# Capstone Project

## Hotel Booking Analysis

EDA by Robin Dubey

# Skepticism in Planning Trips

Vacations, business stay or a casual trip to a new city, hotel bookings are mandatory and each one of us desires to optimize our stay. Optimization, for someone could be booking a hotel where he/she pays less for a good deal, or maybe just getting the only luxurious suite, in a 7-star hotel during a non-peak period.

# Continued . . .

As a customer, I would like to know more details about each of these hotels before the actual booking, like which one is better in terms of long stay-backs, what time of year would be best for me to visit if I am aiming at saving a few bucks, for how long should I stay at a particular hotel to get an optimal price, which one of them is couple-friendly, family-friendly or both.

I might also want to know, if someone from my home country had already stayed in that hotel or not, this builds a sense of confidence and safety visiting a new place.

# Optimizing Our Hotel Booking

- **Step-1**
  Variables Understanding
- **Step-2**
  Data Wrangling
- **Step-3**
  Analyzing Through Data Manipulation & Visualization Tools
- **Deriving Conclusions**

# DATA VARIABLES

- **hotel –** Two hotels are given: Resort Hotel, City Hotel
- **is_canceled-** 1: Canceled, 0: Not canceled
- **lead_time -** gap between booking and arrival
- **arrival_date_year -** arrival year
- **arrival_date_month -** arrival month
- **arrival_date_week_number -** arrival week
- **arrival_date_day_of_month-** arrival date
- **stays_in_weekend_nights -** count of nights the guests booked the hotel during Sat-Sun
- **stays_in_week_nights -** count of nights the guests booked the hotel during Mon-Fri
- **adults -** count of adults
- **children-** count of children
- **babies-** count of babies
- **meal-** meal type (no meal package; BB; HB; FB)
- **country-** country of guests
- **market_segment -** TA: Travel agents, TO: Tour operators
- **distribution_channel**
- **is_repeated_guest-** 1: Yes, 0: No

# DATA VARIABLES (Continued . . .)

- **previous_cancellations-** count of previous bookings that were cancelled by the customer before final booking
- **previous_bookings_not_canceled-** count of no canceled bookings
- **reserved_room_type-** booked room category
- **assigned_room_type-** assigned room category
- **booking_changes-** count of changes made by the customer before final booking
- **deposit_type-** type of deposit made by the customer
- **agent-** travel agent id
- **company-** booking company id
- **days_in_waiting_list-** count of days the booking was in the waiting list before it was confirmed
- **customer_type-** Transient, Contract, Group, Transient-party
- **adr -** average daily rate for the booking
- **required_car_parking_spaces-** count of car parking spaces allotted by the customer
- **total_of_special_requests-** count of special requests made by the customer
- **reservation_status-** status of reservation
- **reservation_status_date-** date corresponding to status of reservation
- **total_stay_nights -** duration of stay including weekend nights and week nights stay
- **price-** total price spent by a guest entity
- **df_country_guests_top10-** Top 10 countries with max visitors

# STEP-1

## Defining New Variables

A few variables are a result of operations performed between 2 or more variables. For example: *distance = (speed ). (time)* here, distance is a new variable which stores a multiplication of speed and time.

**Example:** We are given *average daily rate (adr)* as one variable but the *adr* is defined as an aggregate value but, if we know the adr value per person, then this will help us in getting insight into ‘how much on an average a person needs to spend for daily stay in these hotels?’.

Defining adr per person as *adr_pp*, which stores information of *adr* divided by total count of guests.

# Continued . . .

Similarly, let's define a variable called *total_stay_nights*, which stores the total stay duration of a guest entity, as summation of stays_in_weekend_nights and stays_in_week_nights.

## Picking-Up Right Variables

We need to start with picking-up the right variables from the bucket of data variables, this will actually make our analysis simple and understandable.

Let's know the price that has been spent by a Guest entity* then we need :

- *adr_pp* : average daily rate (adr) value per person
- *adults+children*: number of person in the guest entity **
- *total_stay_nights*: length of stay (stays_in_weekend_nights + stays_in_week_nights)

Total Price Paid by the Guest entity = (adr_ pp). (adults+children). (total_stay_nights)

# STEP-2

Data Wrangling or Data Cleaning is one of the crucial steps while analyzing any dataset. It is generally performed beforehand, to get quality insights. What we generally do in Data Cleaning? Let's Know!

## Null/Missing Value Treatment

In the real-world scenarios, data always comes with imperfection or partialness. These are called Null-values in the data. Treating them well should be our first priority, before deploying any operation on them.

These null values can be treated by either dropping all the null value rows or filling them with the average or most preferably with the mode value of the column. We will prefer the later process by using *fill.na( )* command.

# Continued . . .

### *Removing Unwanted Columns*

**During Picking-Up Right Variables (Refer 2.2), we can decide to drop a column which is not a part of our problem statement. This is generally done by subsetting a dataframe, under the cases of data with billions and millions of rows and thousands of columns.**

**In this case, we are working on a considerably small dataset with just 32 columns, so we will skip dropping any column.**

# STEP-3

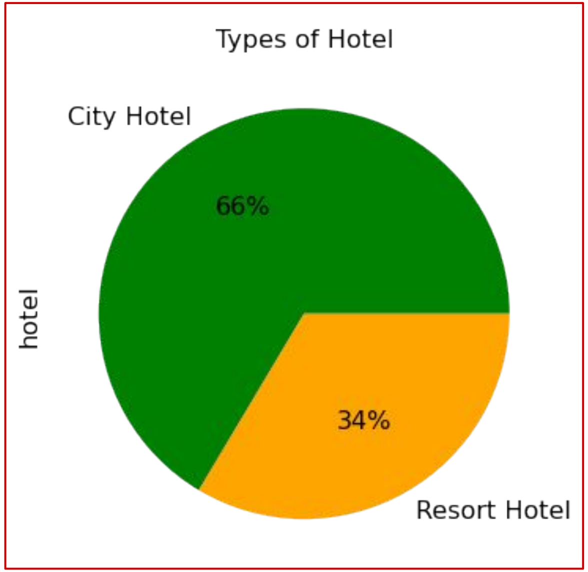
**Data Manipulation** is the modification of information to make it easier to read or more structured.

**Data visualization** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools ..

## Primary Data Visual Comparison Using Matplotlib

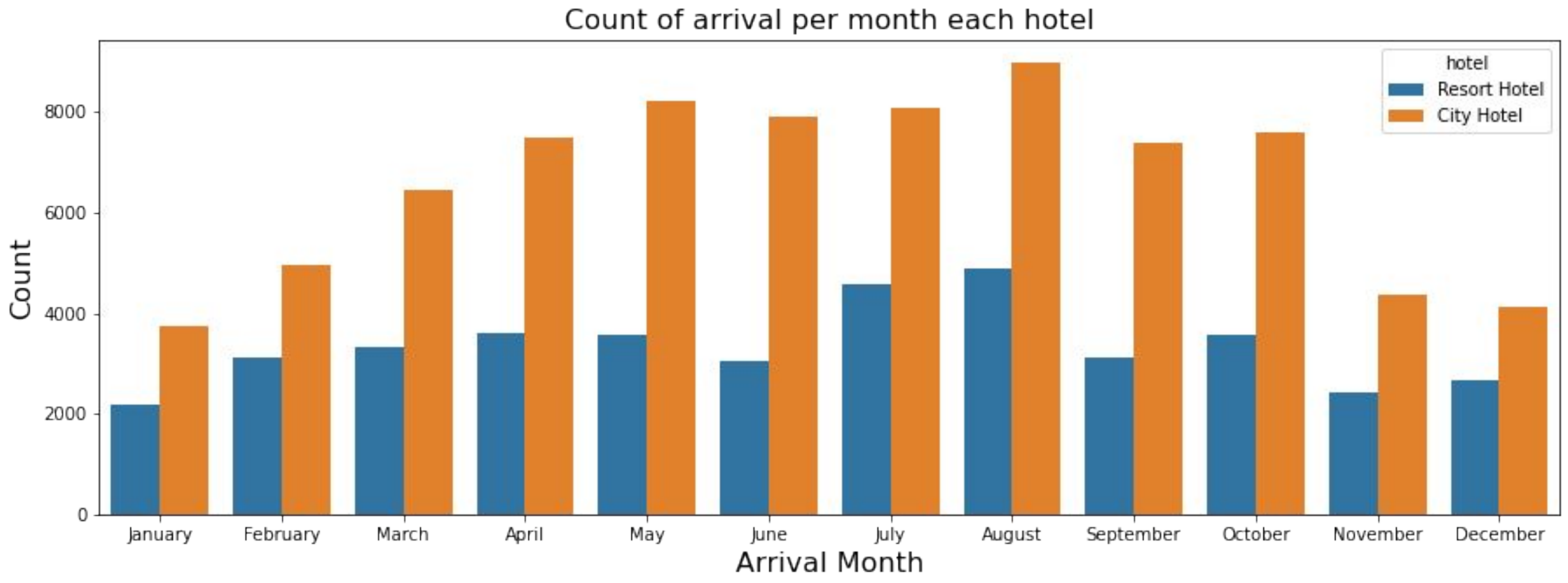
We already know that we are dealing with a 2-Hotel data comparison but, before moving ahead, we would have to see the shares of each *hotel* in the analysis.Let's use the Data Visualization tool i.e. matplotlib (alternatively, seaborn can also be used), for plotting the shares of each hotel data.

# TYPES OF HOTEL PLOT



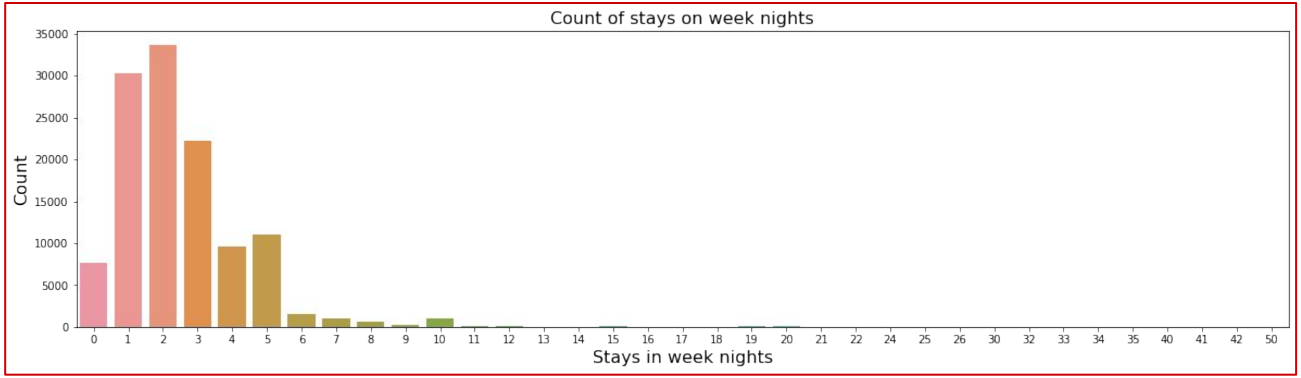
**i.e. the data we will be dealing with has 66 percent share of  City Hotel and the rest is of the Resort Hotel.**

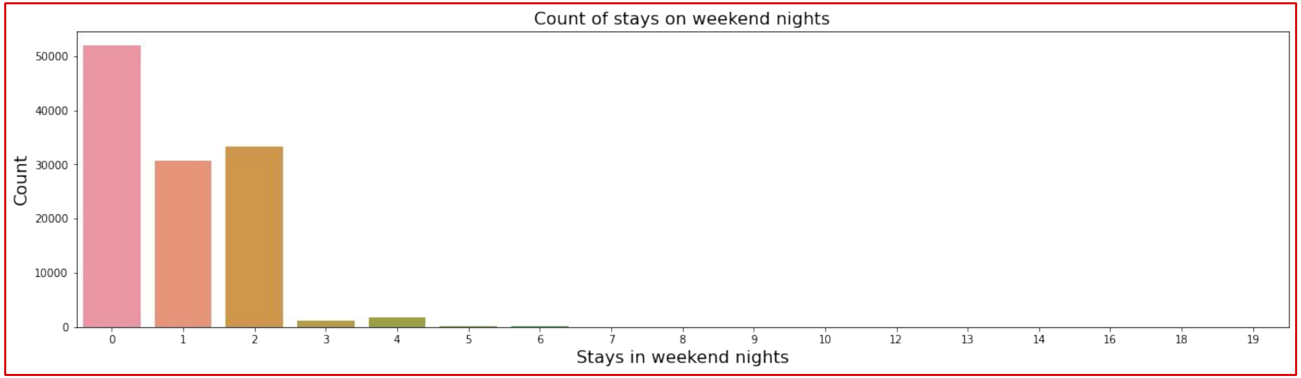
# FABRICATING INSIGHTS



**Analyzing Top Arrival Month**

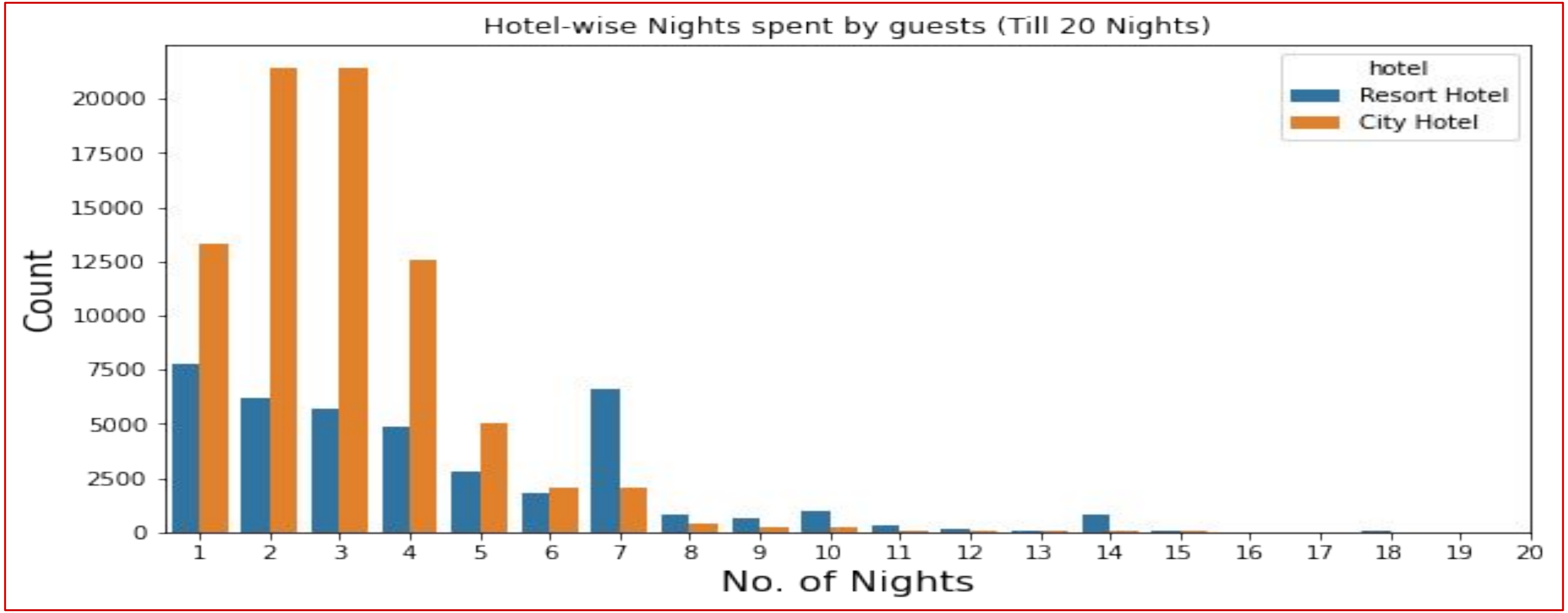
# Continued . . .





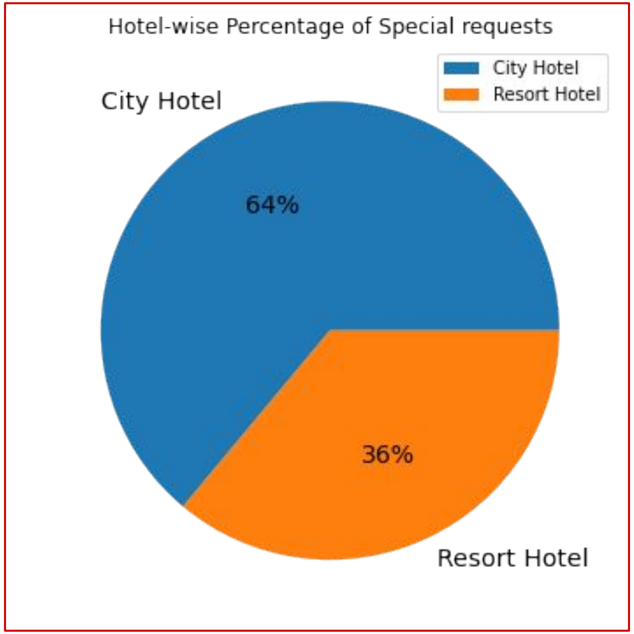
## Analyzing Arrival Weekend Nights/Week Nights

# Continued . . .



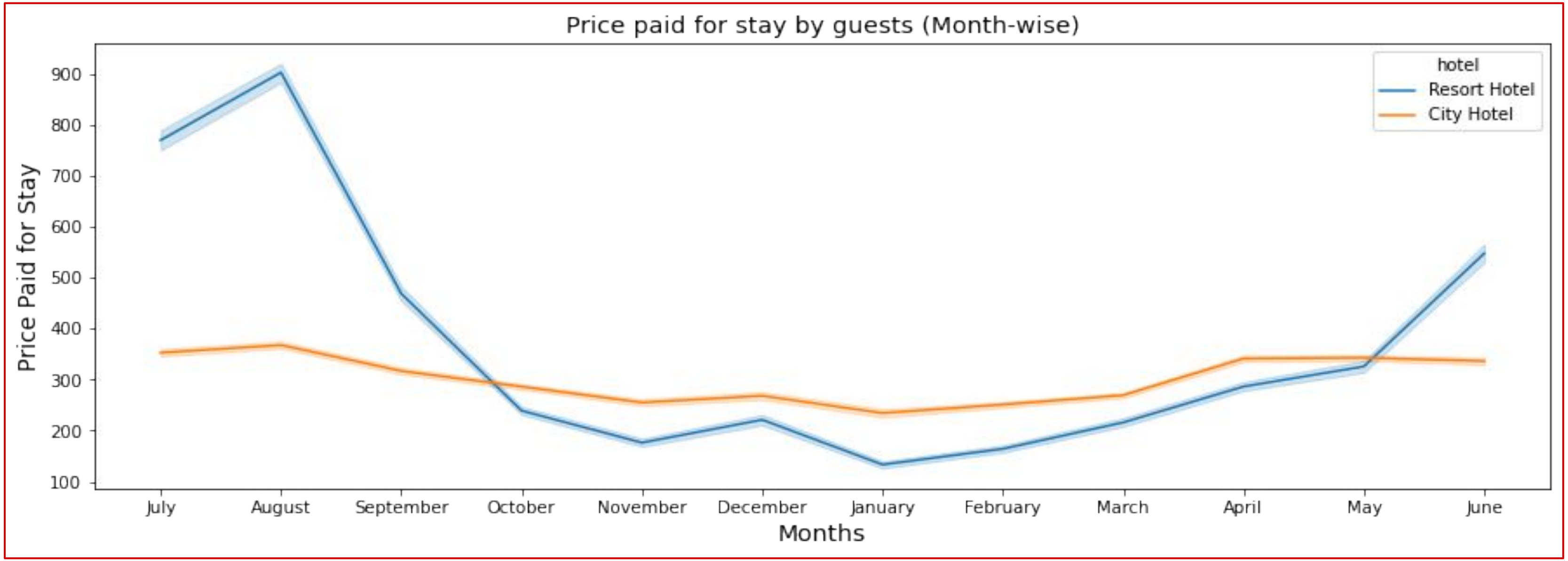
**Analyzing Hotel-wise Nights Spent by guests**

# Continued . . .



Analyzing Hotel-wise Special Request made by Guests

# Continued . . .



## Analyzing Price Paid (Month-wise) by the Guest Entity

# Deriving Conclusions

Hotel Bookings are variable dependent and dynamic in nature. Suppose, if we either add a few thousand data rows in this dataset, the values of each variable will migrate and take a different value or if we remove a few data rows, the same can be observed.

In our case, there are a few conclusions as per the problem statement that could be made.

- The City Hotel is more booked than the Resort Hotel.
- Month of August sees the maximum arrivals for either of the Hotels.
- Guests stay more on week nights than weekend nights.
- For City Hotel, most of the guests preferred staying for an optimal duration of  2 or 3 nights.
- For Resort Hotel, most of the guests preferred staying for an optimal duration of 1 or 7 nights.

# Deriving Conclusions (Continued . . .)

- Guests in the City Hotel make more requests than the guests in the Resort Hotel. Hence, City hotel management should expect having a special request, in every almost 2 bookings out of 3 bookings.
- If you are planning a vacation at City Hotel, you can target June, July, August and September months for a low-cost booking compared to Resort Hotel.
- If you are planning a vacation at the Resort Hotel, you can target months from January to May & October to December for a low-cost booking compared to the City Hotel.
- As, month of August sees maximum footfalls, it also experiences more money spent by a guest entity (or price to the hotel).
- Top 5 countries with the most visitors are with country code PRT, GBR, FRA, ESP, DEU.

# Challenges Faced

- Data Wrangling process was the major challenge for me i.e.  in the step of cleaning the data or in treating the NaN (null) values such that the data doesn’t lose its meaning. Firstly, I dropped all the null values and later on realised that this might have impacted the dataset and yes, I was right.

- Identifying the type of visualisation to be used and making a choice between Matplotlib and Seaborn.

# THANK YOU

Have a great day!