Hotel Booking Analysis

ALMABETTER

Mohammad Jibran
Siddhi Thakur

| Data Science Trainees |

Abstract:

- This data collection comprises booking information for a city hotel and a resort hotel, including dates of booking, duration of stay, number of adults, children, and/or newborns, and available parking spaces, among other things.

- We will use Python to undertake exploratory data analysis in order to gain understanding from the data in this project .

1. Problem Statement

Many factors influence hotel bookings, including the time of year, the number of guests, the distribution channel, and the hotel type.

The major goal of this project is to do exploratory data analysis and derive insights into all of the important elements that influence hotel reservations.

2. Introduction

The hotel industry is the part of the service industry that deals with guest housing. 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.

The Hotel booking data set includes booking due to arrivals dated between July'15 to August'17 which consists of Numerical, Categorical and Binary datatypes.

The hotel types, is canceled, arrival date year, arrival date month, stays in weekend nights, stays on weekend nights, country, market segment, distribution channel, and other columns in the data set allowed us to derive important insights from the data set.

3. Understanding the data

In the datasets, there are 119390 entries and 32 columns. The hotel bookings are mostly found between July 1, 2015 and August 31, 2017, with the customer successfully arriving and canceling the reservation.

```
<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
```

Column Name	Column Descriptions
hotel	The category of hotels, which are two city hotel and resort hotel.
is_canceled	The value of column shows if the booking was canceled or not. Values [0,1], where 0 indicates not canceled and 1 indicates the cancellation.
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_number	Week number of year for arrival date (1-53)

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
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
deposit_type	**No Deposit**– no deposit was made **Non Refund** – deposit was made in the value of the total stay Cost **Refundable** – a deposit was made with a value under the totalcost of stay.
agent	ID of the travel agency that made the booking
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)	Average Daily Rate = sum of all loging transaction/Total no. of staying night
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

4. Data Wrangling

Data wrangling is the process of removing null values and duplicate entries from datasets. The EDA conducted on the clean datasets results in more accurate data interpretation and improved presentation of distinct aspects.

Steps involved:

- ### Importing important libraries

Our major goal in this step was to import all of the necessary libraries to aid us in exploring the issue statement and doing EDA to draw conclusions based on the data collection.

Libraries we used :

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. e.g., 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 Data frames and pandas' libraries. The graphs can also be customized.

- ### Understanding the data set

Following that, we checked the data set. What is the size of the data set? How many columns and rows are there? What are the critical columns for resolving the problem statement? In the data set, how many null values are there.

- ## **Null values Treatment**

Our dataset contains a large number of null values which might tend to disturb our insights. We get rid of them for better understanding. Column named company and agent had the huge number of null values which are eliminated as they are not necessary.

- ## **Exploratory Data Analysis**

After treating the null values, we started with the EDA. We performed EDA and tried to understand the data by asking some questions.

How hotel booking works?

- ● The Hotel booking is done by the guests before visiting the hotel. The hotel receives the booking from the following distribution channels:
  - ○ Direct
  - ○ Corporate
  - ○ TA/TO
  - ○ GDS
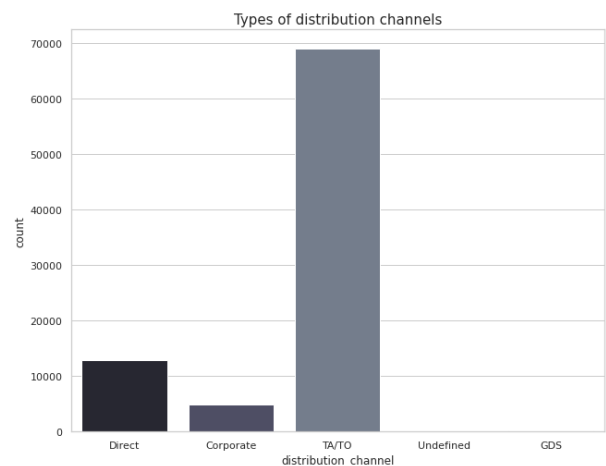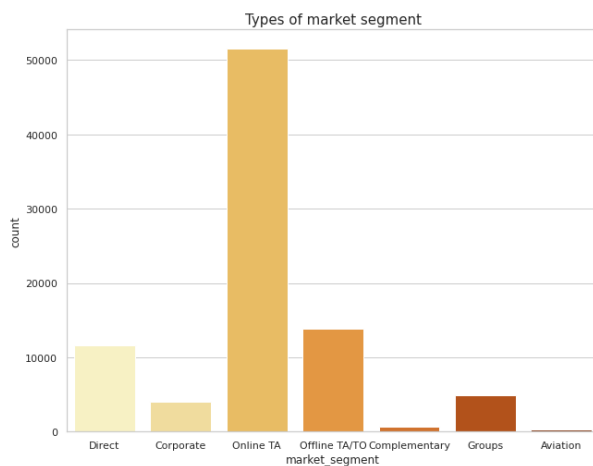
- ● The Hotel then assign a room as per the availability and demand. If the demanded room is unavailable the hotel assigns a different room.
- ● The Hotel receives no deposit, or refundable deposit or non-refundable deposit against the bookings.

Which type of hotel is most preferred by the customers?

- There are two types of hotels i.e., City hotel and Resort hotel and as per given dataset people mostly preferred city hotel.
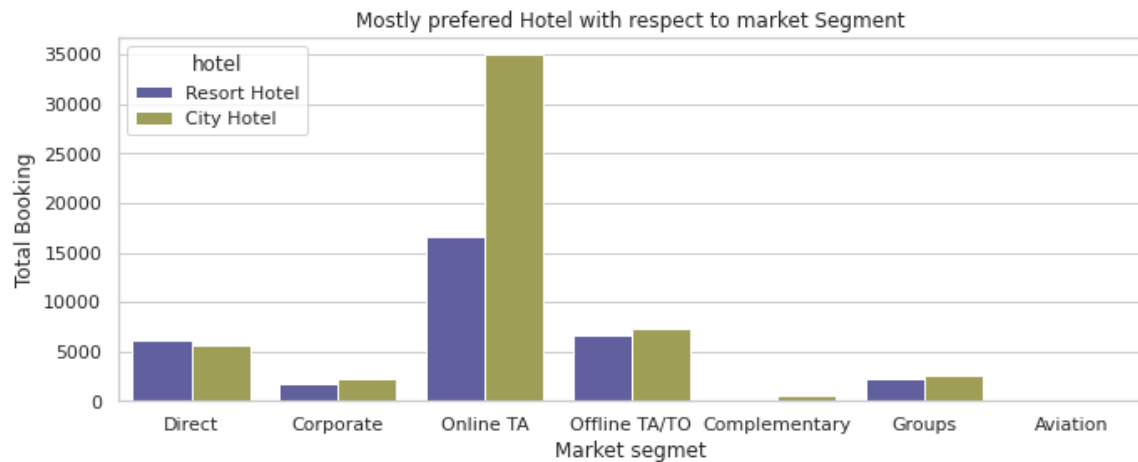

Which is the busiest month for hotels business?

- Booking trends are also increasing in the middle of the year, with August being the busiest. Summer appears to be the busiest season for hotel bookings.
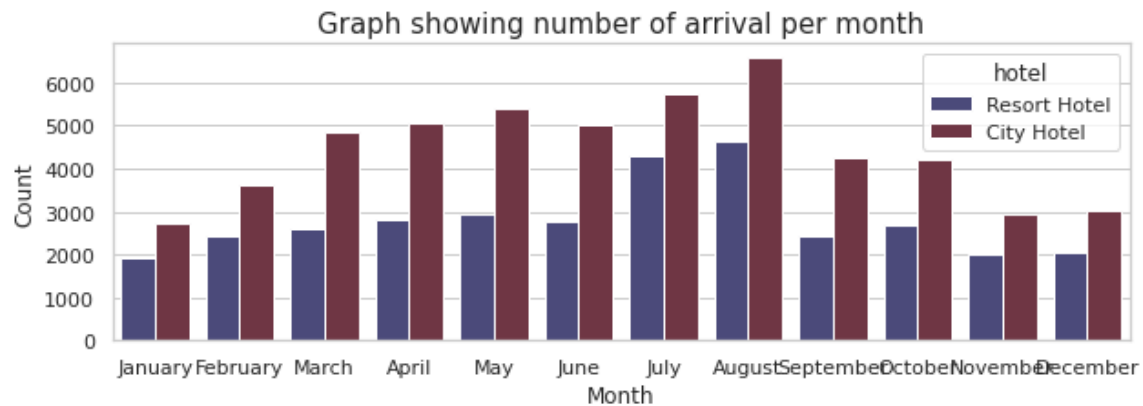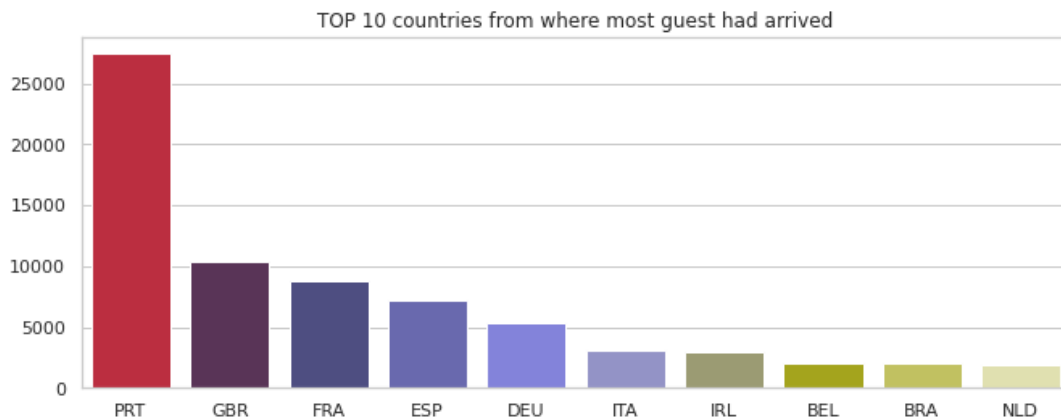
Which is the most preferred business segment?

Customers prefer to book through the 'online TA' section for both types of hotels.


Which country's guest have high number of bookings?

We have the most visitors from western European countries, namely Portugal, France, and the United Kingdom. We can tell the marketing department to focus on people in this area.

Which months have highest and lowest booking rates?

The cost of staying in a resort hotel is exorbitant. It appears to be true since resort hotels place a premium on it. Hotel rates in the city do not fluctuate considerably.
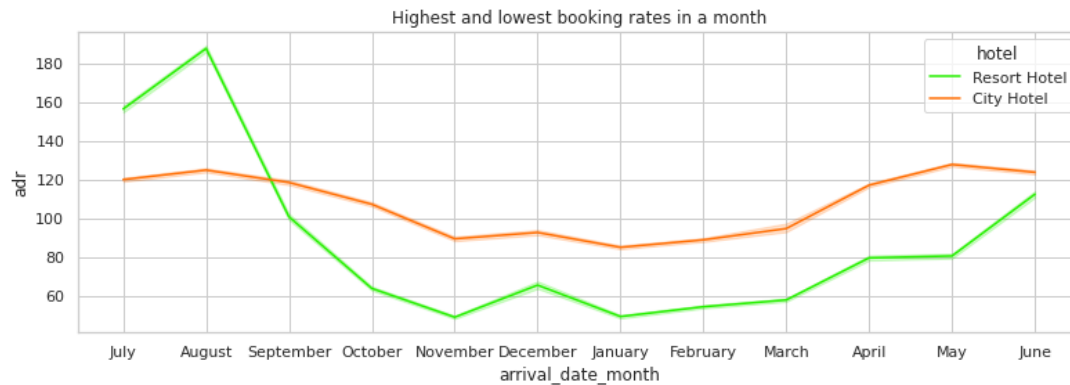
# Do guests with kids have a particular preference over hotels?

It's visible that the majority of the guests with kids prefer city hotels. It may be because of high rates for the resort hotel as well as resort hotels are also less in numbers .
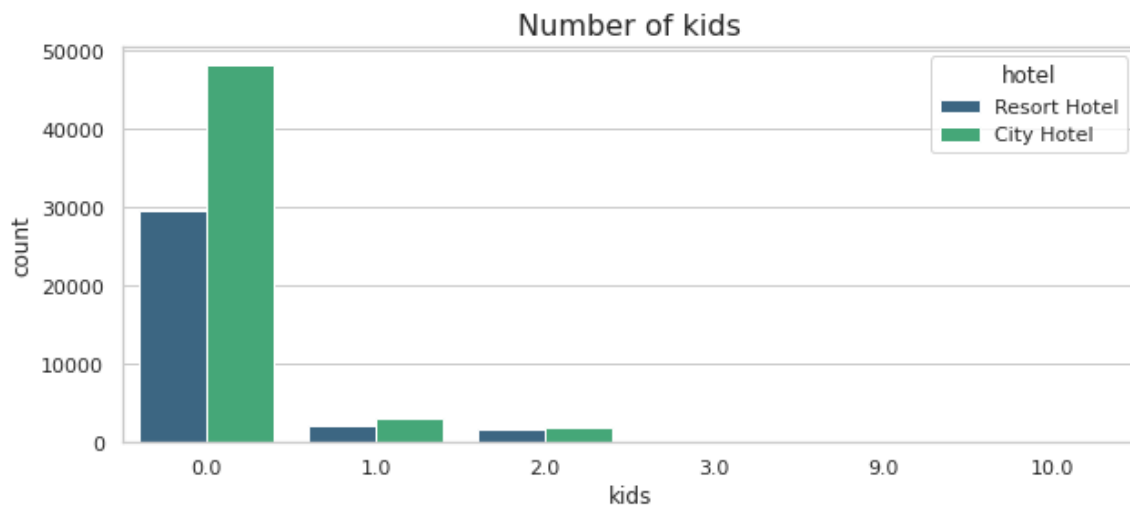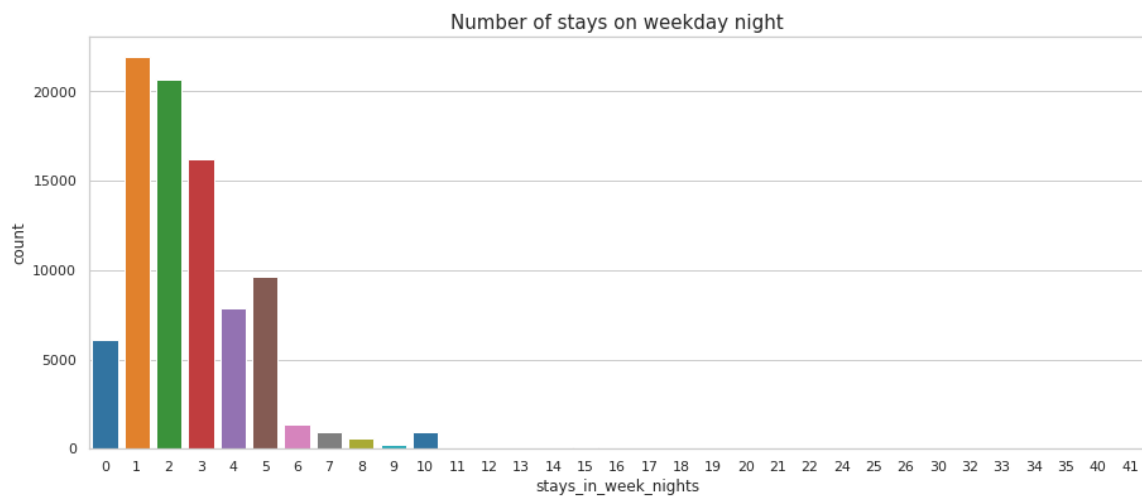
Stay over weekends or weekdays, what guest preferred most?

The majority of the stays appear to be on weekday nights. As a result, whatever we observed for the chart on the day of the month appears to be random.

- **Stay on weekend nights -**

Number of stays on weekend nights

- **Stay on weekdays nights -**
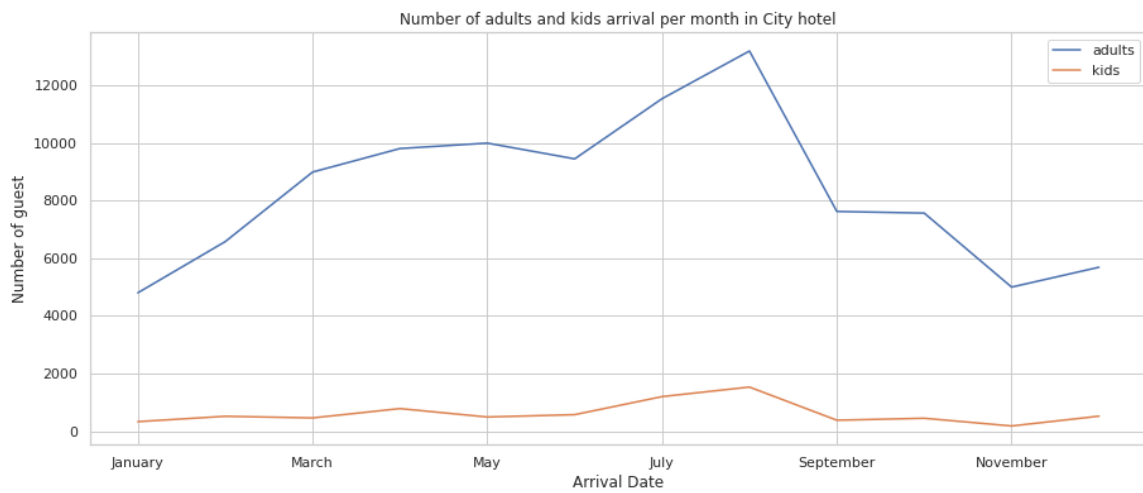

Number of stays on weekday night
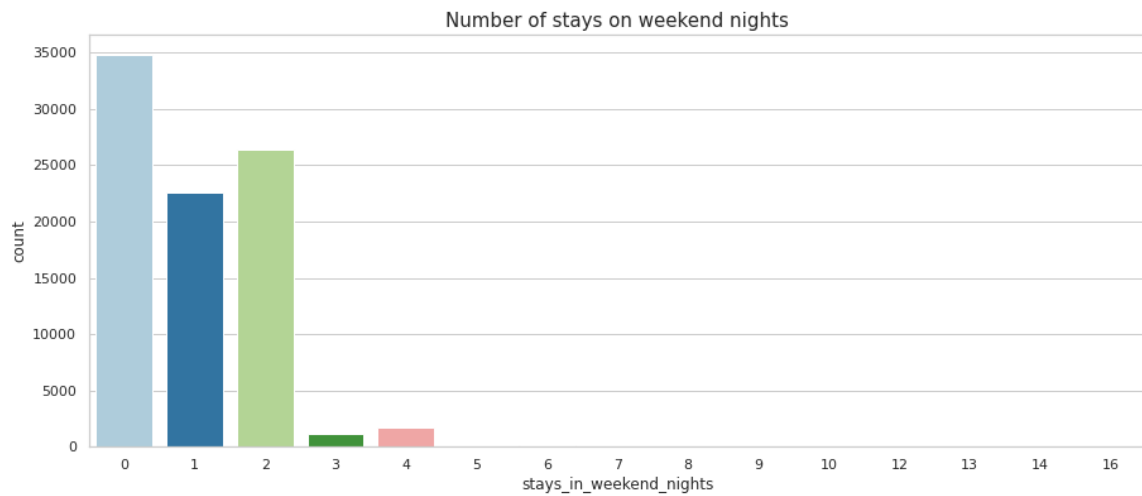
How many adults and kids arrived per month in each hotel?

The number of adults and children appears to be higher in city hotels. For large families city hotels are a preferable option.

For resort hotels –

Number of adults and kids arrival per month in Resort

For City hotels –

Number of adults and kids arrival per month in City hotel
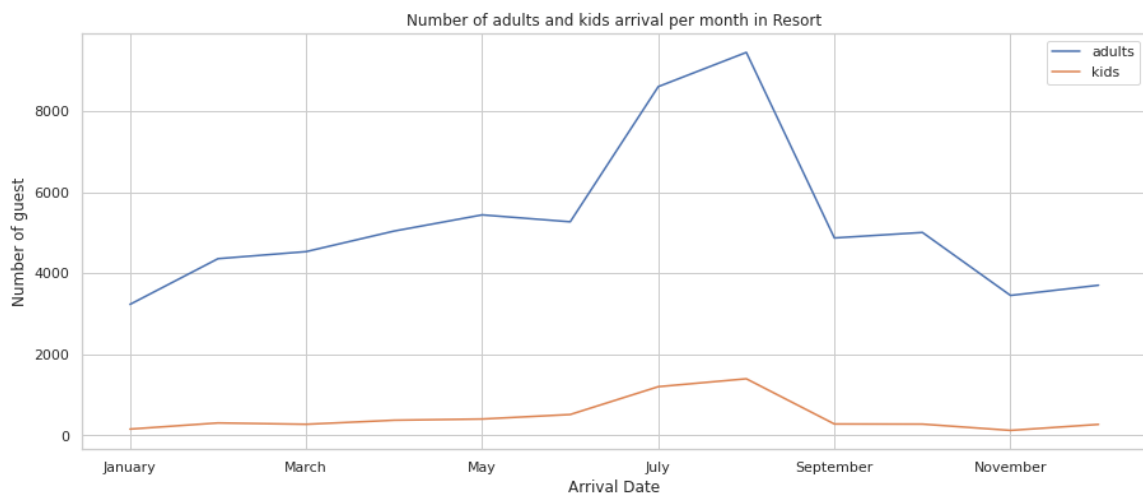
Conclusion

- Majority of People preferred City hotel over Resort hotel. Especially guest with families prefer city hotel. Resort hotels are expensive than city hotels

- In year 2016 all hotel bookings are at their highest.

- August is the busiest month. In August there are 6586 City Hotel Bookings while 4643 Resort hotel booking. While November to Jan is low session for hotels

- Month ending has the least number of arrivals in both types of hotels

- Mostly people don't cancel their bookings but if compared with the resort booking city hotels have high cancellation rate because they have high booking rate as well. West Europe visitors didn't cancel their booking

- Most people don't ask car parking

- Online Travel Agent is most preferred distribution channel for both types

- Most Hotels didn't ask for deposits

- Most of the guest stay in weekdays in hotel as compare to weekends night