**Hotel Booking Analysis**

Saksham Tripathi

Bindu Kovvada

Shivangi Mishra

Deepak Kumar Gautam

Satyajit Sahoo

Data Science Trainees,

AlmaBetter, Bangalore

**Abstract:**

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more. This makes analysing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the

business.

We will be using the data available to analyse the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings. We will be tackling this problem statement in three stages: We will be analysing some key metrics for hotel bookings

like:

**Types of Hotels**

* Resort Hotel
* City Hotel

The database divided into two types of hotels "city" hotels and "Resort" hotels.

**1.Problem Statement**

* Percentage of bookings in each hotel type?
* From which country most guests come?
* Which is the busiest month for hotels?
* Which room type is in most demand ?
* Which room type generate highest average daily rate?
* Which meal type is most preferred meal of

customers?

* How many booking were cancelled?
* Booking cancellation and Repeated guest
* Now we will check whether not getting allotted the same room type as demanded is the cause of cancellation of bookings?
* Let’s see does not getting same room affects the average daily rate?
* Distribution channel and the Days on waiting list
* Which distribution channel has highest cancellation percentage?

Using the results from the above questions,

business can make key decisions regarding

the customer experience they desire to deliver.

**2. Introduction**

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. From it, we can understand the customer’s’ behaviour and it might help us make better decisions.

**Data set name:**

Hotel Booking Database that contains booking information for a city hotel and a resort hotel of various countries from 2015 to 2017.

**Data Shape:**

Rows - 119390

Columns - 32

And the Columns are:

**hotel**: Resort Hotel or 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 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

**arrival\_date\_day\_of\_month**: Day of arrival date

**stays \_in\_weekend\_nights**: Number of weekend nights

**stays\_in\_week\_nights**: Number of week nights.

**adults**: Number of adults

**children**: Number of children

**babies**: Number of babies

**meal**: Type of meal booked.

**country**: Country of origin

**market\_segment**: Market segment designation (TA/TO)

**distribution\_channel**: Booking distribution channel. (TA/TO)

**is\_repeated\_guest**: is a repeated guest (1) or not (0)

**previous\_cancellations**: Number of previous booking that were cancelled by the customer prior to the current booking

**reserved\_room\_type**: Code of room type reserved.

**assigned\_room\_type**: Code for the type of room assigned to the booking.

**booking changes**: Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.

**deposit\_type**: No Deposit, Non-Refund, Refundable.

**agent**: Id of the travel agency that made the booking

**company**: ID of the company/entity that made the booking.

**days\_in\_waiting\_list**: Number of days the booking was in the waiting list before it was confirmed to the customer.

**customer\_type**: type of customer. Contract, group, transient, Transient party.

**adr**: Average Daily Rate as defined by dividing the sum of all lodging transactions by total number of staying nights

**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.

**reservation\_status**: Reservation last status

We will perform exploratory data analysis with python to get insight from the data.

4. **Exploratory Data Analysis:**

Before exploring the dataset, we load the libraries and packages we are going to use as the usual aliases.

● import pandas as pd

● import numpy as np

● import matplotlib.pyplot as plt

● import seaborn as sns

By seeing heatmap we are not able to find any highly correlated columns. Still we are able to find out two results from the heatmap:

1.Lead time and total stay have slight correlation. This means when people want to stay little longer they plan little before than actual arrival.

2.Total guests and average daily rate(adr) has some correlation. This means the when the number of guests increases, adr will also increase.

****

**Cleaning the data:**

1) Removing duplicates.

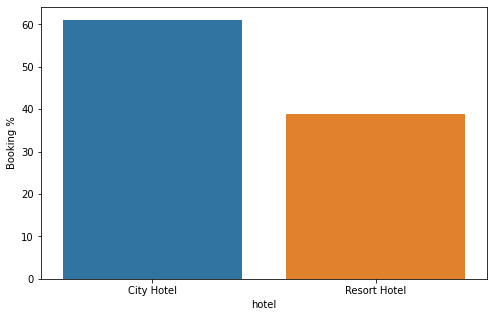
2) Handling missing values:

we are replacing missing values in the agent and company columns with zero. We are also replacing null values for children with mean values and country null values with ‘unknown’.

3) Convert columns to appropriate datatypes.

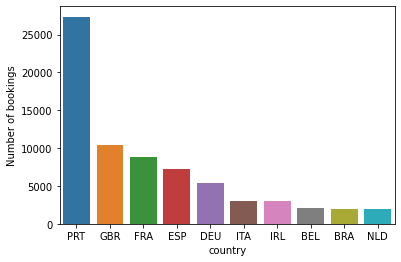
**5. MAJOR OBSERVATIONS FROM EDA**

1. Bookings were more for the City hotel than the Resort hotel.

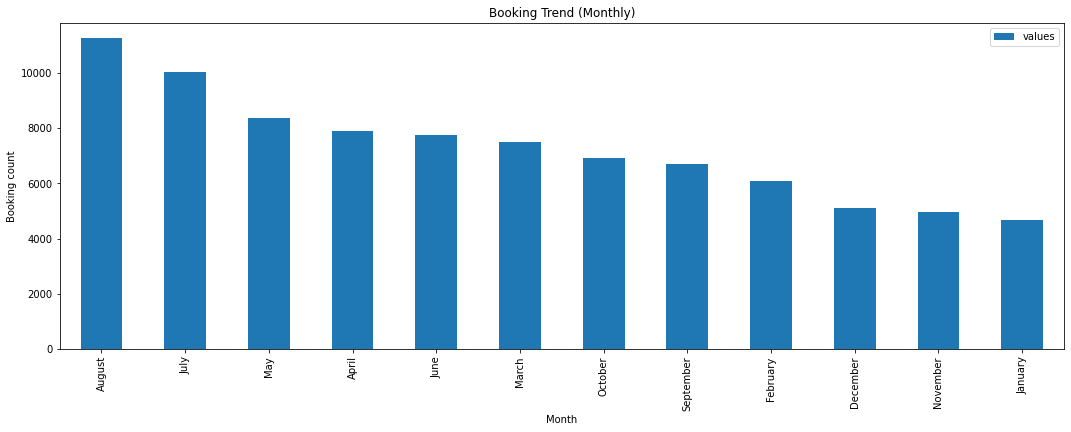


Around 60% bookings are for City Hotel and 40% are for Resort Hotel.

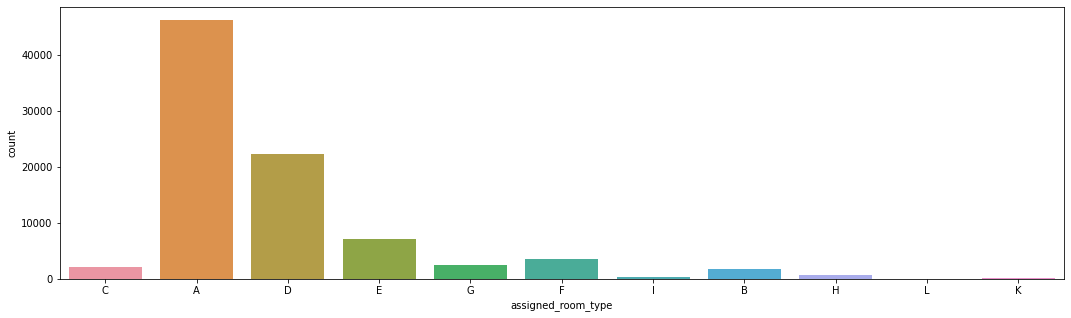
2. As the hotels are in Portugal Europe, the bookings are mostly with European countries, Highest is Portugal with 48.59k bookings.



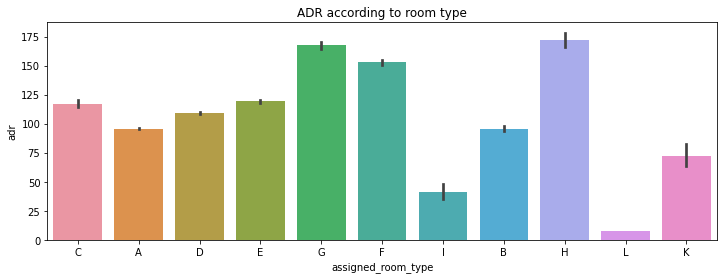
3. Number of bookings made were highest in the month of July and August and lowest in January.



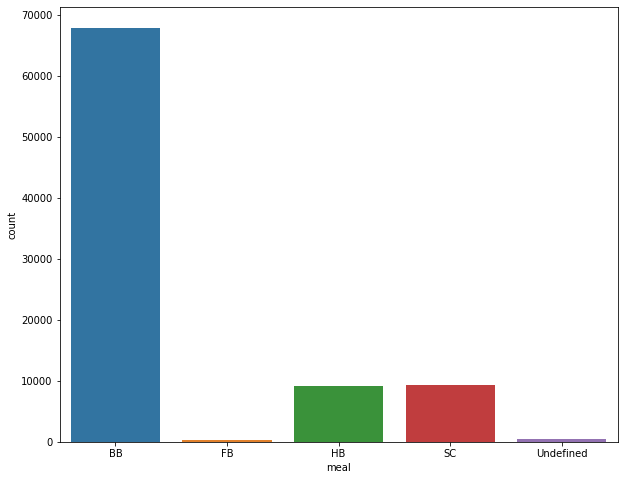
4. Most demanded room type is A. Hotel should increase room type A to increase the revenue.



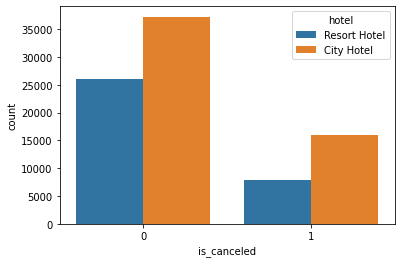
5. Although room type A was on demand and most booked room. Highest revenue was produced by room type H followed by G. We can say that room type H and G are luxury rooms.



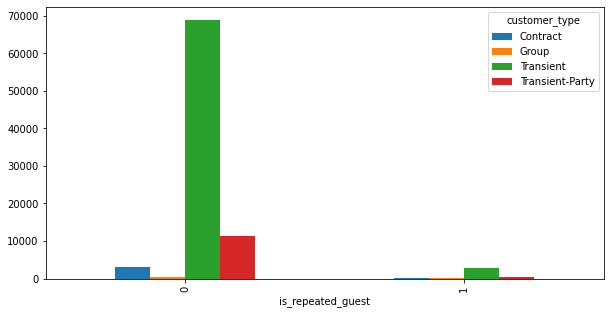
6.As per our analysis, 77% of the bookings are made with bed and breakfast.



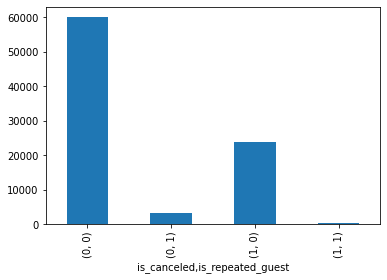
7. We can see that 41.7% of the total bookings were cancelled for City hotel and 21.7% for the Resort hotel.



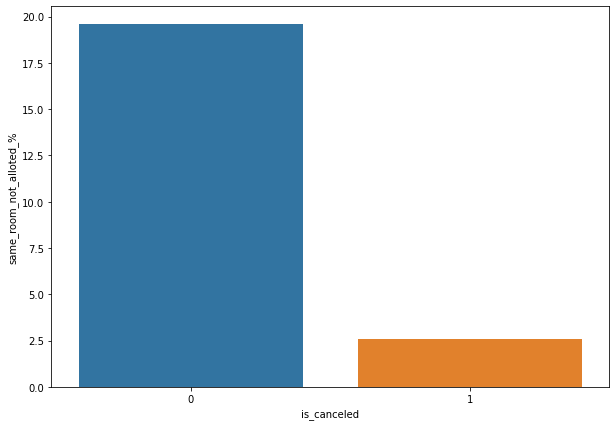
8. Here, we can see that the maximum number of repeated guest are "Transient type" i.e., the "Short-time customers"



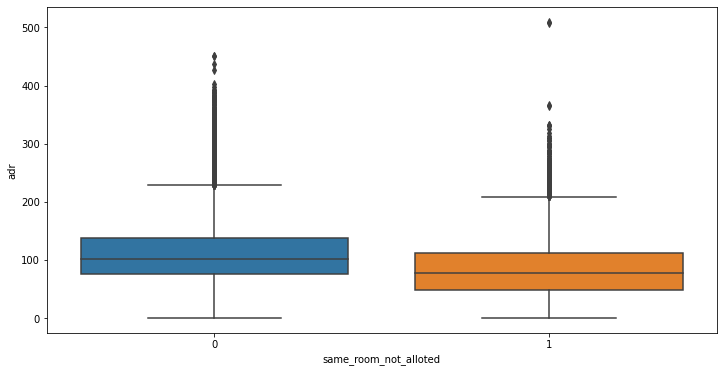
9. From this graph, we can say that when the hotel booking is cancelled and the customer is a repeated guest, the entries are almost '0', indicating that repeated guest is very less likely to cancel his booking with the hotel.



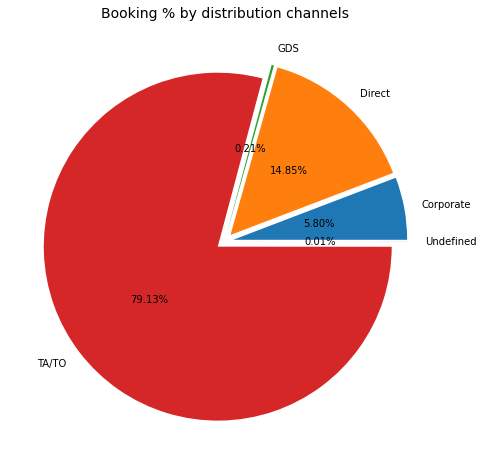
10. Here we see that not getting same room as demanded is not a case of room cancellation. A significant percentage of bookings are not cancelled even after getting different room as demanded.



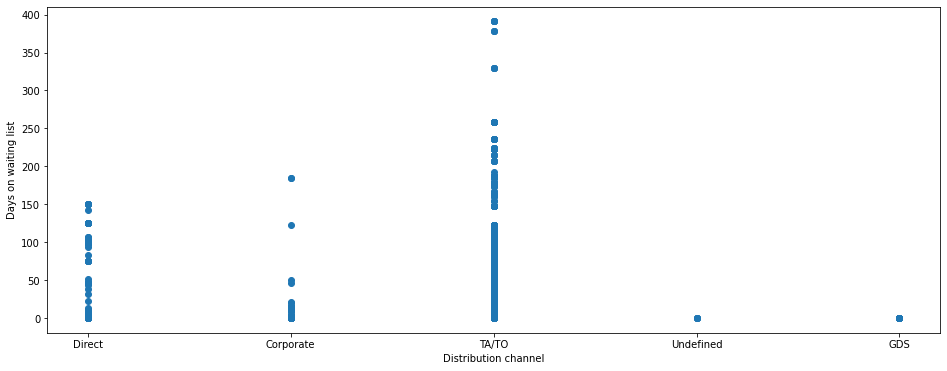
11. Not getting same room do affects the adr, if a guest does not get the desired room, he will pay a bit less.



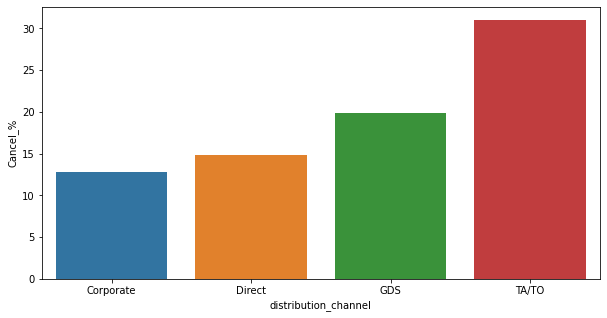
12. Guests use different channels for making bookings out of which most preferred way is TA/TO (i.e., Travel Agents and Tour Operators)



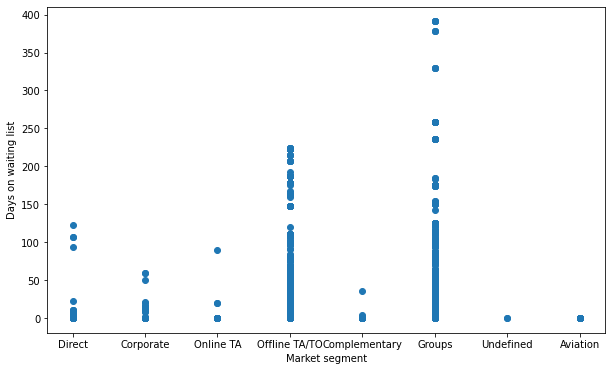
13. We see that the 'Travel Agent' and 'Tour Operators' are the distribution channels for which the highest number of days are there on the waiting list.



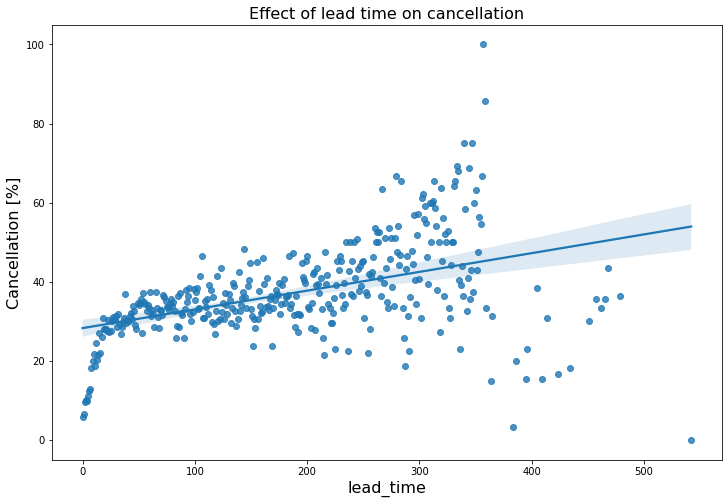
14.TA/TO has highest booking cancellation %. As a result, booking made through TA/TO are 30% more likely to get cancelled.



15. Here, we see that Aviation industry has the minimum number of days on the waiting list. The reason for this could be that when a flight has to land at the location, it has to provide immediate accommodation to all of its working staff such as pilots and air hostages therefore they do not entertain hotels that have a long waiting list. So, in general, the hotel management sees to it that their needs are satisfied immediately and that they have almost no days on the waiting list.



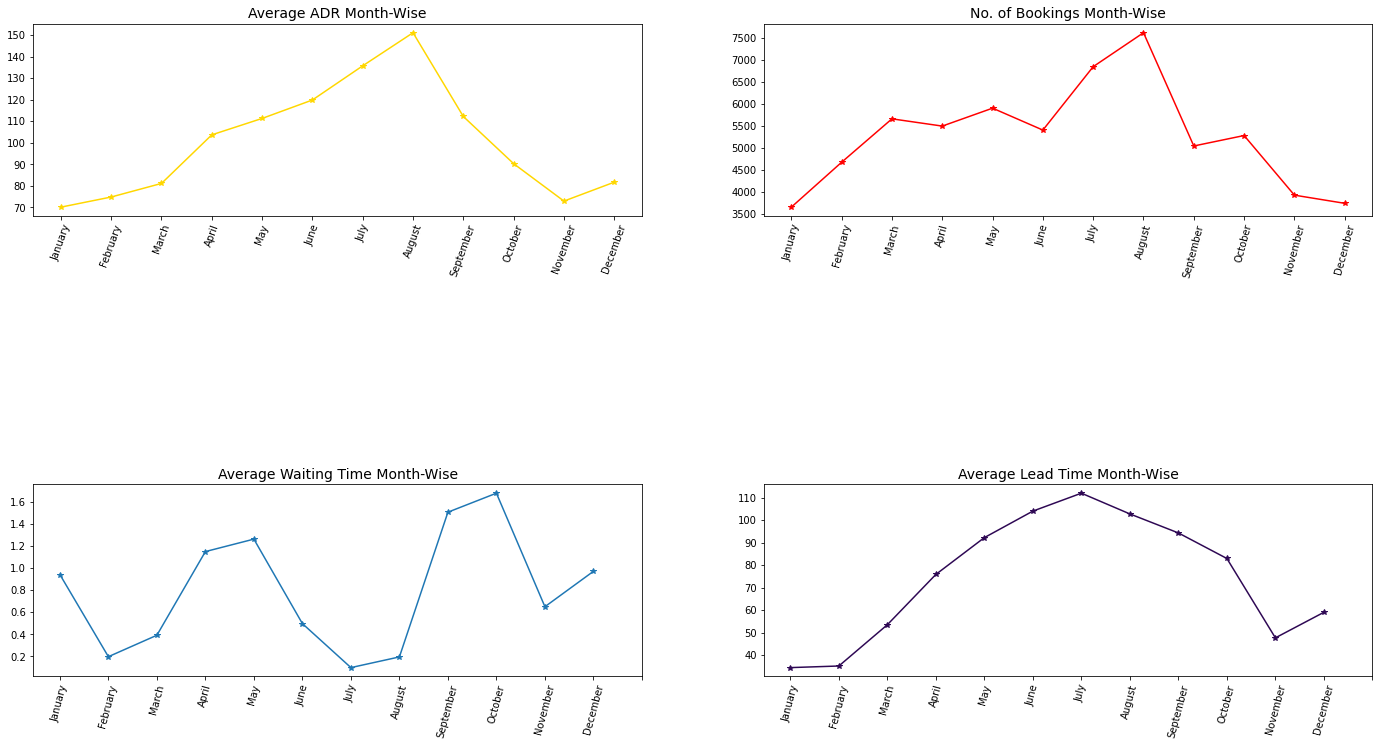
16. We see that the Effect of Lead time on Cancelation is POSITIVE which means that as the lead time increases, so does the number of cancelations.



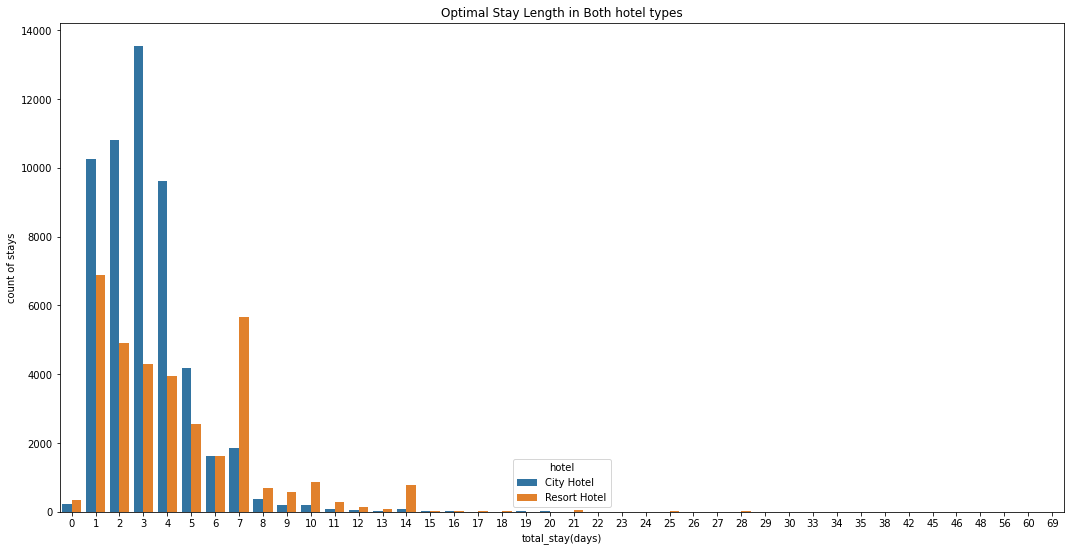
16. Here, we can see that the price and number of bookings are lower in the months of October to February, and the average lead time is also lower in this period compared to other months. The only issue is the wait time, which is slightly longer

from September to October, whereas the waiting time is the lowest and the price is the highest in July and August.

So the best time to book a hotel would be from October to February in order to get a cheaper price as well as some privacy.



17. Most people prefer to stay in hotels for at least 5 days in both types of hotel. However, people tend to stay longer at resort hotels.



**6. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far, we have done EDA, null values treatment, encoding of categorical columns, evaluated the data and performed analysis on various variables and obtained visualization on the given dataset.

1. Higher lead time has higher chance of cancellation.
2. July-August are the busiest and most profitable months for both the hotels.
3. The best time to book a hotel would be between October to February to get for cheaper price.
4. Most people prefer to stay in the hotels for at least five days in both types of hotel. However, people tend to stay longer at resort hotels.
5. Maximum number of repeated guests are "Transient type" i.e., the "Short-time customers"
6. City hotels receive around 60% of bookings, while Resort hotels receive 40%, hence city hotels are busier than Resort hotels.
7. City hotel’s total adr is slightly higher than Resort hotel’s.
8. Most of the guests were from European countries, with the most coming from Portugal followed by Great Britain and France.
9. Guests use different channels for making bookings out of which most preferred way is TA/TO.

9. Almost 30% of bookings via TA/TO are cancelled.

10. Bookings are not affected if they do not receive the same room as reserved. Although different room allotment does lessen the adr.

**7.Challenges:**

**(1) There was lot of duplicate data.**

**(2) Data was present in wrong datatype format.**

**(3) Choosing which visualization techniques to use, was difficult.**

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

**(5) There was lot of Outliers.**

**And many more challenges.**