Capstone Project

Hotel Booking EDA Analysis

Team Members Name:

1. Mayank Gupta
2. Dheeraj Patel
3. Vishal Kumar Yadav
4. Khushboo Yadav
5. Rozi Fatma

Content:

Problem Statements:

- In this project, we are analyzing the data of Hotel Bookings which contain different types of hotels data like City hotel and Resort Hotel and there are lots of information included like when will the hotel be booked, cancellation of hotel booking, date of booking, types of customers, length of stay, number of available parking spaces etc.
- In the type of Hotel Industry, so many factors help in the growth of the hotel business, which is also volatile.
- The main object behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management which can perform various campaigns to boost the business and performance.

AI

Data Preparation

Hotel Booking Analysis and its data?

- The purpose of our project was to gather and analyze detailed information about hotels in order to provide insights and estimate the profit.
- The majority of Revenue Management research on demand forecasting and prediction issues is conducted in the tourism and travel-related industries.
- We have given two hotel data sets. i.e., the resort hotel is one of the hotels, and the city hotel is the other. There are 32 columns and 119390 rows.
- With out industry-specific data, it is impossible to completely understand the requirements and peculiarities of the remaining tourism and travel sectors, such as hospitality, cruising, theme parks, etc. To help overcome this restriction, two hotel datasets with demand data are given.
- Hotels will be able to identify the issue that is causing customers to cancel their bookings, as well as the reason for the cancellations, by utilizing the predictive
- It would be fantastic if the hotel management team could identify the root cause and develop a better strategy.
- The goal of our project was to collect and analyze detailed hotel information in order to provide insights and estimate profit.

Data Summary:

Field	Description
Hotel	H1= Resort Hotel H2=City Hotel
is_cancelled	If the booking was cancelled(1) or not(0)
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number for arrival date
arrival_dat_day	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number of adults
children	Number of children
babies	Number of babies
meal	Kind of meal opted for
country	Country code
market_segment	Which segment the customer belongs to
Distribution _channel	How the customer accessed the stay corporate booking/Direct/TA.TO
is_repeated_guest	Guest coming for first time or not
previous_cancellation	Was there a cancellation before
previous_bookings	Count of previous bookings
reserved_room_type	Type of room reserved
assigned_room_type	Type of room assigned
booking_changes	Count of changes made to booking
deposit_type	Deposit type
agent	Booked through agent
days_in_waiting_list	Number of days in waiting list
customer_type	Type of customer
required_car_parking	If car parking is required
total_of_special_req	Number of additional special requirements
reservation_status	Reservation of status
reservation_status_date	Date of the specific status

Data Collection and Understanding:

After collecting data it's very important to understand your data. So we had hotel booking which has 119390 rows and 32 columns. So, let's understand the columns.

Data Description:-

Hotel : Different type of Hotels.

is_canceled : The value indicates whether or not the reservation has been cancelled.

lead_time : How far in advance the reservation was made

arrival_date_year : Year of customer arrival.

arrival_date_month : Which month of the year did the customer visit

arrival_date_week_number: Which week of the year

arrival_date_day_of_month : The month in which the customer visited the hotel.

stays_in_weekend_nights : Customer stayed or booked to stay in hotel during weekend nights.

stays_in_week_nights : The customer stayed or planned to stay in a hotel on a weekend night.

adults : Number of adults

children : number of children.

babies : Number of babies.

meal : Type of meal booked.:

country : Country of origin of customer.

market_segment : where the bookings came from.

distribution_channel : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” .

is_repeated_guest : Value indicating if the booking name was from a repeated guest (1) or not (0).

previous_cancellations : The number of previous bookings that the customer cancelled prior to the current booking.

previous_bookings_not_canceled : Number of previous bookings that were cancelled by the customer prior to the current booking.

reserved_room_type : The number of previous bookings cancelled by the customer prior to the current booking.

assigned_room_type : The code for the room type assigned to the booking. Because of this, the assigned room type may differ from the reserved room type.

***booking_changes *:** Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.

deposit_type : Indicates whether or not the customer paid a deposit to secure the reservation.

agent : The ID of the travel agency that made the reservation.

company : ID of the company/entity that made the reservation or is responsible for paying the reservation.

days_in_waiting_list : The number of days the reservation was on the waiting list before being confirmed to the customer.

customer_type : Booking type, assuming one of four categories.

***adr *:** The average daily rate is calculated by dividing the total number of staying nights by the sum of all lodging transactions.

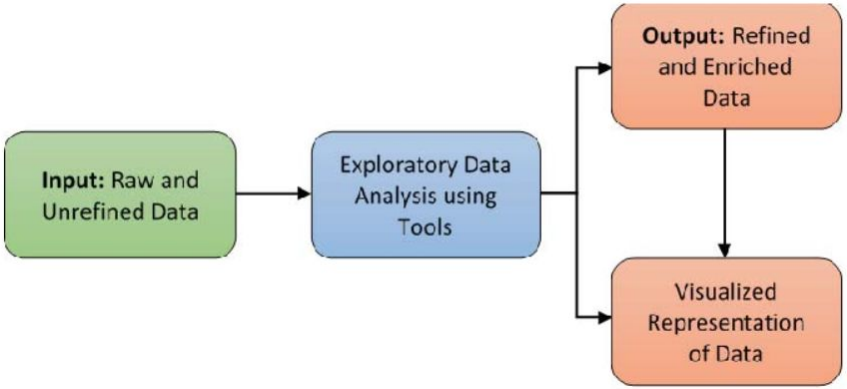
required_car_parking_spaces : The number of parking spaces needed by the customer.

***total_of_special_requests *:** The number of customer special requests (e.g. twin bed or high floor).

reservation_status : Last reservation status in one of three categories: Canceled - the customer cancelled the reservation; Check-out: the customer checked out of the hotel. No show: the customer did not check in to the hotel and informed the hotel of the reason.

reservation_status_date : The date on which the most recent status was set. This variable, in conjunction with the Reservation Status, can be used to determine when the booking was cancelled or when the customer checked out of the hotel.

Flowchart and EDA Analysis



Cleaning of Data and Manipulation

Removing of Duplicate values

Duplicate Values

```
# Dataset Duplicate Value Count
print(main_df.shape)
main_df.duplicated().sum()
```

```
(119390, 32)
31994
```

- Here we find that the total no. of duplicate values in our data is 31,994

```
[8] #We find that there are some duplicates in our data.So We will remove the duplicates.
main_df.drop_duplicates(inplace=True)
main_df.shape
```

```
(87396, 32)
```

So after removing the duplicates from the dataset,our dataset shape has 87396 rows and 32 columns.

```
[10] #we shall check the removal of duplicate values.
main_df.duplicated().sum()
```

```
0
```

Dealing with Null Values

Missing Values/Null Values

```
# Missing Values/Null Values Count
null_counts = main_df.isnull().sum()
null_counts
```

```
children                        4
babies                          0
meal                            0
country                       452
market_segment                  0
distribution_channel            0
is_repeated_guest               0
previous_cancellations          0
previous_bookings_not_canceled  0
reserved_room_type              0
assigned_room_type              0
booking_changes                 0
deposit_type                    0
agent                       12193
company                     82137
```

Dealing with Null Values Contd…

```
# Missing Values/Null Values Count
null_counts = main_df.isnull().sum()
null_counts
```

- In this, we have found that 4 columns have null values and in which the 'company' and 'agent' have highest null values. So that, we are removing these columns by using of drop method.

```
[ ] # Removing the null data
    main_df = main_df.drop(columns = ['agent', 'company'])
```

- Now we will just need to remove 488 rows with the NaN values. 488 rows out of 119390 is negligible hence I will just remove. This can be done using data.dropna(axis = 0)

```
[ ] main_df = main_df.dropna(axis=0)

    # Check to see if there are any more NaN data
    main_df.isnull().sum()
```

- Now, we have clean Data in which there is no null values present in any column so that, the analysis will be more accurate.

Converting Data Type

```
#dtypes of each column
main_df.dtypes
```

```
hotel                           object
is_canceled                      int64
lead_time                        int64
arrival_date_year                int64
arrival_date_month              object
arrival_date_week_number         int64
arrival_date_day_of_month        int64
stays_in_weekend_nights          int64
stays_in_week_nights             int64
adults                           int64
children                       float64
babies                           int64
meal                            object
country                         object
market_segment                  object
distribution_channel            object
is_repeated_guest                int64
previous_cancellations           int64
previous_bookings_not_canceled   int64
reserved_room_type              object
assigned_room_type              object
booking_changes                  int64
deposit_type                    object
days_in_waiting_list             int64
customer_type                   object
adr                            float64
required_car_parking_spaces      int64
total_of_special_requests        int64
reservation_status              object
reservation_status_date         object
dtype: object
```

We know that the data type of children can not be a float type. So that, we need to convert it into the integer.

```
[62] #Conversion of float into integer
     main_df['children'] = main_df['children'].astype('int64')
     main_df['children']
```

```
Name: children, Length: 86940, dtype: int64
```

DATA VISUALIZATION

Let’s take some insights from our data

1 What is the count of each type of Hotels ?

2 Which of the two hotels is preferred by customers, and in which year most hotels were booked?

9 When the hotel gets more guest i.e., in weekdays or weekends?

7 Which month is the most profitable for hotel bookings?

3 What is the booking rate according to the population?

4 Which form of distribution do customers prefer most?

5 Which type of hotel is mostly preferred by adults , children or babies?

6 Which type of hotel bookings are mostly cancelled?

8 Which hotel produces maximum revenue?

10 The maximum number of guests are from which country?

11 Which distribution route and market segments has given adr the most boost in terms of revenue?

12 In which month do the hotels have the highest ADR?

13 What is the reason for cancellation of bookings?

14 What is the number of repeated customer in hotel bookings?

15 Does a longer waiting period result in cancelled bookings?

What is the count of each type of Hotels ?

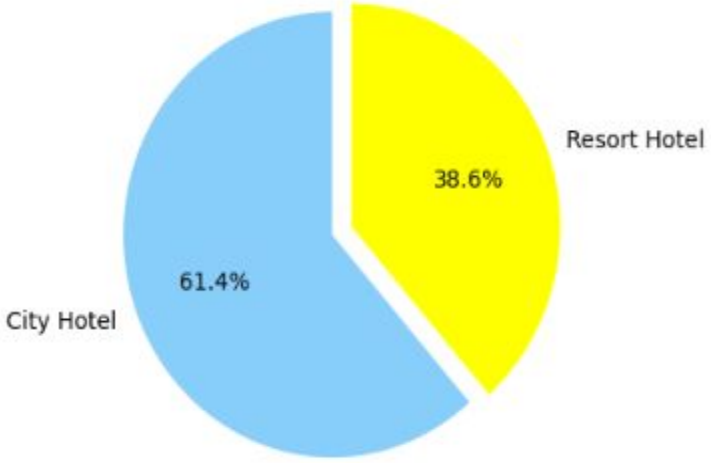
```
[99] main_df['hotel'].value_counts()

City Hotel      53417
Resort Hotel    33522
Name: hotel, dtype: int64
```

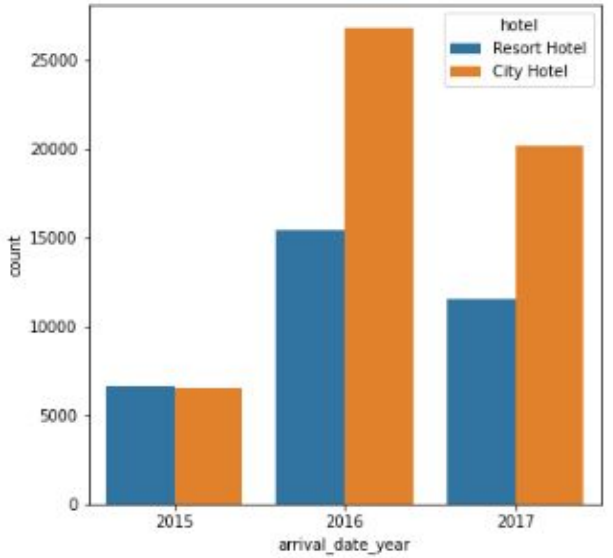
```
[101] #Counting of Hotels
hotel_type_count = main_df.groupby('hotel')['hotel'].count()
```

```
# Enlarging the pie chart
plt.rcParams['figure.figsize'] = 6,6

# Indexing labels. tolist() will convert the index to list for easy manipulation
labels = main_df['hotel'].value_counts().index.tolist()

# Convert value counts to list
sizes = main_df['hotel'].value_counts().tolist()

# As the name suggest, explode will determine how much each section is separated from each other
explode = (0, 0.1)

# Determine colour of pie chart
colors = ['lightskyblue','yellow']

plt.pie(sizes, explode = explode, labels=labels, colors=colors, autopct='%1.1f%%',startangle=90, textprops={'fontsize': 14})
```

Data Visualization of hotel type

Which of the two hotels is preferred by customers, and in which year most hotels were booked?

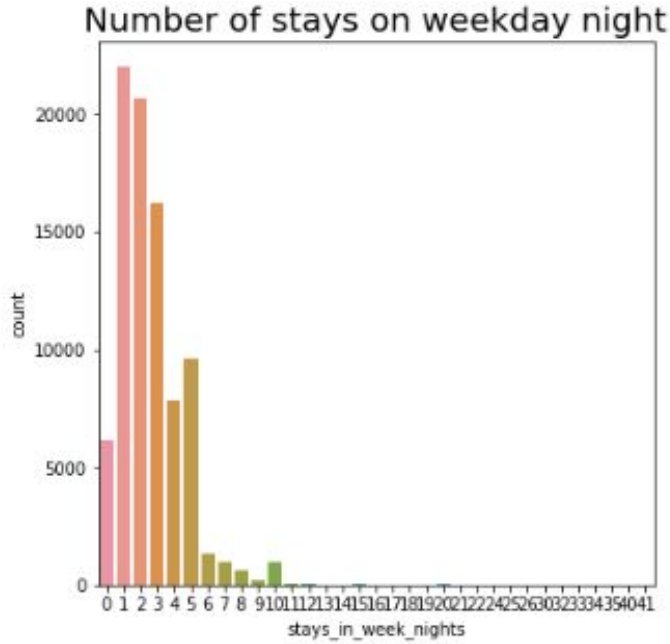




- We have found there are two types of hotel in the given dataset.
- It seems that a huge proportion of hotels was city hotel. Resort hotel tend to be on the expensive side and most people will just stick with city hotel.
- The most of the hotels are booked in the year **2016**

Data Visualization

When the hotel gets more guest i.e., in weekdays or weekends?

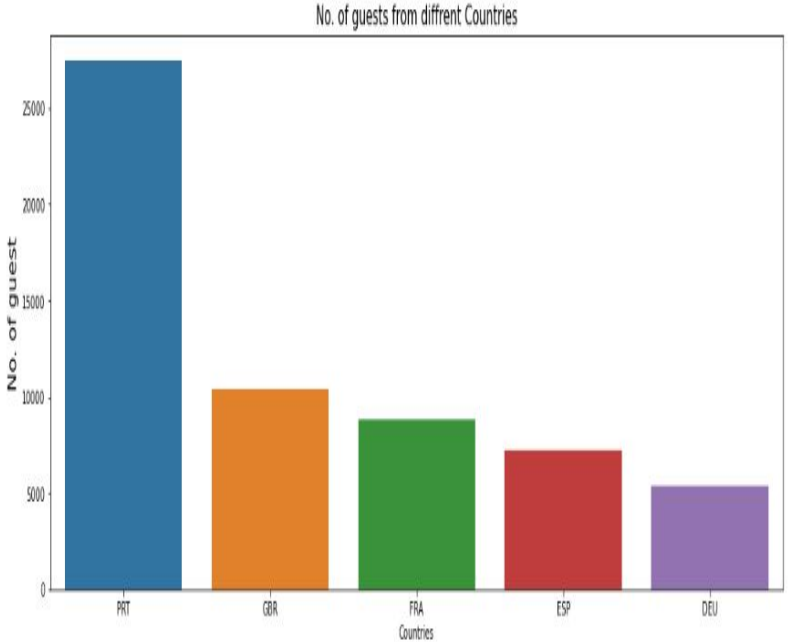


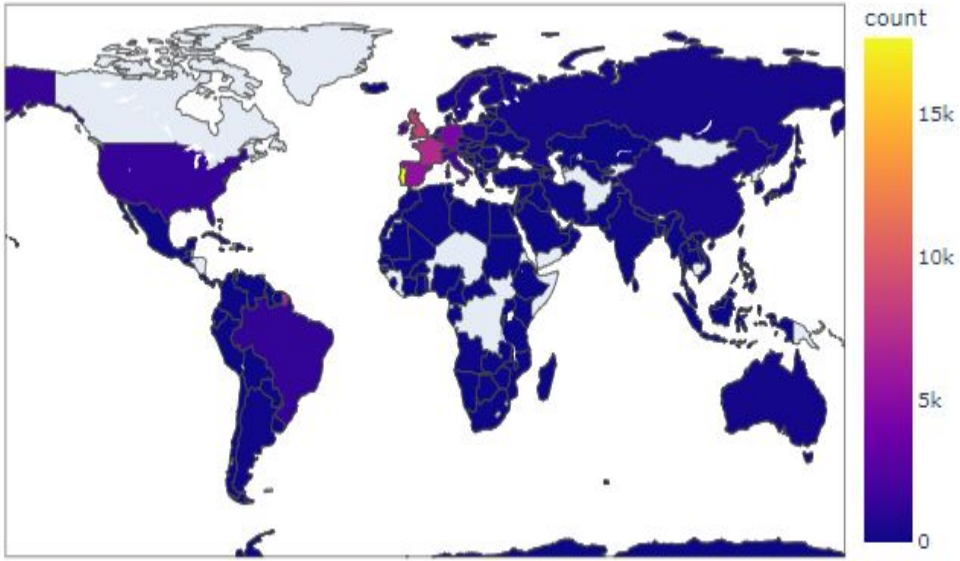
What do we see this time?

- It seems that majority of the stays are over the weekend's night.

Data Visualization

The maximum number of guests are from which country?





Observation : More than 25000 people, or the majority of the attendees, are from Portugal.

Abbreviations for nations:
PRT- Portugal, **GBR**- United Kingdom, **FRA**- France, **ESP**- Spain, **DEU** - Germany

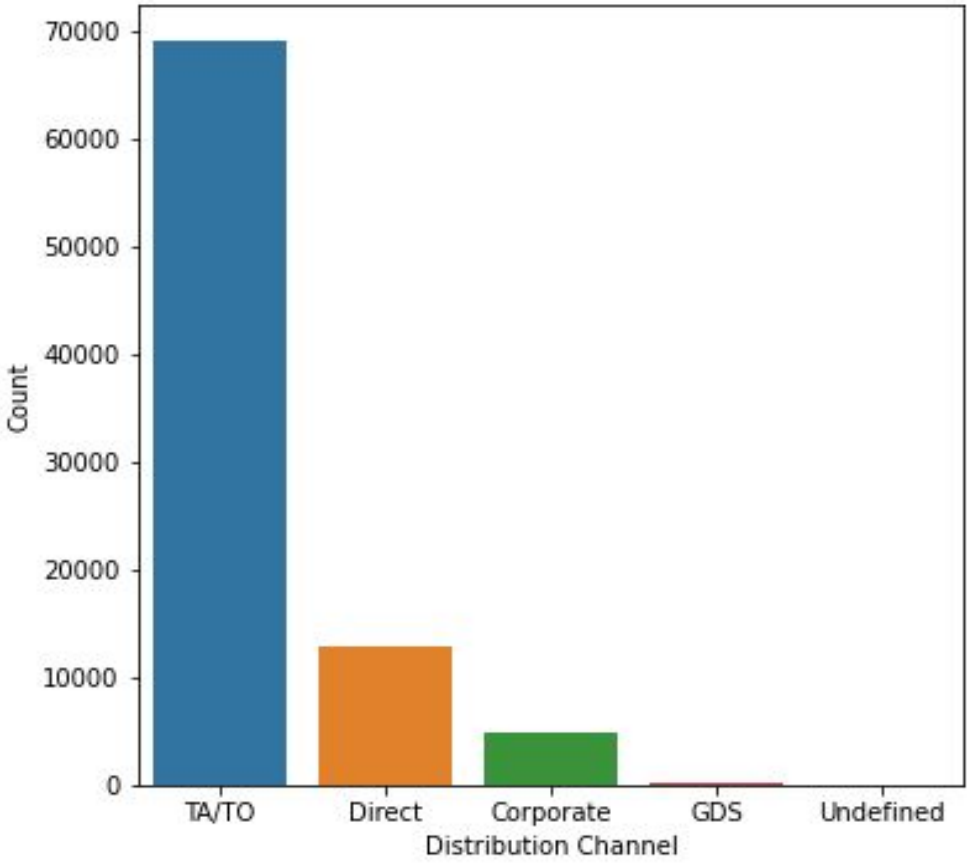
Data Visualization

Which hotel produces maximum revenue?

- According to the above figure, City hotel has more average revenue than resort hotel

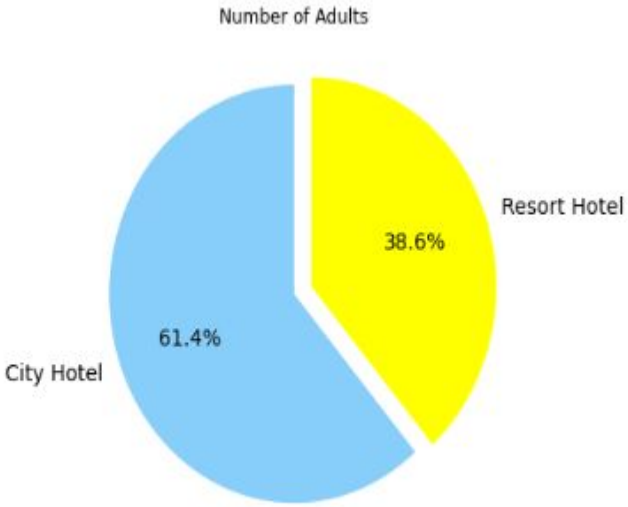
Data Visualization

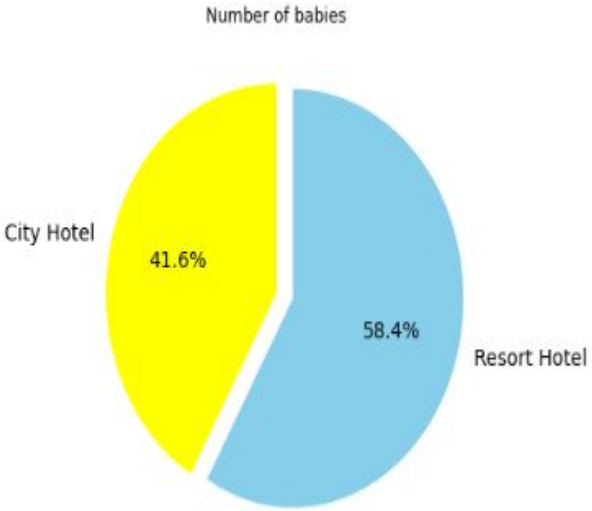
Which form of distribution do customers prefer most?



- TA/TO are the customers' chosen distribution channels.
- In order to grow their business, hotels might partner with these agents and operators or promote using them as a medium.

What is the booking rate according to the population?

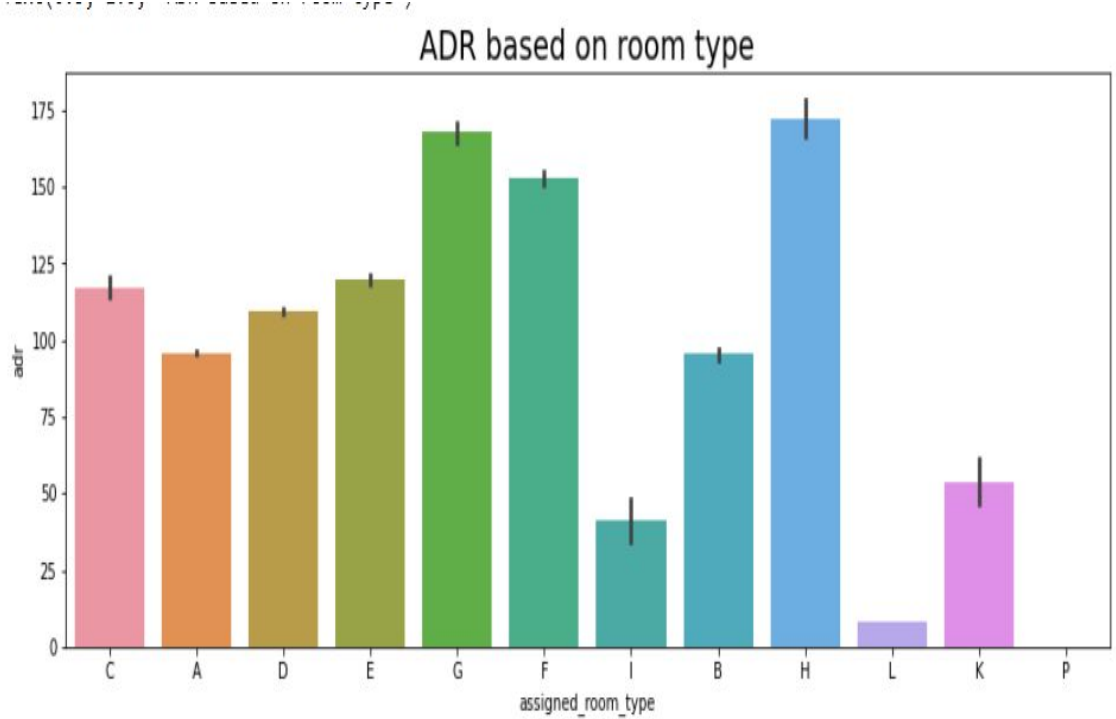




- Resort hotels are better choice for large families having babies.
- It seems that majority of the visitors who travel in pair, prefer City hotels.

Data Visualization

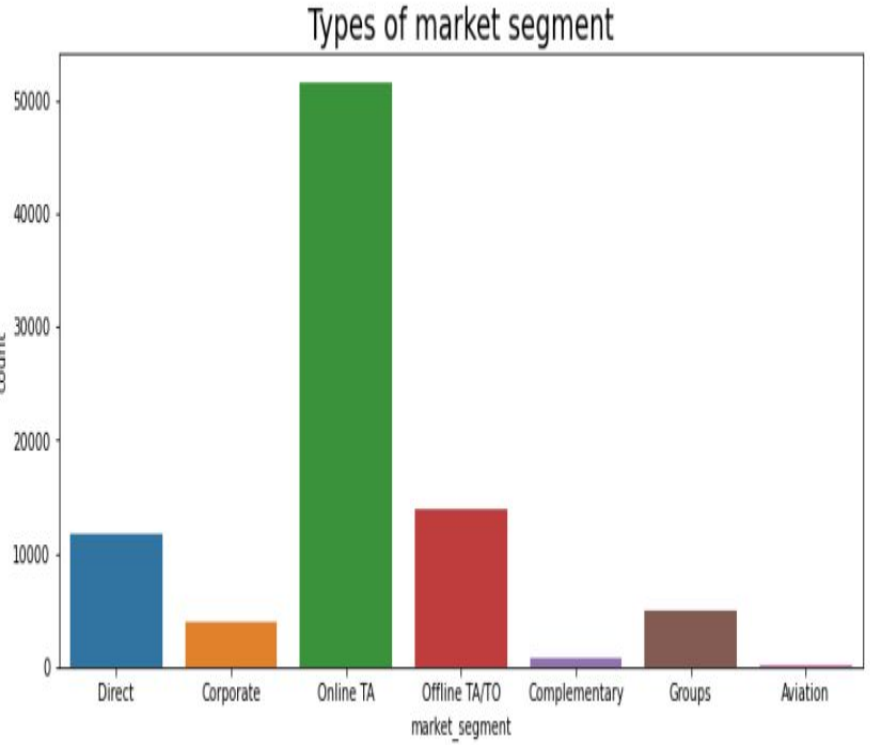
Which room type has the highest average daily rate?

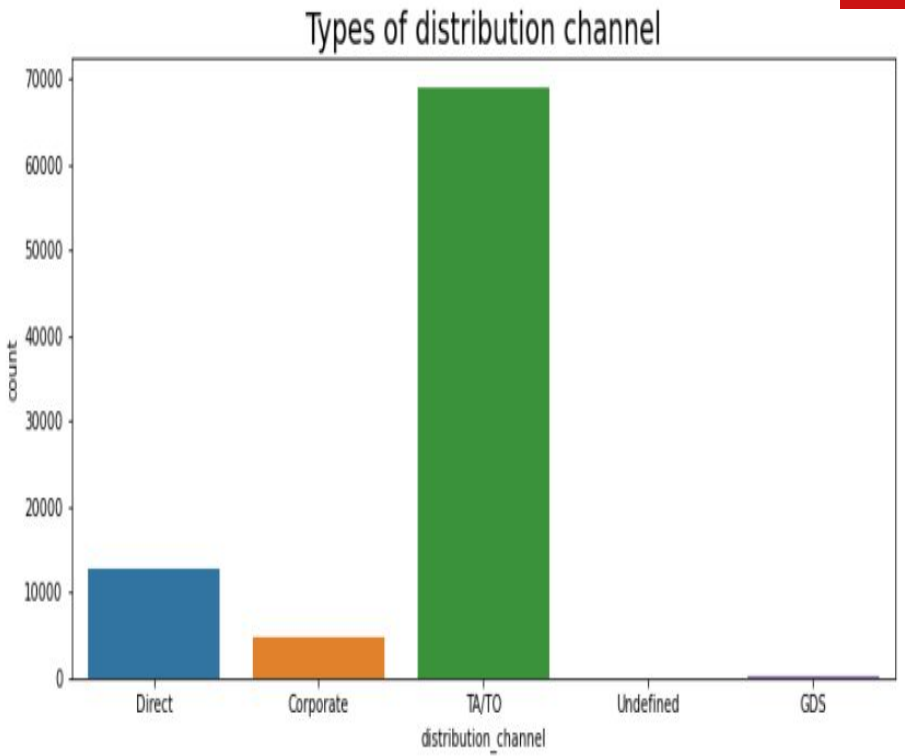


H type has the highest Average daily rate followed by G type and F type

Data Visualization

From which Channel the guest come in the hotel location?

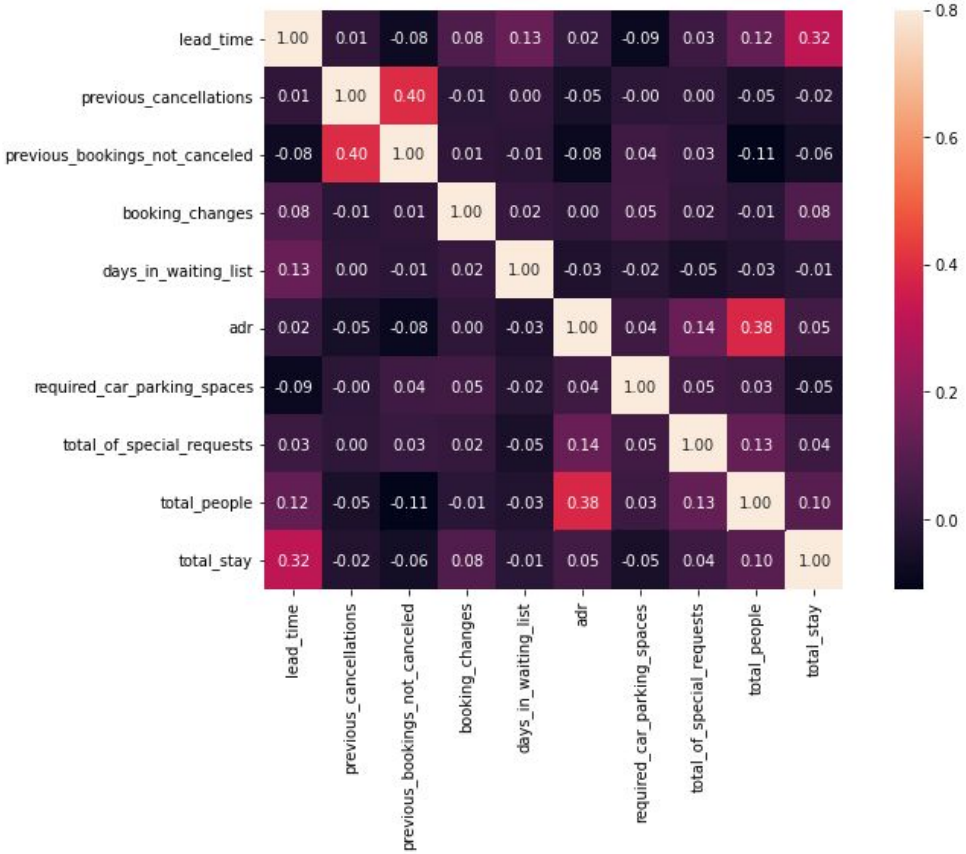




- Majority of the distribution channels and market segments involve travel agencies (online or offline).
- We can target our marketing area to be on these travel agencies website and work with them since majority of the visitors tend to reach out to them.

Data Visualization

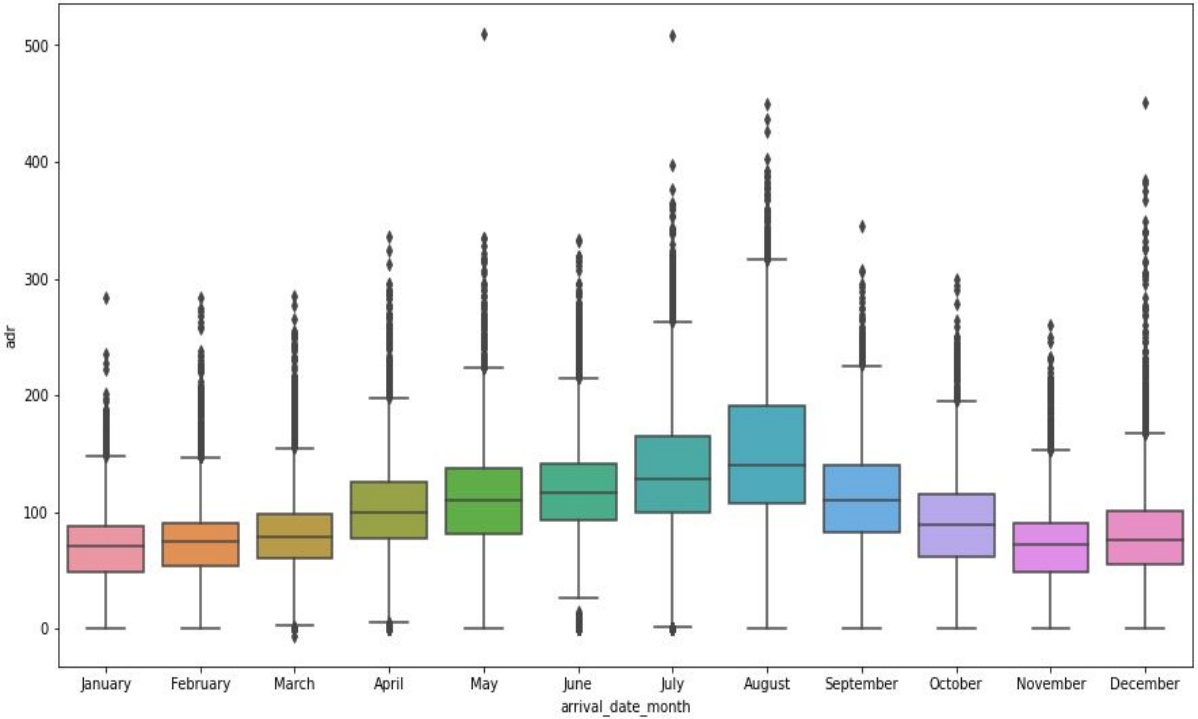
What will give after finding the correlation between the numerical data ?



- Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the the actual arrival.
- adr is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.

Data Visualization

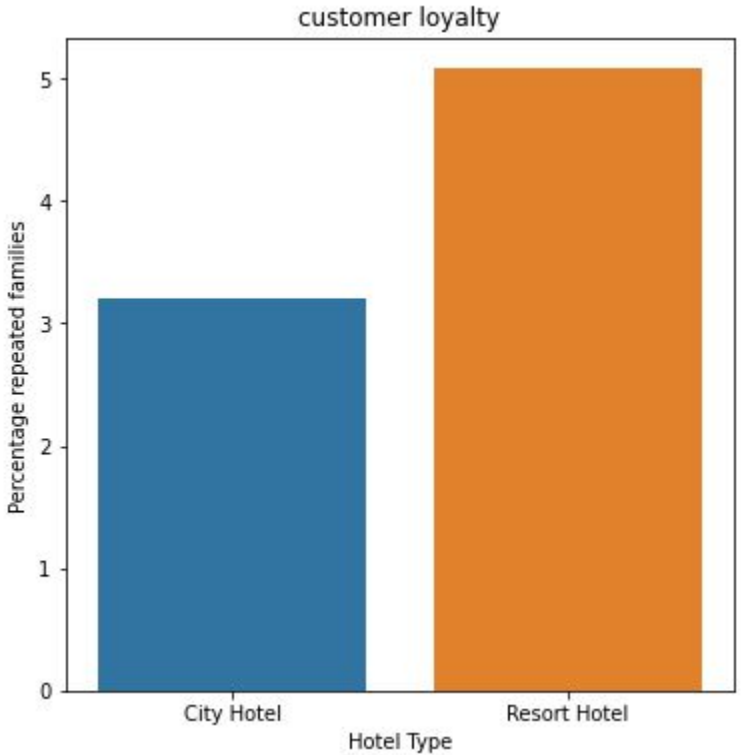
Which month generated more revenue across the year 2015, 2016, and 2017?



Avg adr rises from beginning of year upto middle of year and reaches peak at **August** and then lowers to the end of year. But hotels do make some good deals with high adr at end of year also.

Data Visualization

Which hotel has a higher rate of returning customers?



From the above graph it is clear that highest rate of returning customers are from the resort hotel.

Conclusion

With this whole analysis we found the following points:

- Resort hotels are more expensive compared to the City hotels. A huge portion of the hotels is City hotel.
- Resort hotels are better choice for large families.
- It seems that majority of the visitors who travel in pair, prefer City hotels.
- In the year **2016**, highest hotel bookings were registered.
- In the month, **July** to **August** highest hotel bookings were found.
- It seems that majority of the stays are over the weekday night.
- In the hotel bookings we have found that a huge number of visitors are coming from western europe, namely France, UK and **Portugal** are among the highest.
- Majority of the distribution channels and market segments involve travel agencies (online or offline).
- We observed that the high rate of cancellations is due to 'no deposit' policy.
- We need to focus on that customers who visited first time in the hotel but not booking the hotel again.

kulfyapp.com

Thank You