Capstone project

EDA on Hotel Booking
Analysis by

“Suraj kad”

(suraj.kad.90@gmail.com)

Data Science Trainees

Alma better

Let's Catch The Defaulters

a) Defining Problem statement

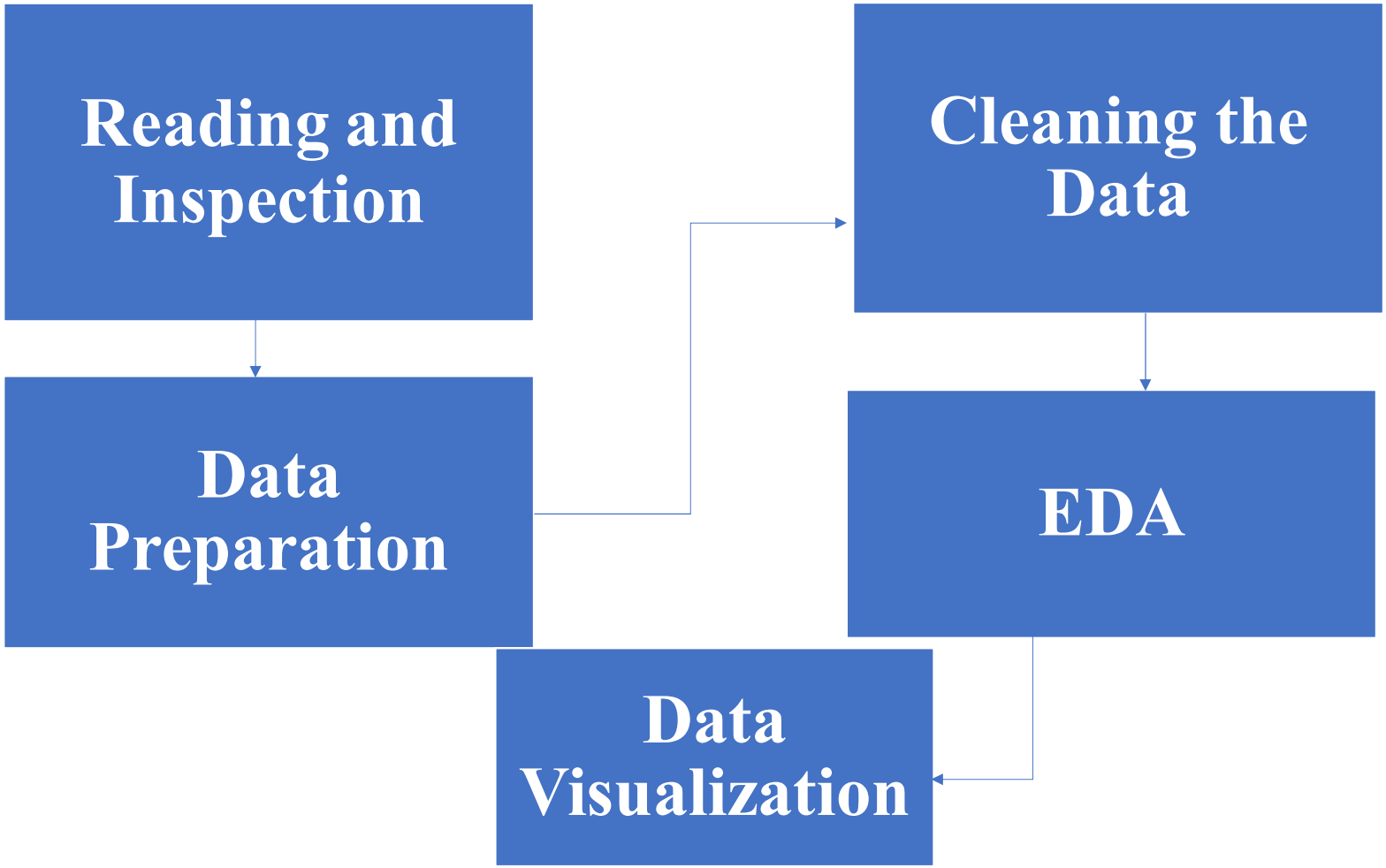
b) Reading and Inspection

c) Cleaning the Dataset

d) Exploratory Data Analysis

e) Data Visualization

DATA PIPELINE



Data Pipeline :

A data pipeline is a set of tools and processes used to automate the movement and transformation of data between a source system and a target repository.

Data Reading:

read and write tabular data using pandas functions.

Data preparation:

Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (*known as* structured data) for further analysis, often through the use of spreadsheet or statistical software.

Cleaning the Data: :

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Exploratory data analysis:

Once the datasets are cleaned, they can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data.

Data visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Problem Statement

- Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?
- This hotel booking dataset can help you explore those questions!
- 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.
- All personally identifying information has been removed from the data. Explore and analyze the data to discover important factors that govern the bookings

Data Collection And Understanding

- After collecting data it's very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let's understand this 32 columns.

Data Description:

- 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.
- distribution_channel : Booking distribution channel.
- is_repeated_guest : is a repeated guest (1) or not (0)
- 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.
- 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 the 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 (e.g. twin bed or high floor)
- reservation_status : Reservation last status.

Reading and Inspection

```
hotel
```

Python

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	s
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	
...	...	...	...	...	...	...	...	
119385	City Hotel	0	23	2017	August	35	30	
119386	City Hotel	0	102	2017	August	35	31	

```
hotel.head()
```

Python

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_i
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	

5 rows × 32 columns

- Read and write Hotel Booking (**tabular)** data using pandas functions

Pandas **head()** method is used to return top n (5 by default) rows of a data frame for hotel booking dataFrame.

Reading and Inspection

```
hotel.info()

Output exceeds the size limit. Open the full output data in a
<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
```

```
(hotel.shape)
                                                                    Python
(119390, 32)

hotel.describe()
                                                                    Python
```

lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	
390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	1193
104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	
106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	
0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	
18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	
69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	
160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	
737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	

- The info() method **prints information about the hotel booking DataFrame**. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

- Shape attribute in Pandas **enables us to obtain the shape of a DataFrame**. In this DataFrame has a shape of (119390, 32) , this implies that the DataFrame is made up of 119390 rows and 32 columns of data.
- The describe() method **returns description of the data in the DataFrame.** If the DataFrame contains numerical data, the description contains these information for each column

Reading and Inspection

```
hotel.isnull().sum()
```

```
Output exceeds the size limit. Open the full output data in a text editor
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           488
market_segment                      0
```

- We will count total number of NaN data present in hotel booking dataset and find out the number of NaN or missing values in each columns.
- As we can see in the output**, hotel.isnull().sum()**
  Function has return **country** columns **488 NaN** data are present.

Cleaning the Data

The dataset contains Missing values . Drop unecessary columns: Lets drop columns with high missing values.

```
hotel=hotel.drop(['agent','company'],axis=1)
```

Python

Country has 488 rows with the NaN values. 488 rows out of 119390 is negligible hence we will just remove.

```
hotel = hotel.dropna(axis = 0)
hotel.isnull().sum()
```

Python

```
Output exceeds the size limit. Open the full output data in a text editor
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
```

Data Preparation

Lets Rename the columns for better readability

```
hotel.columns = ['Hotel', 'Canceled', 'LeadTime', 'ArrivingYear', 'ArrivingMonth', 'ArrivingWeek','ArrivingDate', 'Weeke
                 'WeekStay', 'Adults', 'Children', 'Babies', 'Meal','Country', 'Segment', 'DistChannel','RepeatGuest', 'Pre
                 'PrevBook', 'BookRoomType','AssignRoomType', 'ChangeBooking', 'DepositType', 'WaitingDays',
                 'CustomerType', 'ADR','ParkSpace', 'SpecialRequest','Reservation', 'ReservationDate']
```

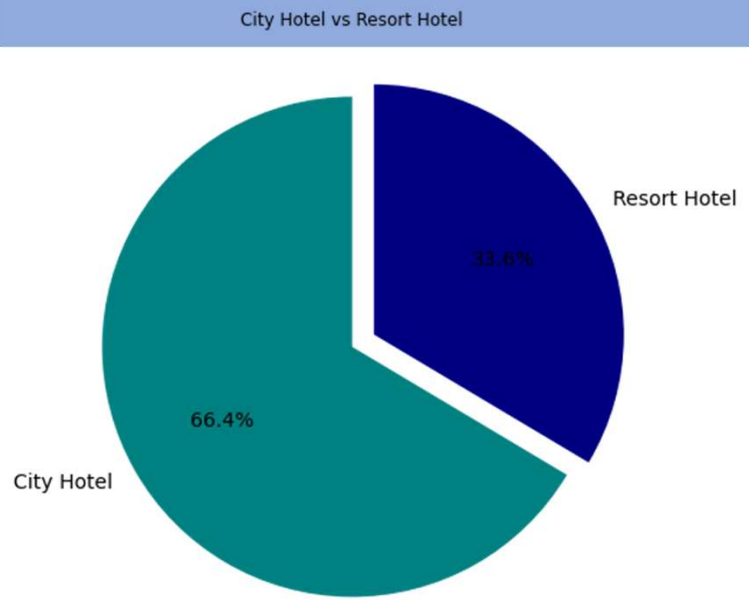
```
#convert the datatypes to string
hotel['ArrivingYear'] = hotel['ArrivingYear'].astype('str')
hotel['ArrivingMonth'] = hotel['ArrivingMonth'].astype('str')
hotel['ArrivingDate'] = hotel['ArrivingDate'].astype('str')

hotel['Canceled'] = hotel['Canceled'].astype('str')
hotel['RepeatGuest'] = hotel['RepeatGuest'].astype('str')
```

```
# Lets convert arrival date to datetime
hotel['Arrival Date'] = hotel['ArrivingDate'] + '-' + hotel['ArrivingMonth'] + '-' + hotel['ArrivingYear']
hotel['Arrival Date'] = pd.to_datetime(hotel['Arrival Date'], errors='coerce')
```

- This method is quite useful when we need to rename some selected columns because we need to specify information only for the columns which are to be renamed

- we'll convert each value of a column of integers to string using the **astype(str)** function

EDA

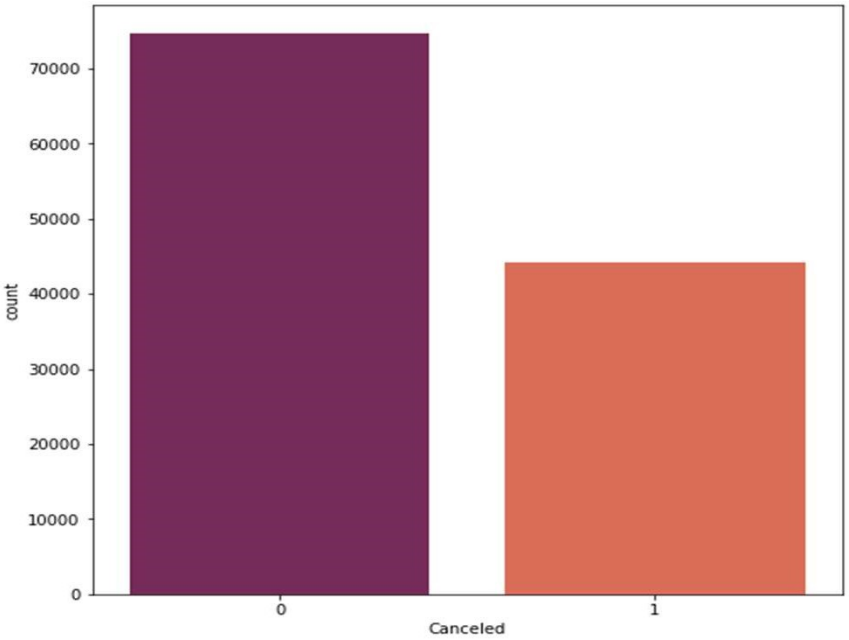


- Around 66.4% bookings are for City hotel and 33.6% bookings are for Resort hotel.
- Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.

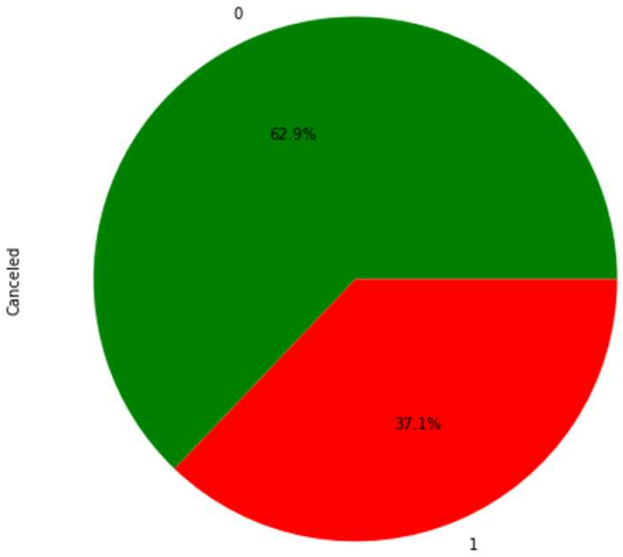
• Total stay length and lead time are slightly correlated. This may means that for longer hotel stays, people generally plan little before the actual arrival.

• adr is slightly correlated with total_people, which makes sense as more no. of people means more service to deliver, therefore more adr.

Booking Cancelations



- Majority of bookings were not canceled, still some half of the bookings were canceled



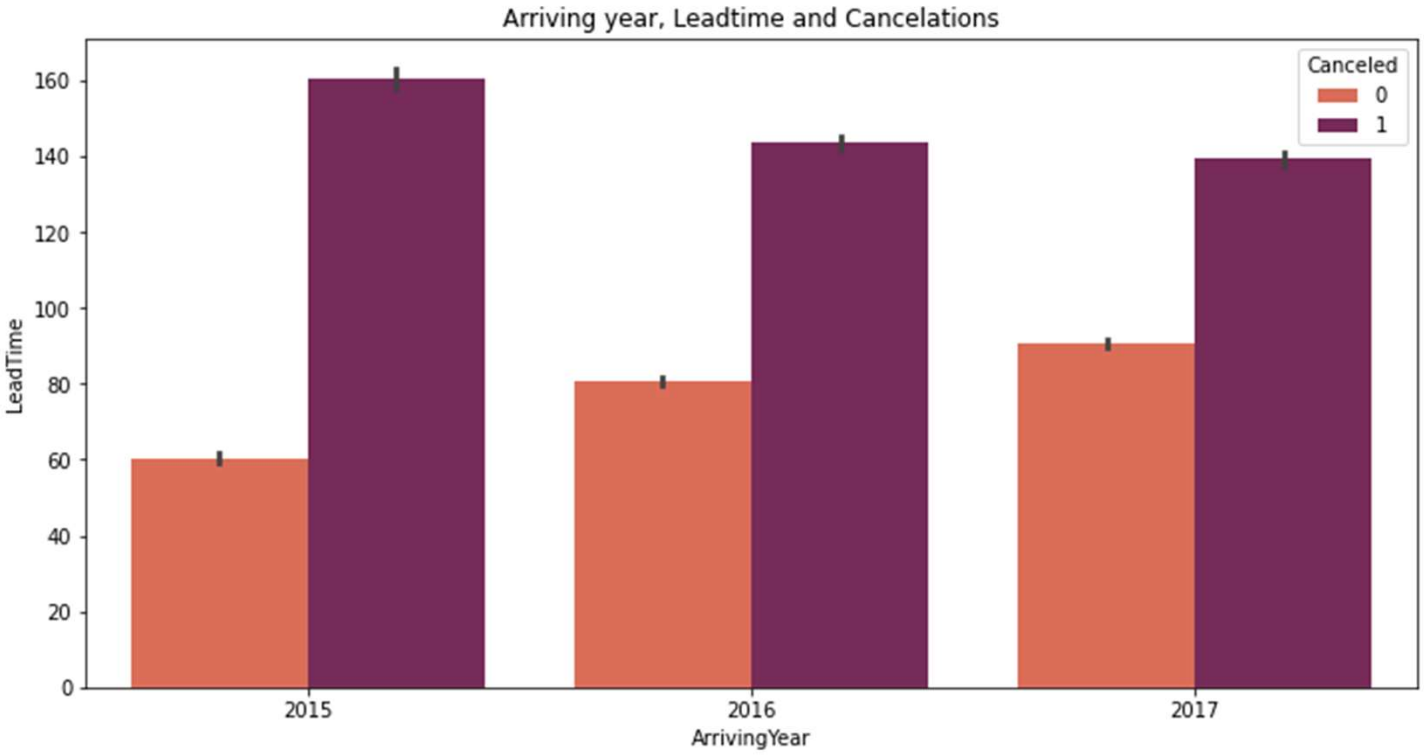

```
Total Bookings canceled
-----------------------------------
0    74745
1    44153
Name: Canceled, dtype: int64
-----------------------------------
***********************************
Cancelation percentage in both hotels
-----------------------------------
0    0.628648
1    0.371352
Name: Canceled, dtype: float64
```

- According to the pie chart, 63% of bookings were not canceled and 37% of the bookings were canceled at the Hotel.

Cancelation Rate in City Hotel & Resort Hotel

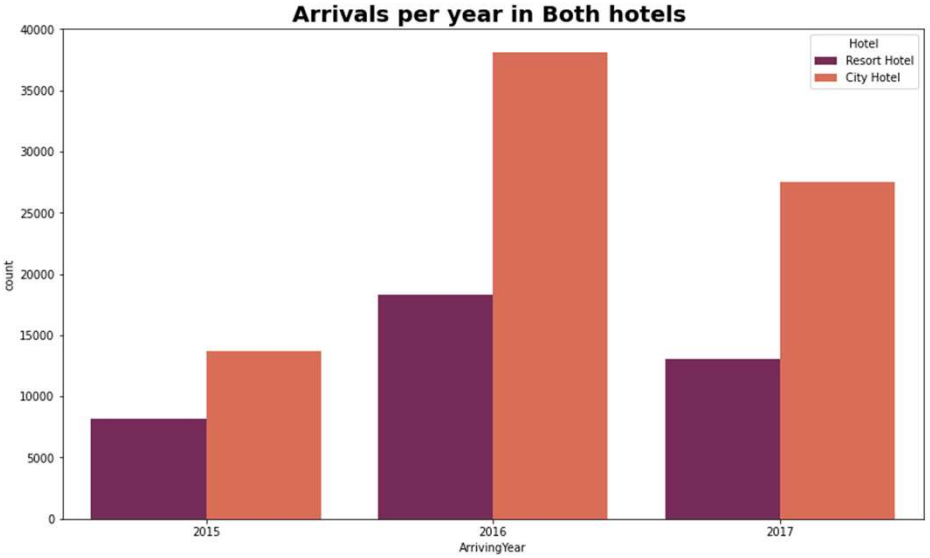
- Most bookings were in city hotel
- Cancelations in Resort hotel is less compared to city hotel

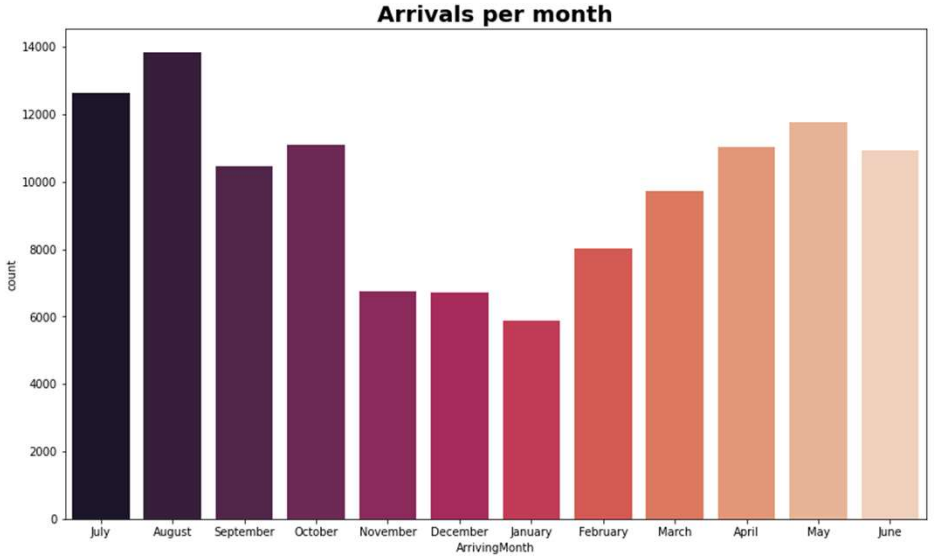
Arriving year, Leadtime and Cancelations



For all the 3 years, bookings with a lead time less than 100 days have fewer chances of getting cancelled, and lead time more than 100 days have more chances of getting cancelled.

Arrivals per Year & Arrivals per Month





- We can see most of the bookings were in the year 2016 and bookings were done in City hotel
- Most bookings were done in the month of May , June, July, August add Codded Markdown

- From the month of July to August the number of bookings increased and in August, City Hotel got most number of guests.

Arrivals per Year & Arrivals per Month

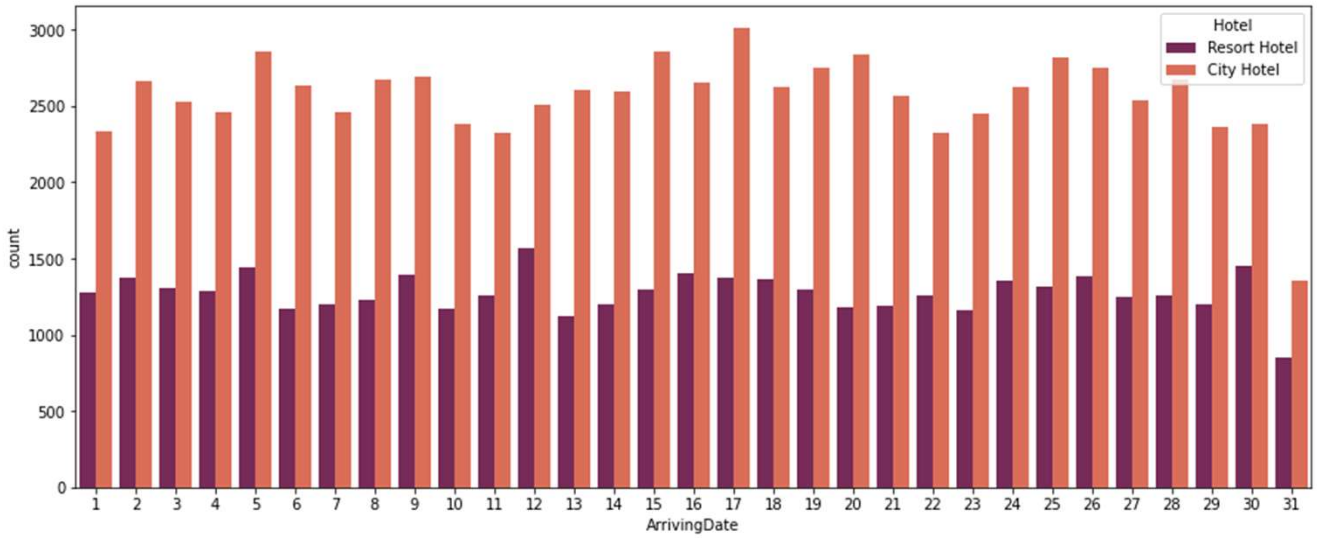
```
hotel.ArrivingMonth.value_counts(normalize=True)
✓ 0.8s
August        0.116503
July          0.106209
May           0.099068
October       0.093315
April         0.092895
June          0.091902
September     0.088033
March         0.081911
February      0.067385
November      0.056788
December      0.056586
January       0.049404
Name: ArrivingMonth, dtype: float64
```

47% bookings were done in 2016, 34% in 2017 and 18 percent in 2015. We can see increasing tendency in bookings year wise

```
hotel.ArrivingYear.value_counts(normalize=True)
✓ 0.1s
2016    0.474651
2017    0.341503
2015    0.183847
Name: ArrivingYear, dtype: float64
```

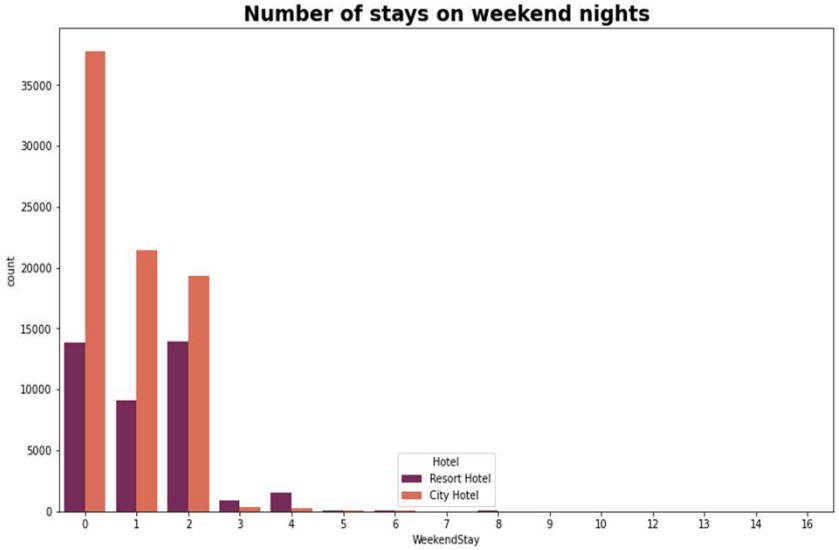
August is the most occupied (busiest) month with 11.6% bookings and January is the most unoccupied month with 4.9% bookings.

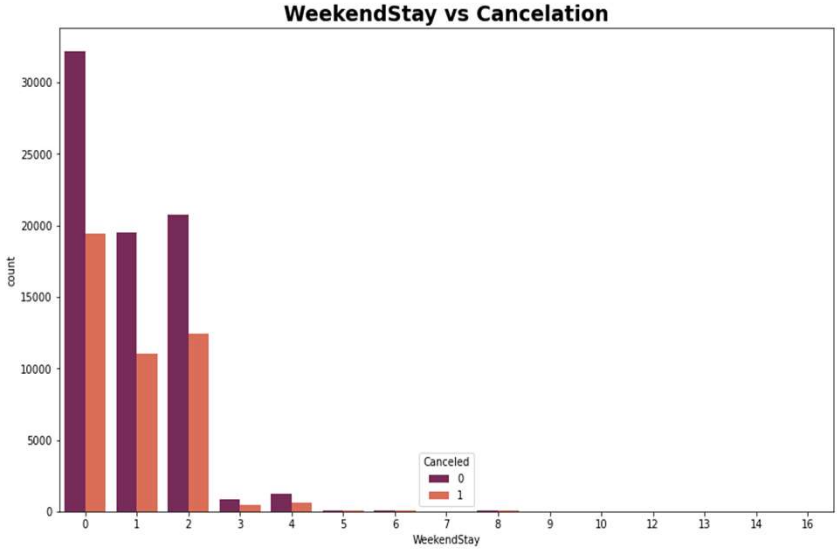
Arriving Date



- Month end day has very less arrivals
- Bookings are more in City hotel

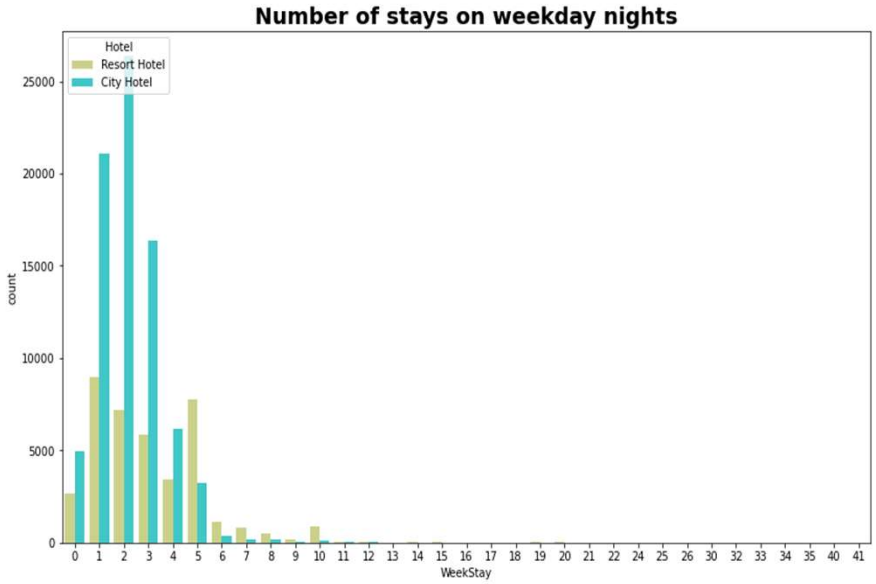
Weekend Stays and Cancellations

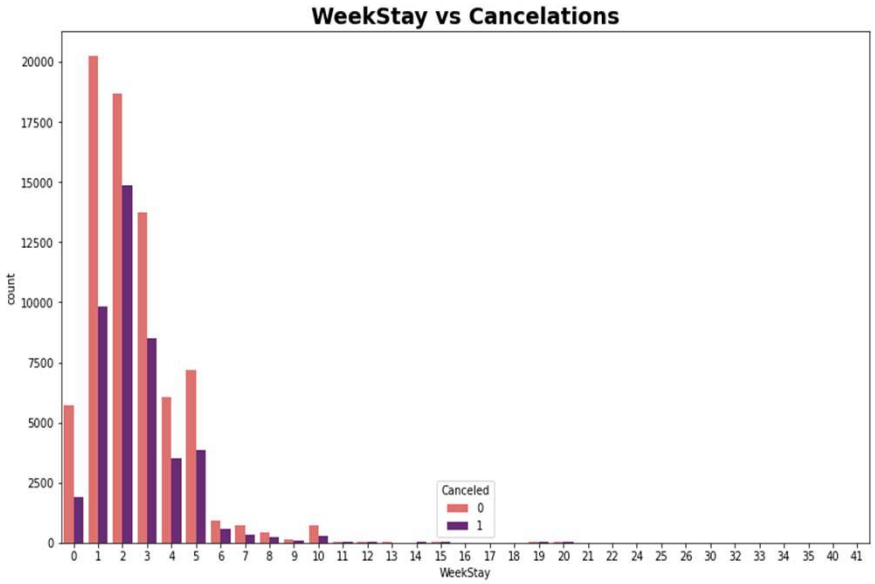




- In the first graph we can see that most of the weekend nights were booked in City Hotel
- Second plot shows most of weekend nights which were booked were not canceled

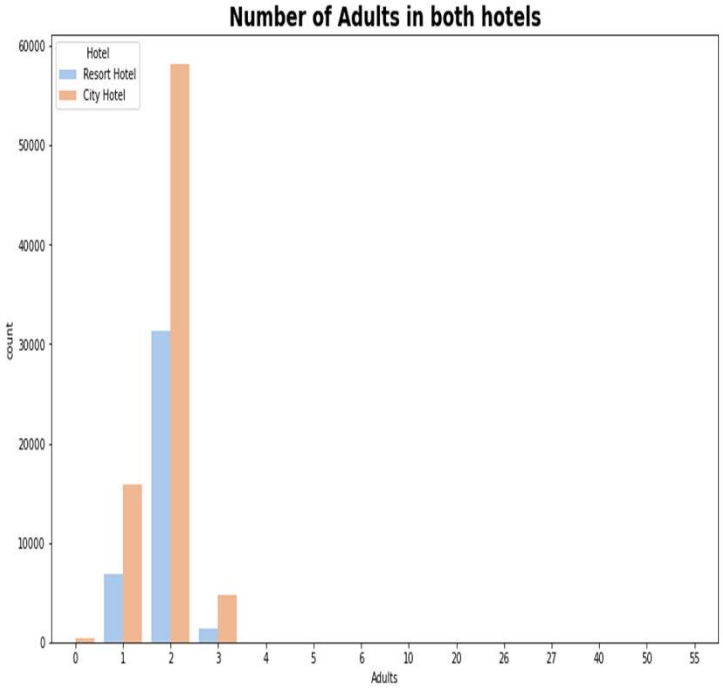
Week Stays and Cancellations

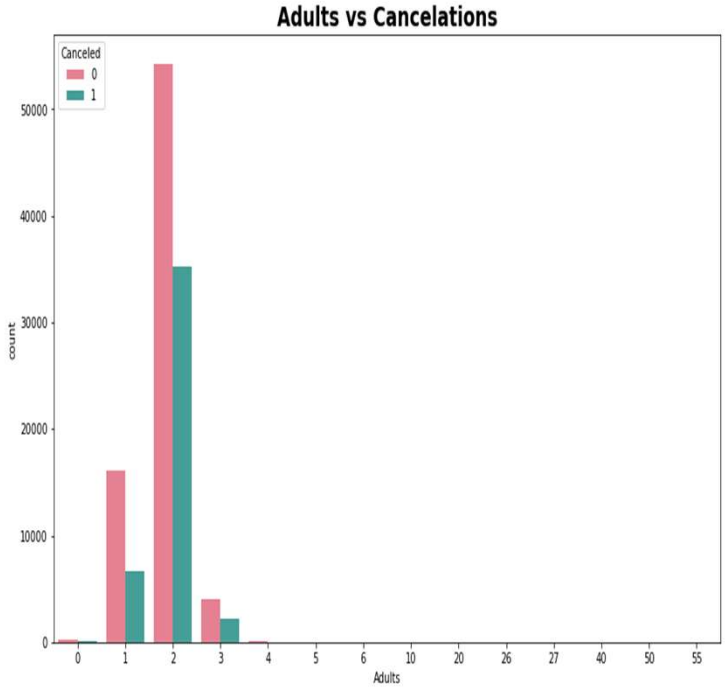




- Weekday night stays were more in City Hotel
- Less cancelations were observed

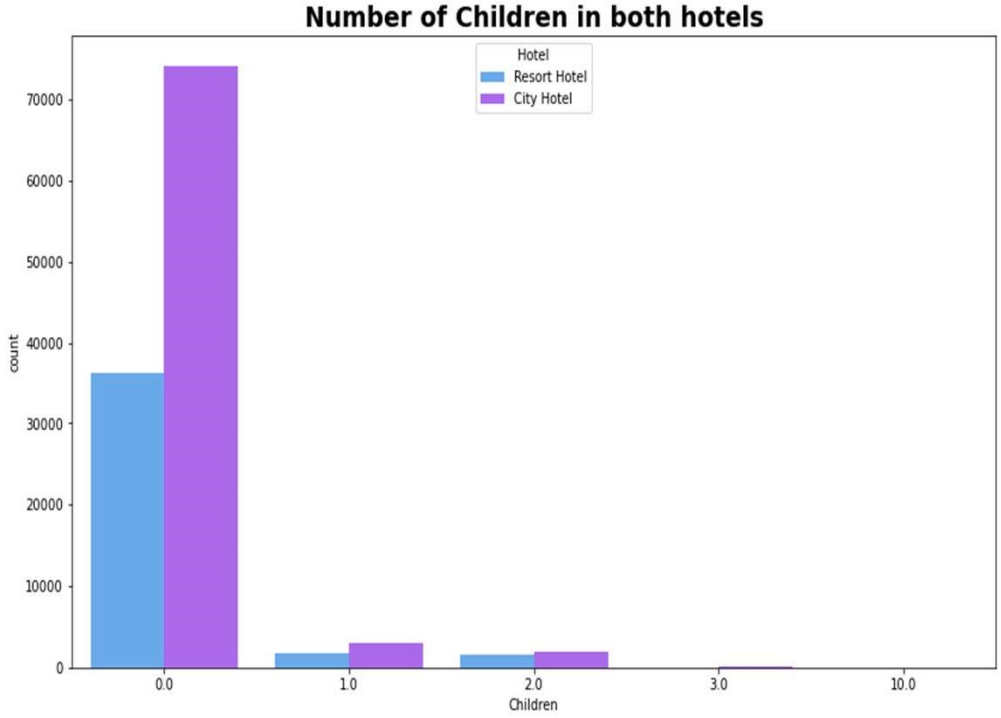
Bookings and Cancellation (Adults)

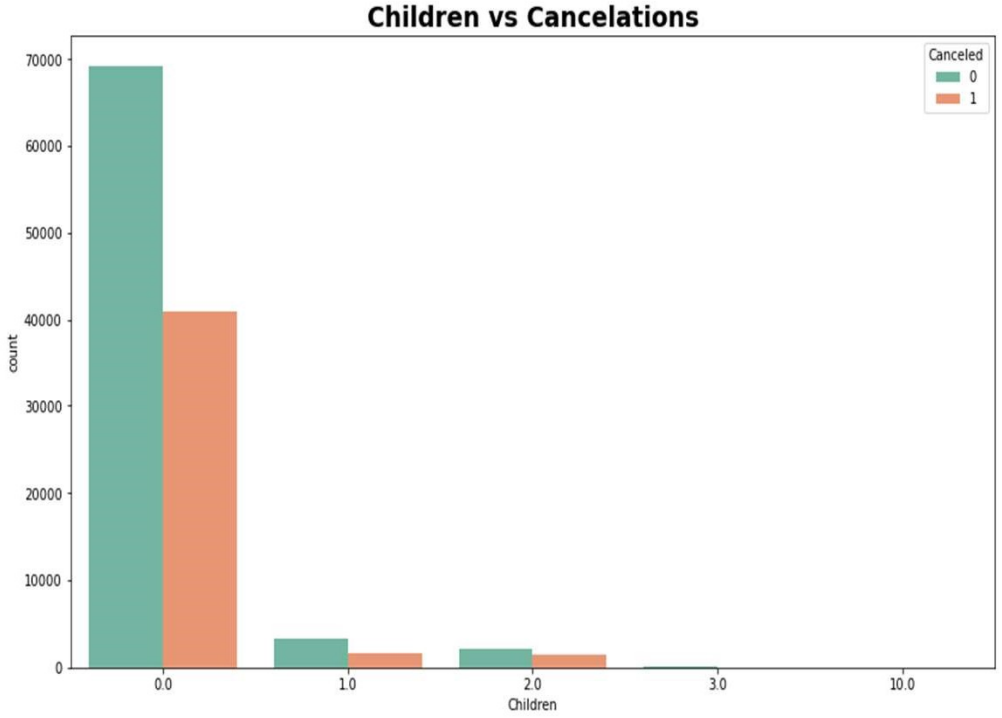




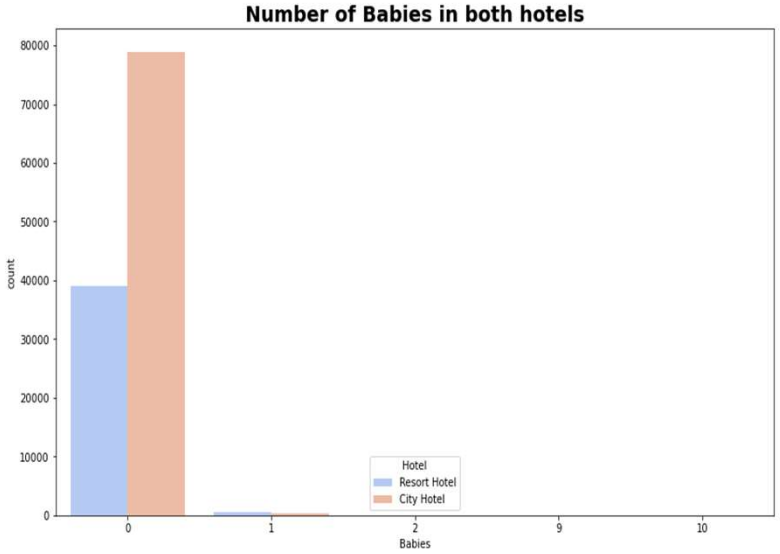
Adults who were 2 in number are more and preferred city hotel rather than resort hotel, infarct more than half the visitors even canceled the bookings

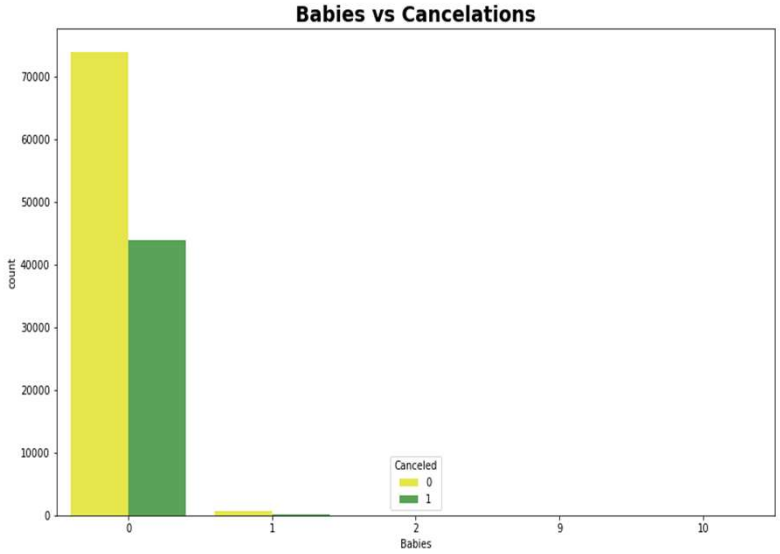
Bookings and Cancellation (Children)





Bookings and Cancellation (Babies)



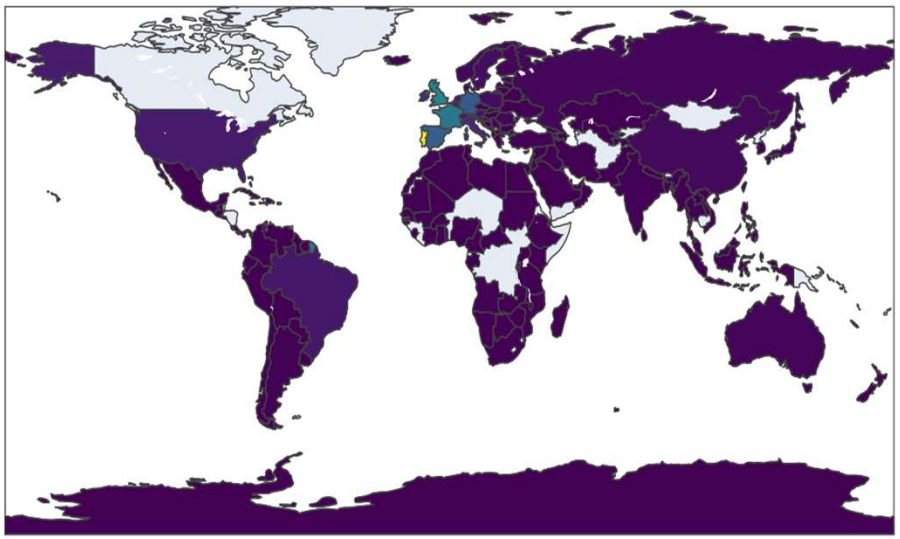


* Most visitors were arrived in pair with no children/ Babies and preferred City hotel over resort hotel

* visitors who had 1 or 2 children also preferred city hotel

Home Country of Visitors

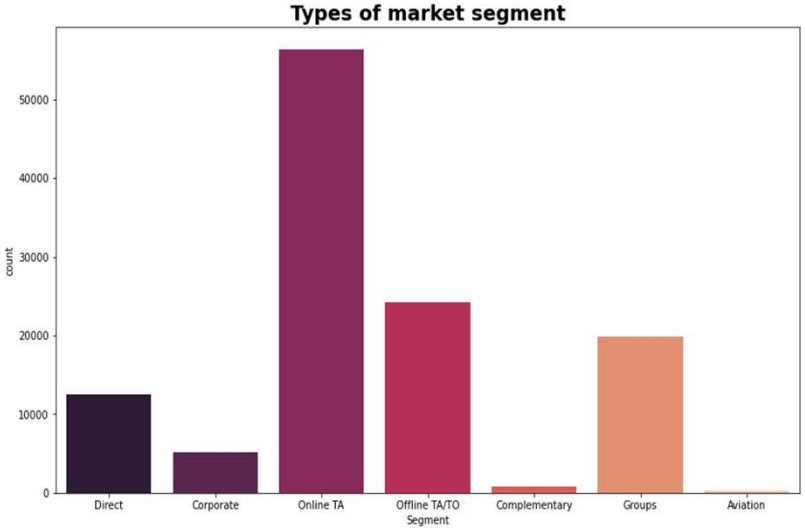
Home country of visitors



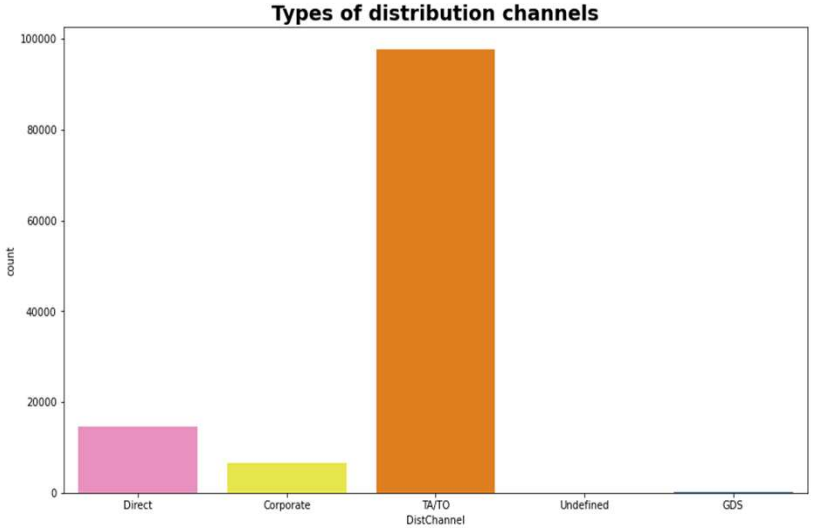

```
hotel.Country.value_counts(normalize=True)

PRT     0.408636
GBR     0.102012
FRA     0.087596
ESP     0.072062
DEU     0.061288
          ...
DJI     0.000008
BWA     0.000008
HND     0.000008
VGB     0.000008
NAM     0.000008
Name: Country, Length: 177, dtype: float64
```

Most of the visitors are from western Europe, namely France, UK and Portugal being the highest.

Market Segment & Distribution Channels



- Majority Distribution channels and Market segment were Travel agencies wither offline/online. So better focus more on this



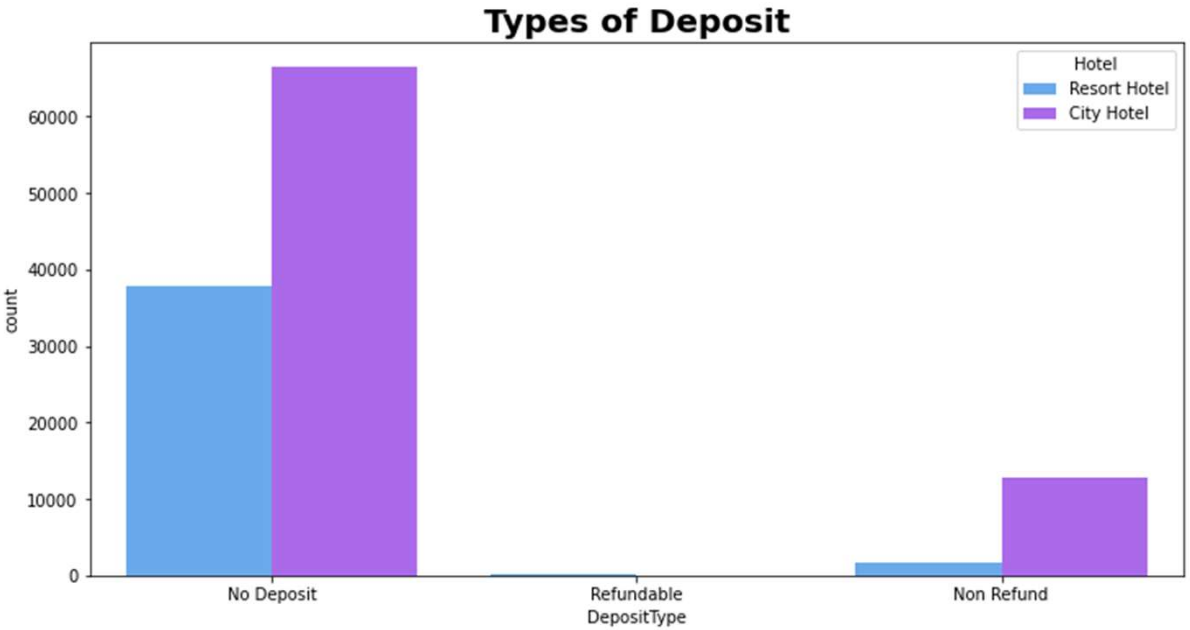
- Here we can see that the most of guest are making reservation through TA/TO channels which is travel agency and tour operator.
- Than the second most used channel is direct.
- Channel which is mostly used for early booking of hotels is also TA/TO.

Market Segment & Distribution Channels

Around 47% of bookings are made via Online Travel Agents, almost 20% of bookings are made via Offline Travel Agents and less than 20% are Direct bookings without any other agents.

```
hotel.Segment.value_counts(normalize=True)
✓ 0.1s
Online TA        0.474373
Offline TA/TO    0.203199
Groups           0.166580
Direct           0.104695
Corporate        0.042986
Complementary    0.006173
Aviation         0.001993
Name: Segment, dtype: float64
```

Types Of Deposit

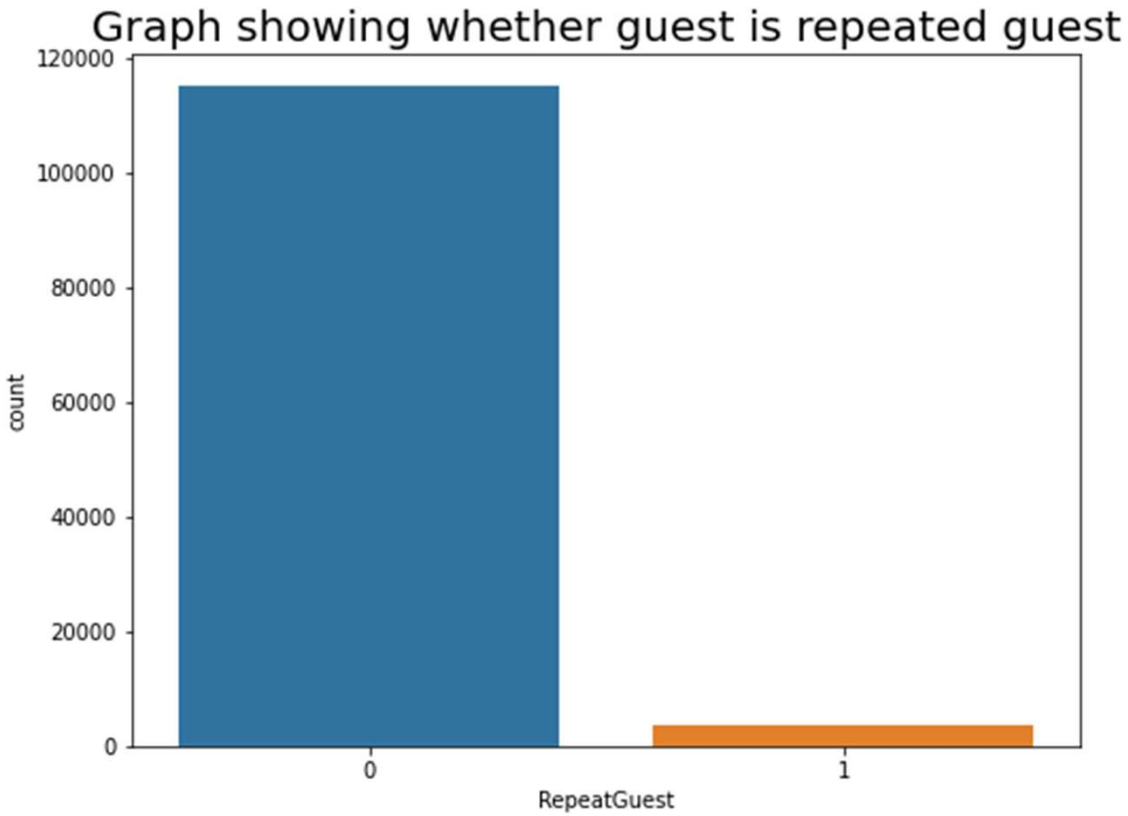



```
hotel.DepositType.value_counts(normalize=True)
```

```
No Deposit    0.876070
Non Refund    0.122567
Refundable    0.001363
Name: DepositType, dtype: float64
```

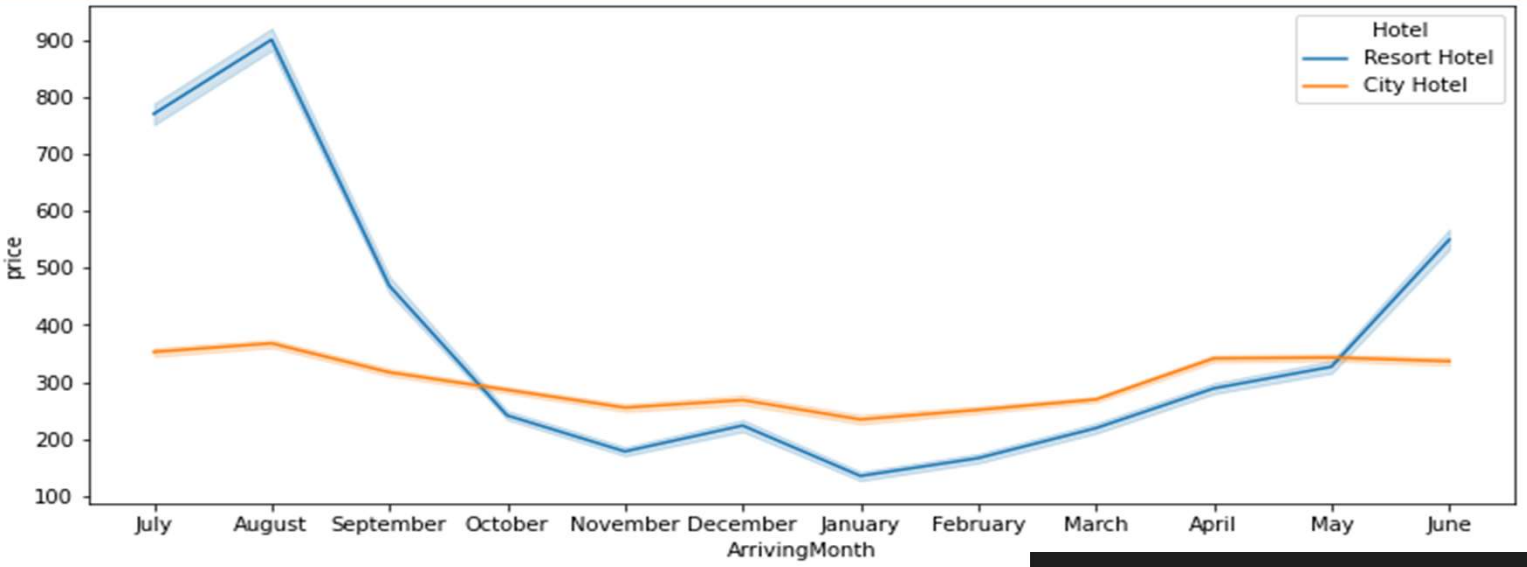
87.6% of the bookings are done without any Deposit12.2% of the bookings are done with Non-Refundable Deposit and 1.3% are Refundable Deposits

Repeat guest



- Low number of repeated guests.
- A need to target repeated guests since they have booked before.
- Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than City Hotel.

Price per Month for Resort Hotel and City Hotel



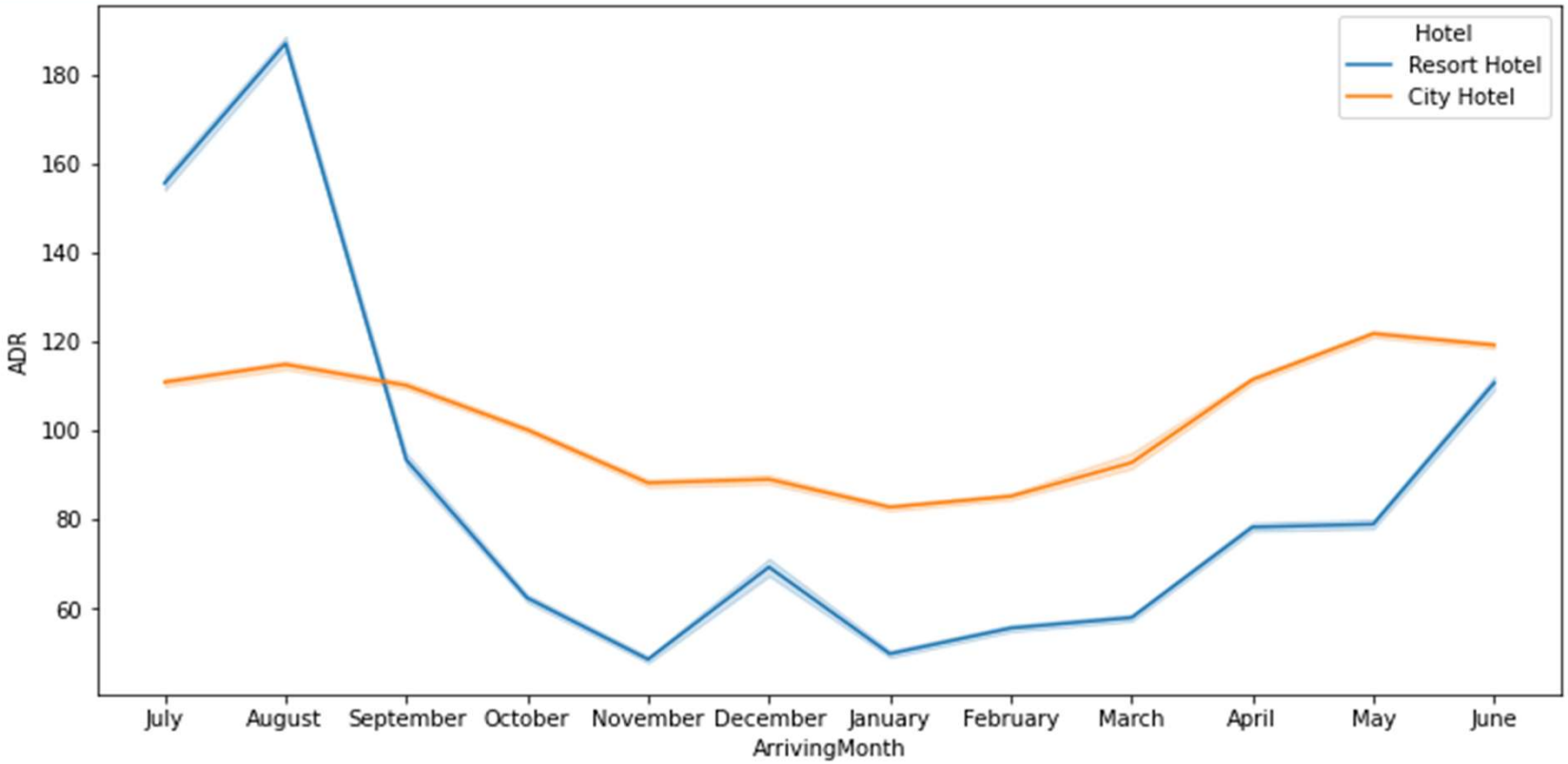
Prices of Resort Hotel are much higher and fluctuates timely during the months.

Prices of City Hotel do not fluctuate that much.

Looking into prices per month per hotel

- average daily rate = $\frac{SumOfAllLodgingTransaction}{TotalNumberOfStayingNight}$
- average daily rate per person = $\frac{ADR}{Adults+Children}$
- We will need to find out average daily rate per person

ADR: Average Daily Rate



For Resort Hotel, ADR is more expensive during July, August & September and for City Hotel, ADR is slightly more during March, April & May.

Conclusion

- Firstly, higher lead time has higher chance of cancellation. Also, history of previous cancellations increases chances of cancellation.

• Secondly, The City hotel has more guests during spring and autumn, when the prices are also highest, In July and August there are less visitors, although prices are lower. Thus, customers can get good deal on bookings in July and August in city hotel.

• Guest numbers for the Resort hotel go down slightly from June to September, which is also when the prices are highest. Thus, these months should be avoided for bookings.

• Thirdly, May to August is the peak season of bookings. Both hotels have the fewest guests during the winter.

• Fourthly, No deposit cancellations are high compared to other categories but these should not be discouraged per se as bookings in this category are also very high compared to non refundable type bookings.

Data Summary

- Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
- We also realise that the high rate of cancellations can be due high no deposit policies.
- We should also target months between May to Aug. Those are peak months due to the summer period.
- Majority of the guests are from Western Europe. We should spend a significant amount of our budget on those area.
- Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
- Set Non-refundable Rates, Collect deposits, and implement more rigid cancellation policies.
- Encourage Direct bookings by offering special discounts
- Monitor where the cancellations are coming from such as Market Segment, distribution channels, etc.

Thank You