guests are from Western Europe. So target this area for more customers.
- The majority of reservations convert into successful transactions.
- More bookings occurred on weekdays vs. weekends.