Capstone Project - I

EDA on Hotel Booking Data By

Siddharth Ray

Avishek Patra

(Cohort Istanbul)

AlmaBetter

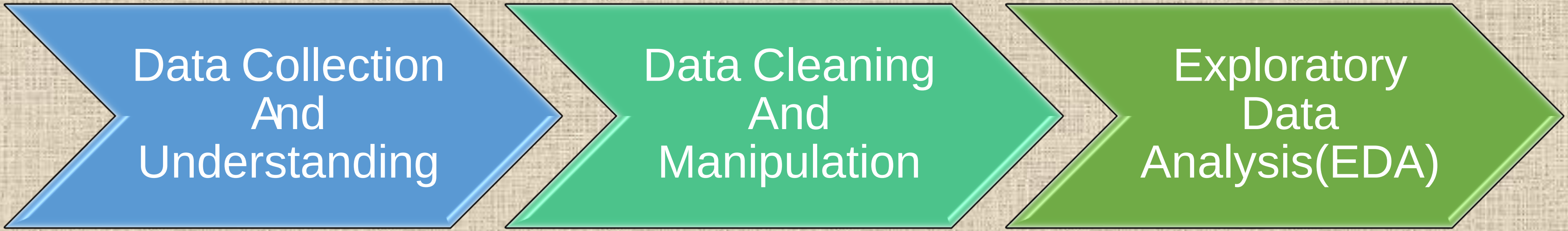
Problem Statement

- In this project we will be analyzing Hotel Booking data. This data set contains booking information for a City hotel and a Resort hotel along with information on various booking criteria such as booking season, pricing data, length of stay, number of adults, children and babies, parking spaces, market segment and many more.

- Primary objective is to explore and inspect the dataset and discover important features using Exploratory Data Analysis that can govern bookings and help hotels penetrate deep into market, thereby attracting more customers.

- Secondary objective is to help the customers in deciding the best period to visit places while availing low accommodation cost benefits.

Work Flow



Data Collection And Understanding

Data Cleaning And Manipulation

Exploratory Data Analysis(EDA)

- EDA will be divided into following 3 analysis:

1. Univariate analysis: Univariate analysis is the simplest of the three analyses where the data, you are analyzing is only one variable.

2. Bivariate analysis: Bivariate analysis is where you are comparing two variables to study their relationships.

3. Multivariate analysis: Multivariate analysis is similar to bivariate analysis but you are comparing more than two variables.

Data Collection and Understanding

After collecting the 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 these 32 columns.

Data Description:

- **hotel** : Resort Hotel or City Hotel
- **is_cancelled** : Value indicating if the booking was cancelled (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

Data Collection and Understanding

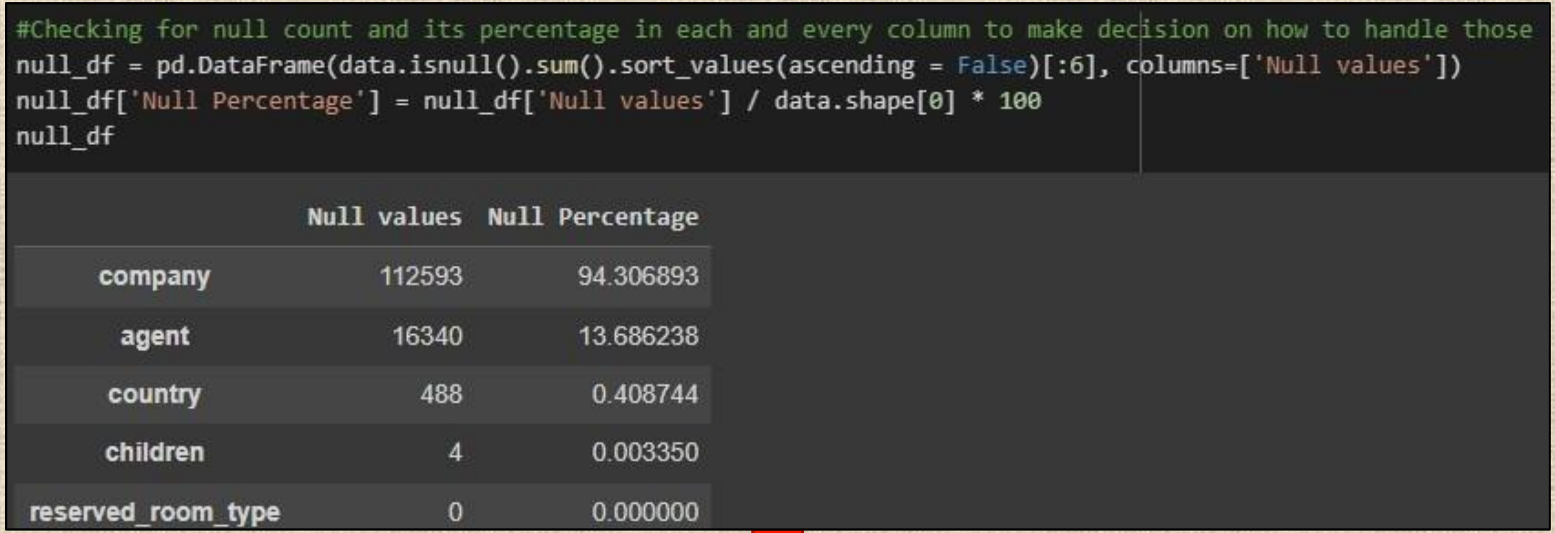
- **adults** : Number of adults
- **children** : Number of children
- **babies** : Number of babies
- **meal** : Type of meal booked
- **country** : Country of origin
- **market_segment** : Market segment designation. (TA/TO)
- **distribution_channel** : Booking distribution channel. (TA/TO)
- **is_repeated_guest** : is a repeated guest (1) or not(0)
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
- **previous_bookiings_not_cancelled** : 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

Data Collection and Understanding

- **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 that booking was in the waiting list before it was confirmed to the customer
- **customer_type** : Type of customer. Contact, Group, Transient, Transient party
- **adr** : Average Daily Rate is 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_request** : Number of special requests made by the customer(e.g., twin bed or high floor)
- **reservation_status** : Reservation last status

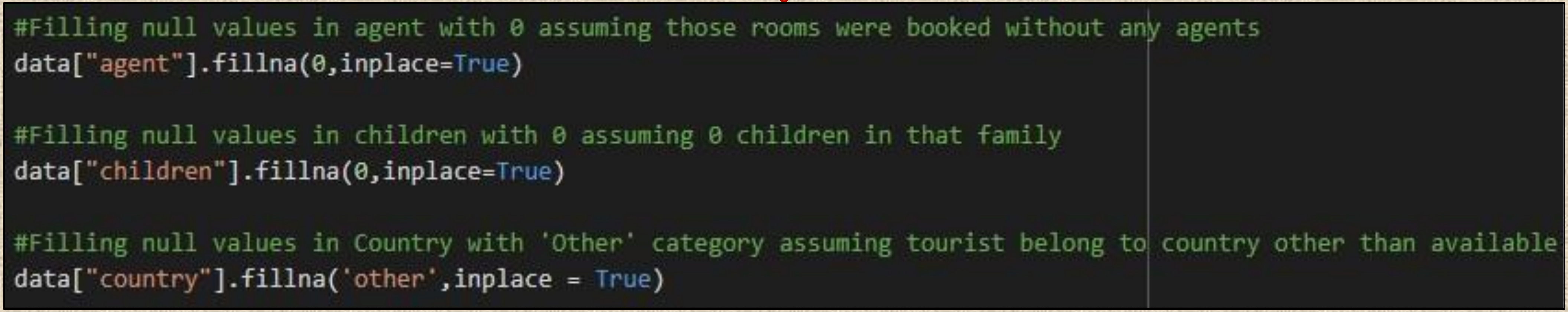
Data Cleaning and Manipulation

I. **Handling Null values:** Columns company, agent, country and children



```python
#Checking for null count and its percentage in each and every column to make decision on how to handle those
null_df = pd.DataFrame(data.isnull().sum().sort_values(ascending = False)[:6], columns=['Null values'])
null_df['Null Percentage'] = null_df['Null values'] / data.shape[0] * 100
null_df
```

	Null values	Null Percentage
company	112593	94.306893
agent	16340	13.686238
country	488	0.408744
children	4	0.003350
reserved_room_type	0	0.000000



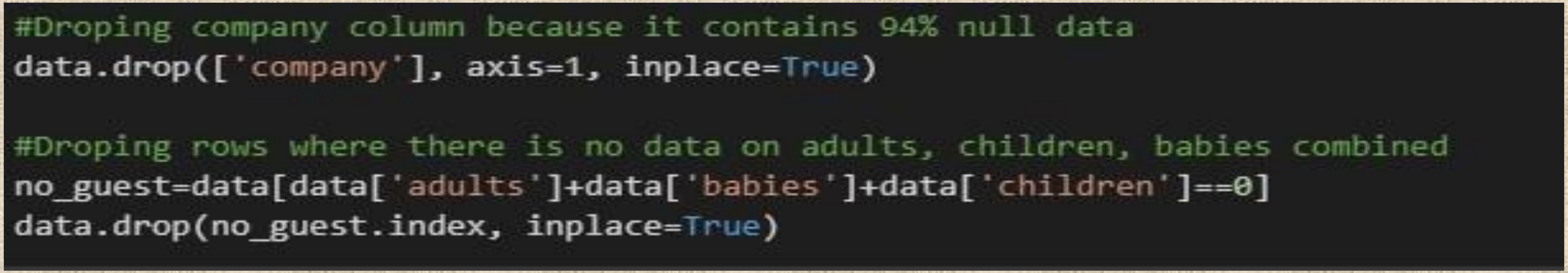
```python
#Filling null values in agent with 0 assuming those rooms were booked without any agents
data["agent"].fillna(0,inplace=True)

#Filling null values in children with 0 assuming 0 children in that family
data["children"].fillna(0,inplace=True)

#Filling null values in Country with 'Other' category assuming tourist belong to country other than available
data["country"].fillna('other',inplace = True)
```

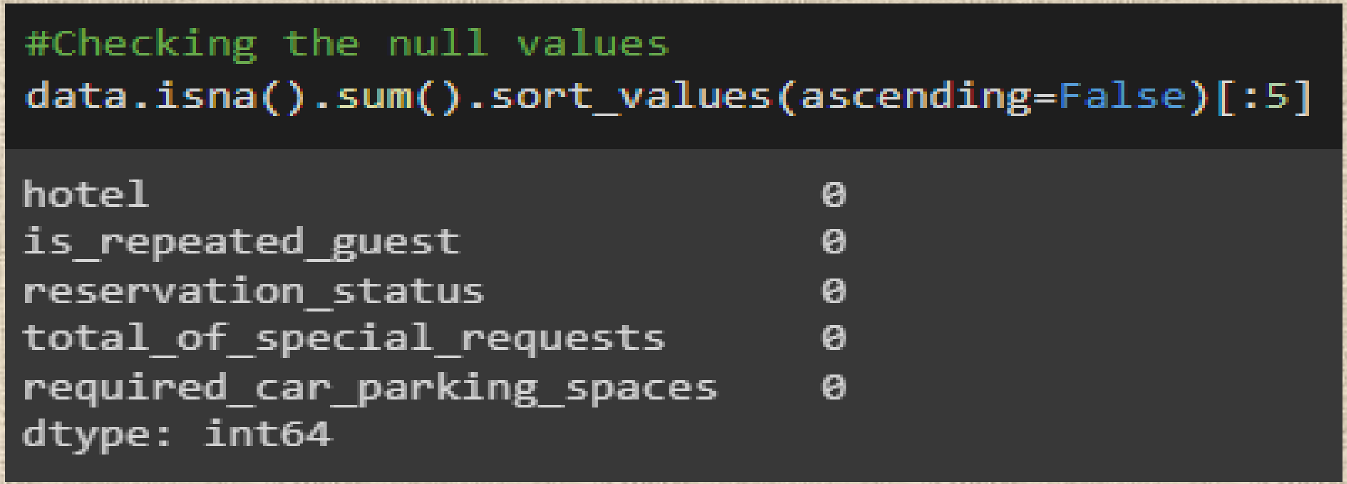
Data Cleaning and Manipulation

II. **Dropping irrelevant rows and columns**



```
#Droping company column because it contains 94% null data
data.drop(['company'], axis=1, inplace=True)

#Droping rows where there is no data on adults, children, babies combined
no_guest=data[data['adults']+data['babies']+data['children']==0]
data.drop(no_guest.index, inplace=True)
```

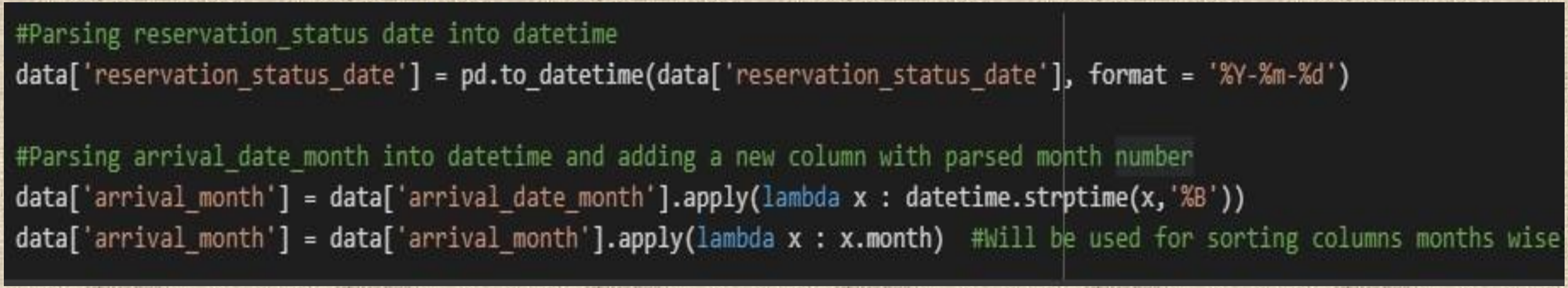


```
#Checking the null values
data.isna().sum().sort_values(ascending=False)[:5]
```

```
hotel                          0
is_repeated_guest              0
reservation_status             0
total_of_special_requests      0
required_car_parking_spaces    0
dtype: int64
```

Data Cleaning and Manipulation

III. Parsing date in string to date time format

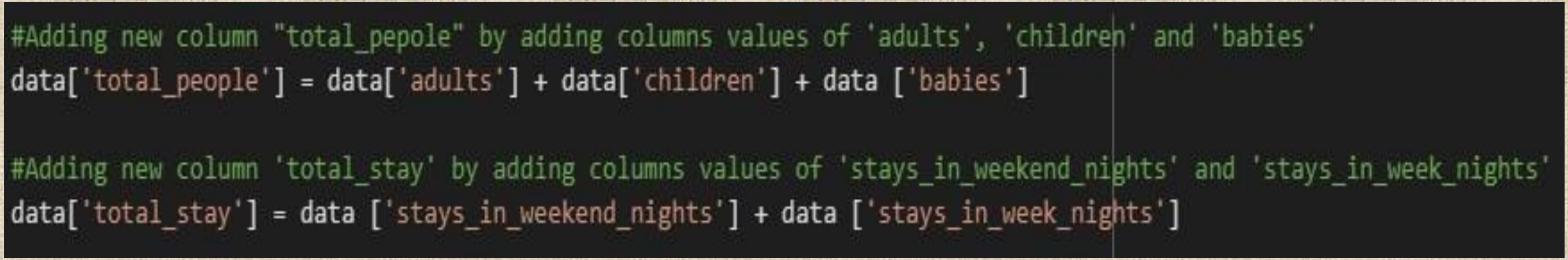


```
#Parsing reservation_status date into datetime
data['reservation_status_date'] = pd.to_datetime(data['reservation_status_date'], format = '%Y-%m-%d')

#Parsing arrival_date_month into datetime and adding a new column with parsed month number
data['arrival_month'] = data['arrival_date_month'].apply(lambda x : datetime.strptime(x,'%B'))
data['arrival_month'] = data['arrival_month'].apply(lambda x : x.month)  #Will be used for sorting columns months wise
```

IV. Feature Engineering

a. ‘total_people’ = total of adults, children and babies
b. ‘total_stay’ = total of weekend nights and weekdays nights

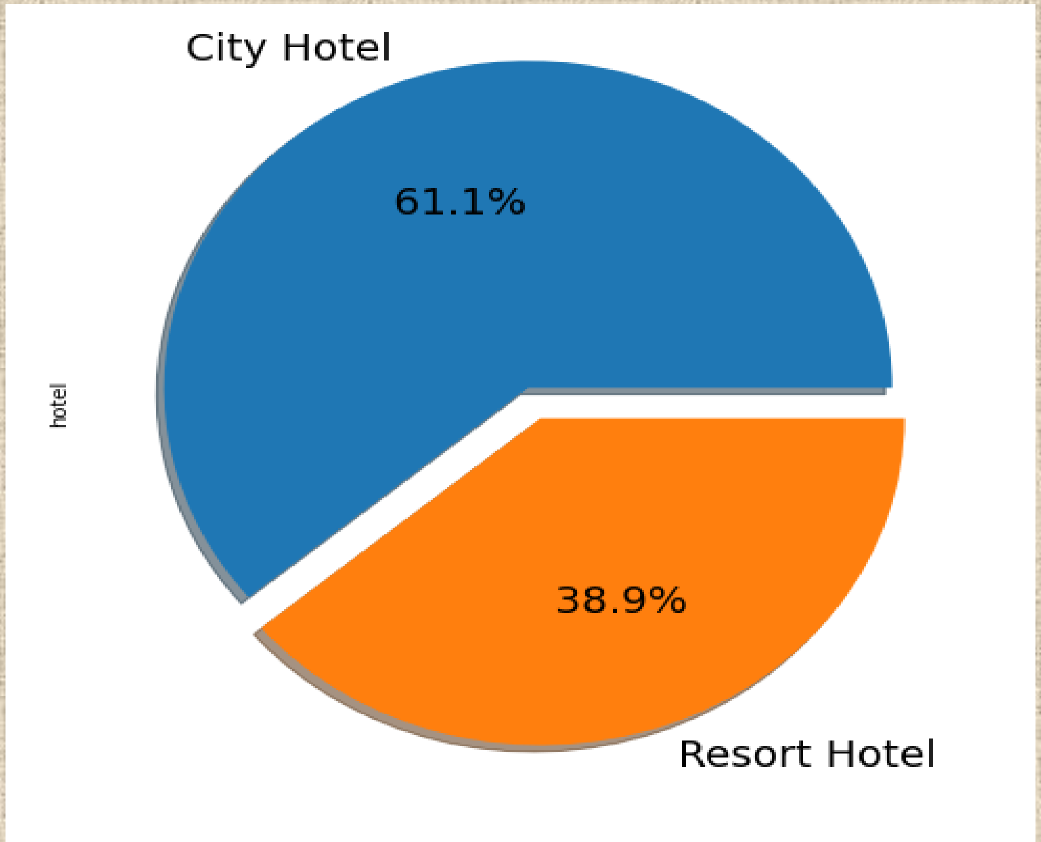


```
#Adding new column "total_pepole" by adding columns values of 'adults', 'children' and 'babies'
data['total_people'] = data['adults'] + data['children'] + data ['babies']

#Adding new column 'total_stay' by adding columns values of 'stays_in_weekend_nights' and 'stays_in_week_nights'
data['total_stay'] = data ['stays_in_weekend_nights'] + data ['stays_in_week_nights']
```

❑ Exploratory Data Analysis (EDA):

Which type of hotel is mostly preferred by the guests?

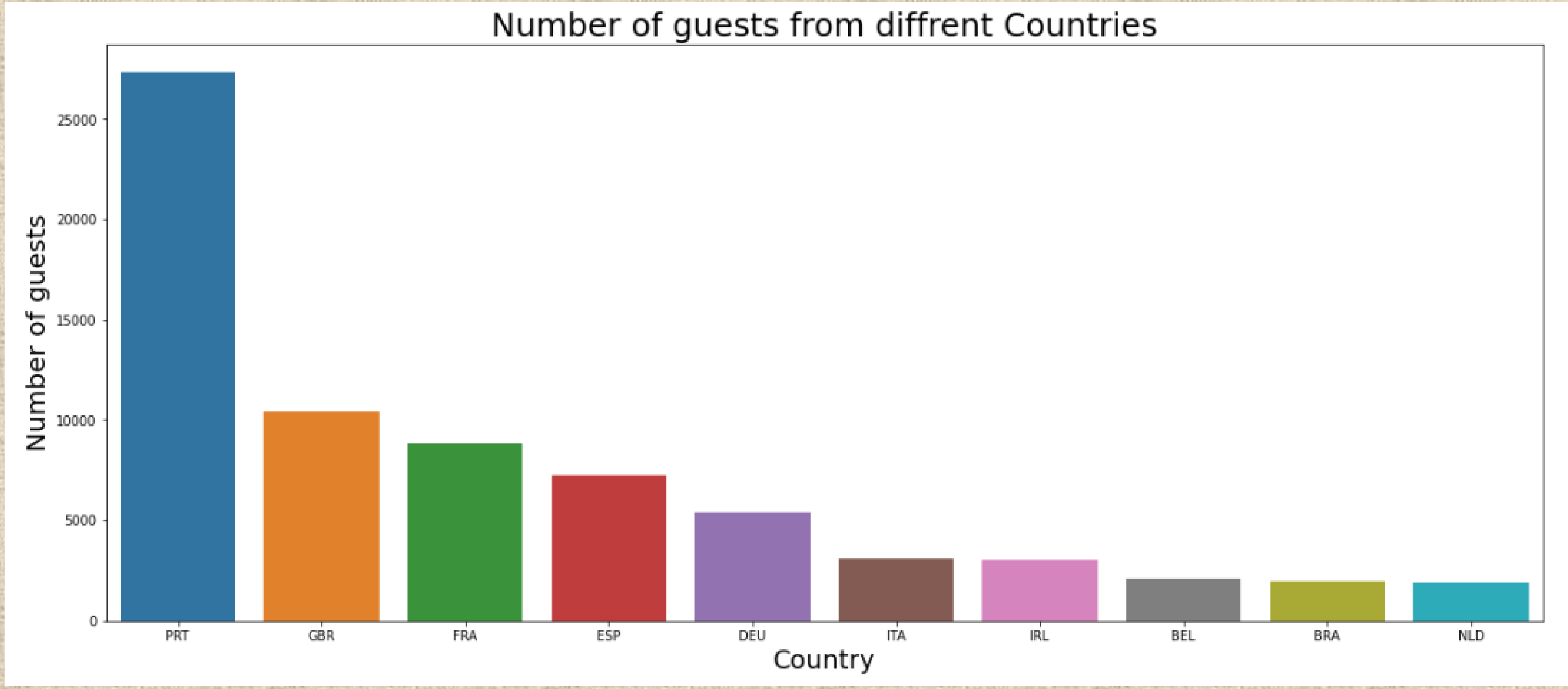


INFERENCE:

- Majority of the guest prefer City Hotel over Resort Hotel
- 2/3rd of total guest prefer City Hotel

❑ Exploratory Data Analysis (EDA):

From which country most guests are coming?

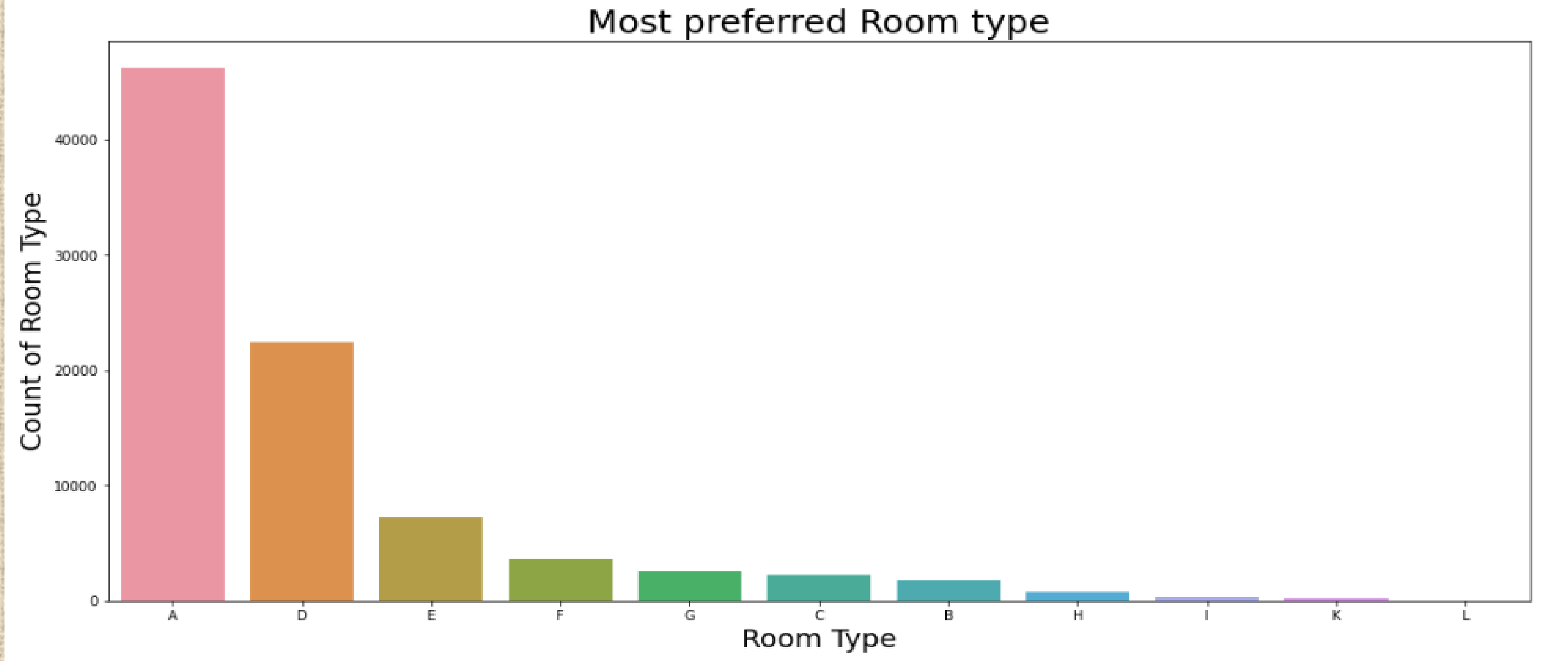


INFERENCE:

- From plotted bar plot, its evident that most of guest visiting the City hotels and Resort hotels are from Portugal and other European countries namely Britain, France, Spain and Germany. Among which Portugal takes the lion’s share with more than 25000 customers.

❑ Exploratory Data Analysis (EDA):

Which is the most preferred room type by the customers?

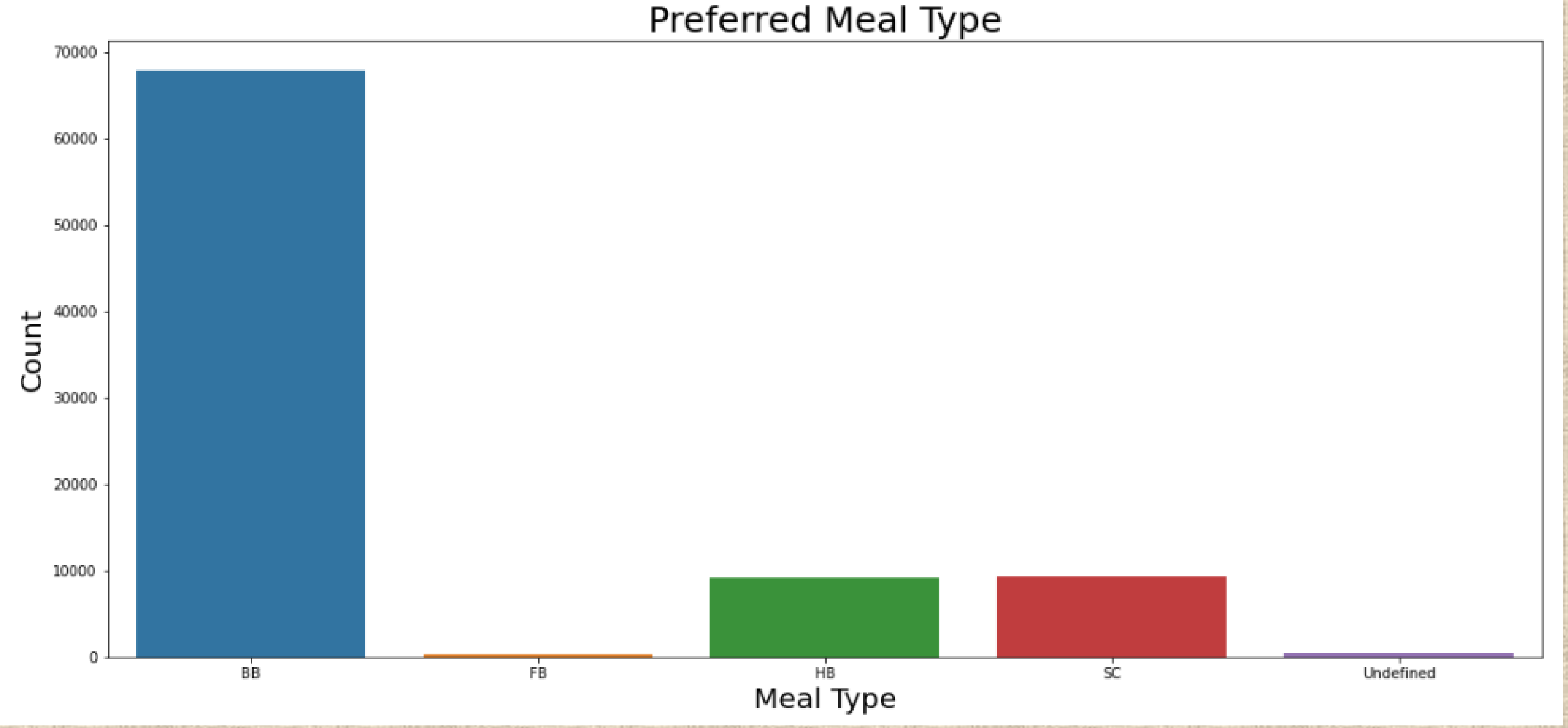


INFERENCE:

- Most demanded room type is A, followed by D and E.

❑ Exploratory Data Analysis (EDA):

Which type of food is mostly preferred by the guests?

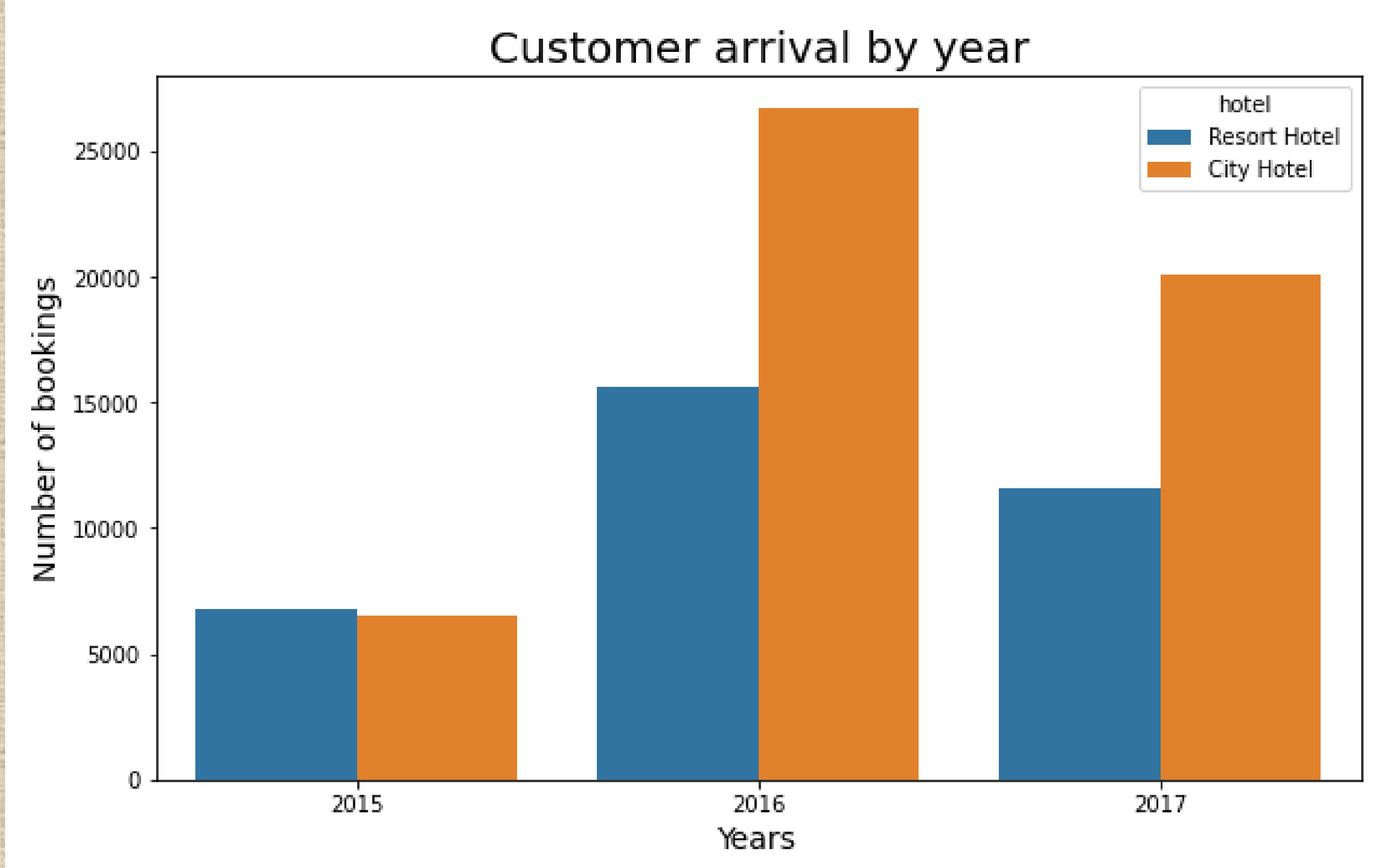


INFERENCE:

- The most preferred meal type by the guests is BB(Bed and Breakfast)
- HB- (Half Board) and SC- (Self Catering) are equally preferred.

❑ Exploratory Data Analysis (EDA):

Which year had the highest bookings?

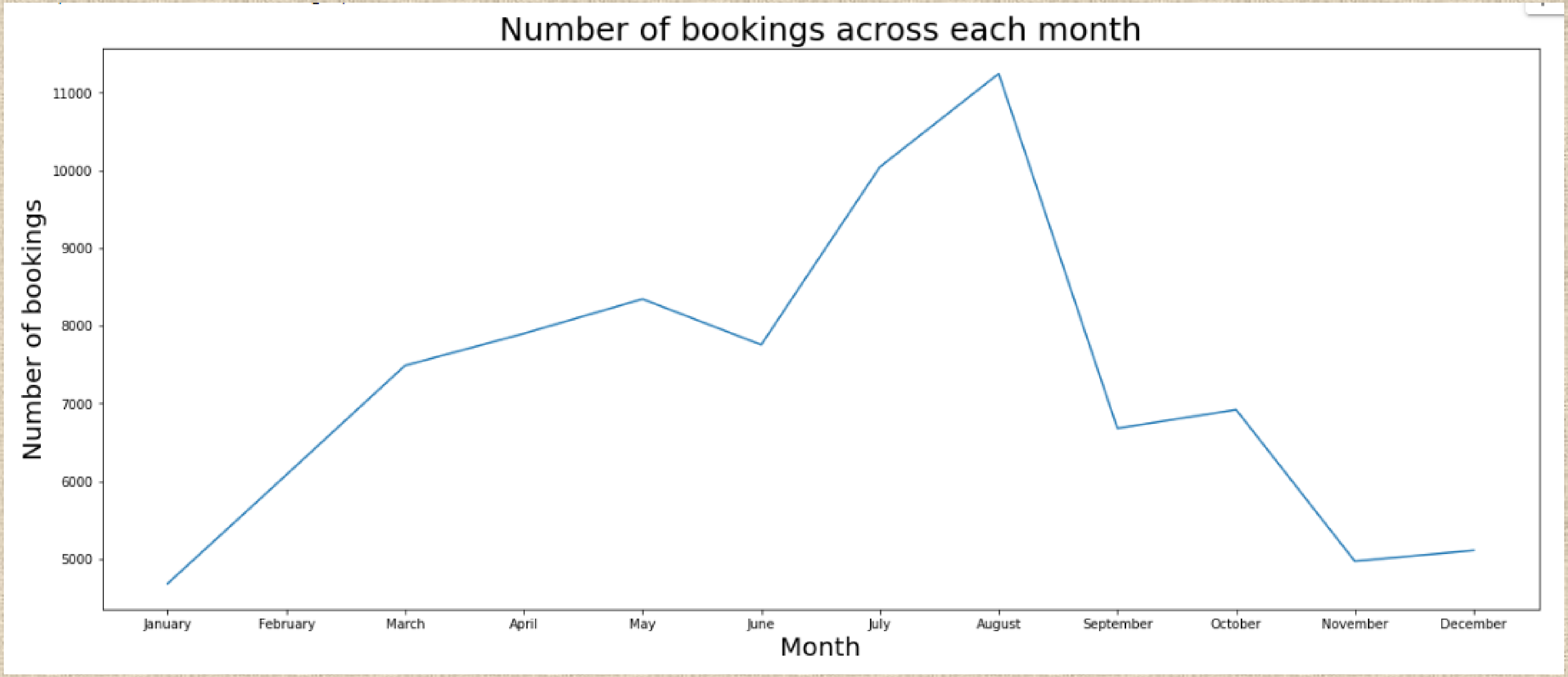


INFERENCE:

- As we can see that 2016 was the year where number of hotel booking was highest followed by total bookings in 2017 and 2015.
- Overall City hotel had the highest number of bookings compared with Resort Hotel.

❑ Exploratory Data Analysis (EDA):

In which month most of the bookings happened?

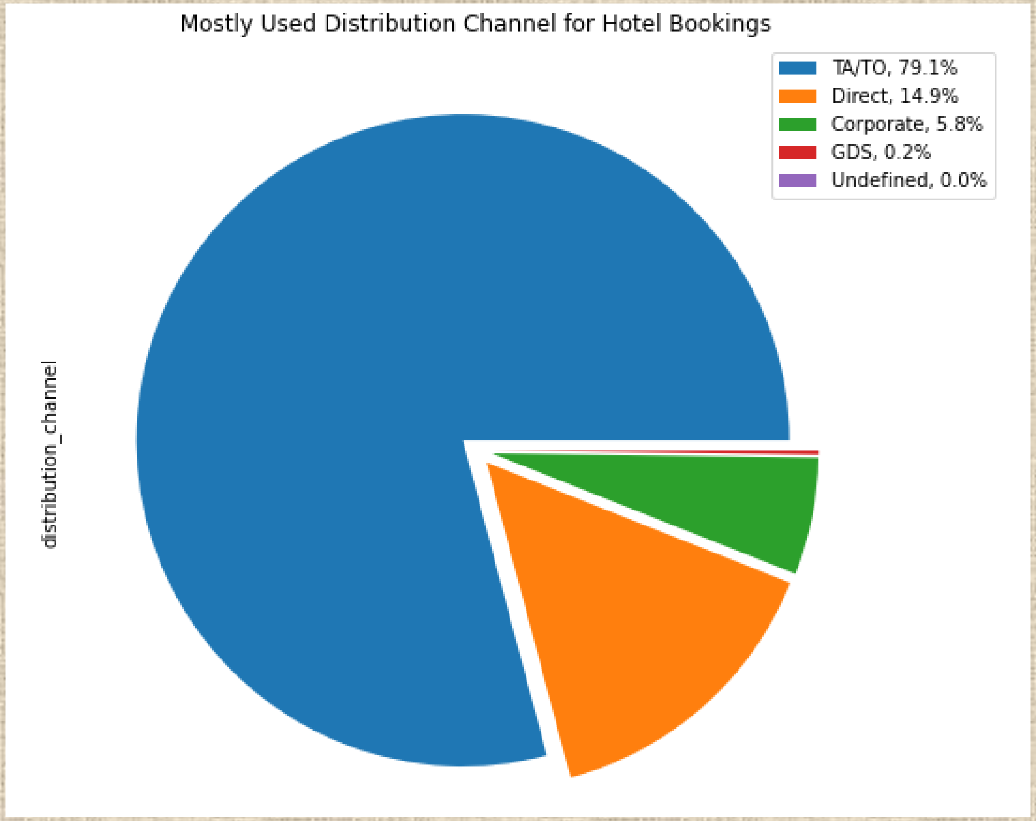


INFERENCE:

- Peak visiting season is from mid June to August because of summer breaks in Europe.
- Off season is from November to February because of cold weather throughout Europe.

❑ **Exploratory Data Analysis (EDA):**

Which Distribution channel is mostly used for hotel bookings?

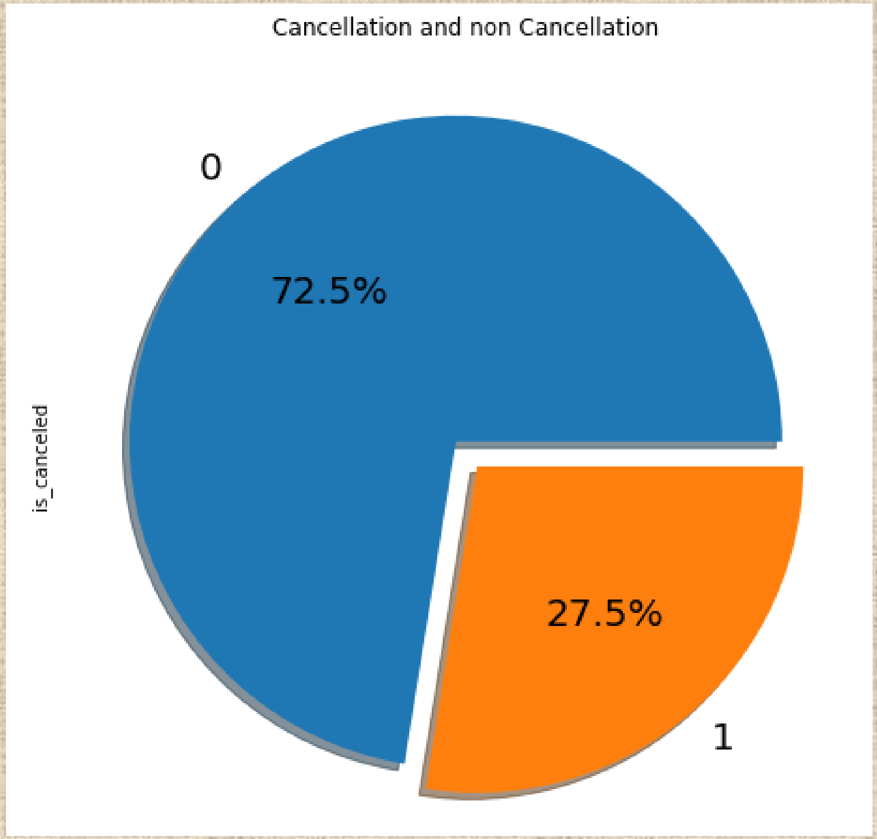


INFERENCE:

- As we can see, Resort hotel and City hotel are getting most of bookings from travel agency and tour operators. May be in future they will be monopolize the entire booking channel.

❑ Exploratory Data Analysis (EDA):

What is the percentage of cancellation?

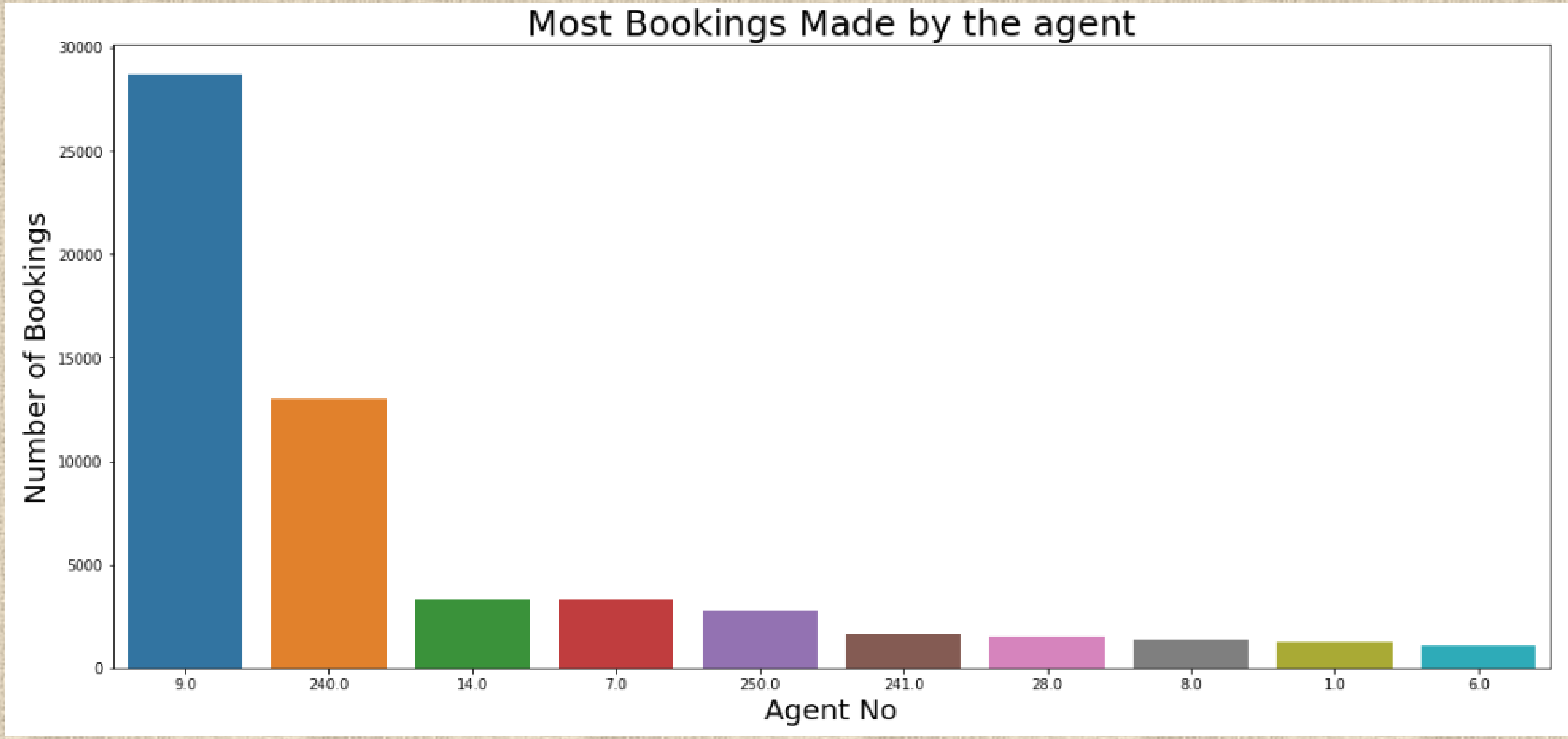


INFERENCE:

- 27.5 % of the bookings were cancelled, while 72.5% were not cancelled.

❑ Exploratory Data Analysis (EDA):

Which Agent made the most bookings?

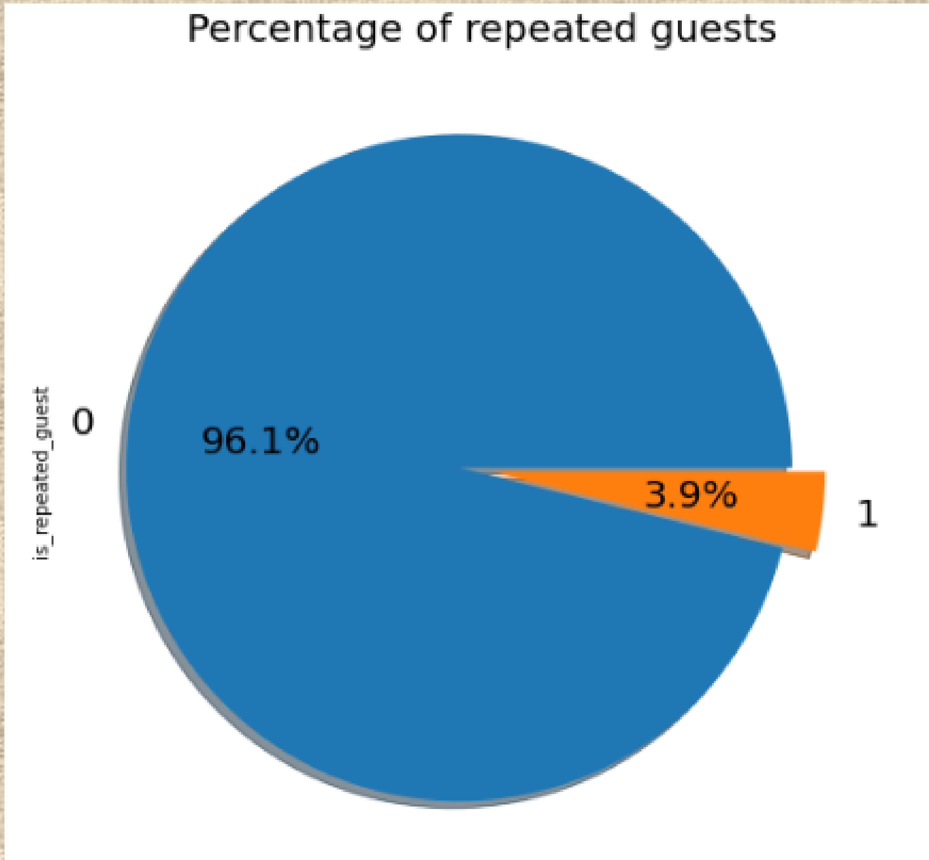


INFERENCE:

- Agent ID No. 9.0 made most of the bookings, followed by 240.0

❑ Exploratory Data Analysis (EDA):

What is the percentage of repeated guests?

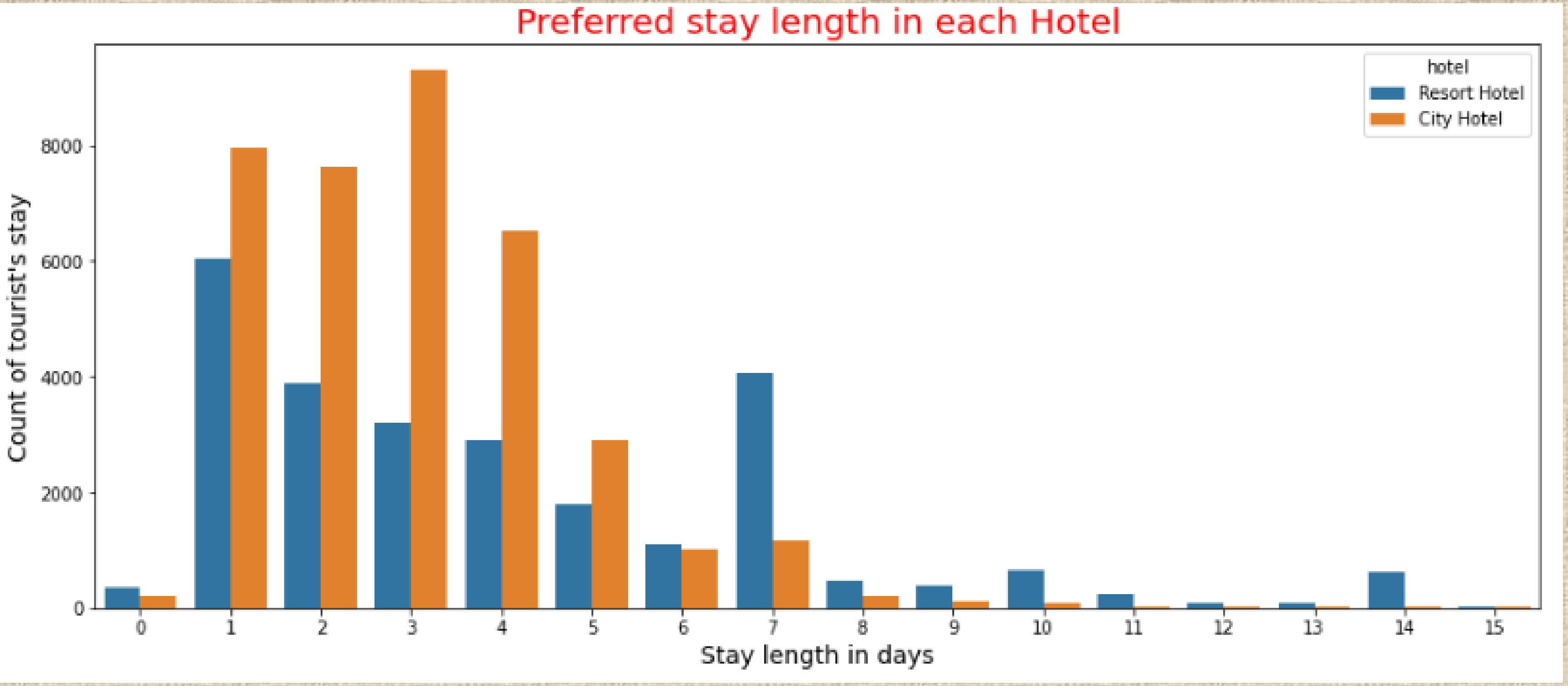


INFERENCE:

- Repeated guests are very few which only comprise of **3.9 %**.

❑ Exploratory Data Analysis (EDA):

What is the Optimal length of stay in both types of hotels ?

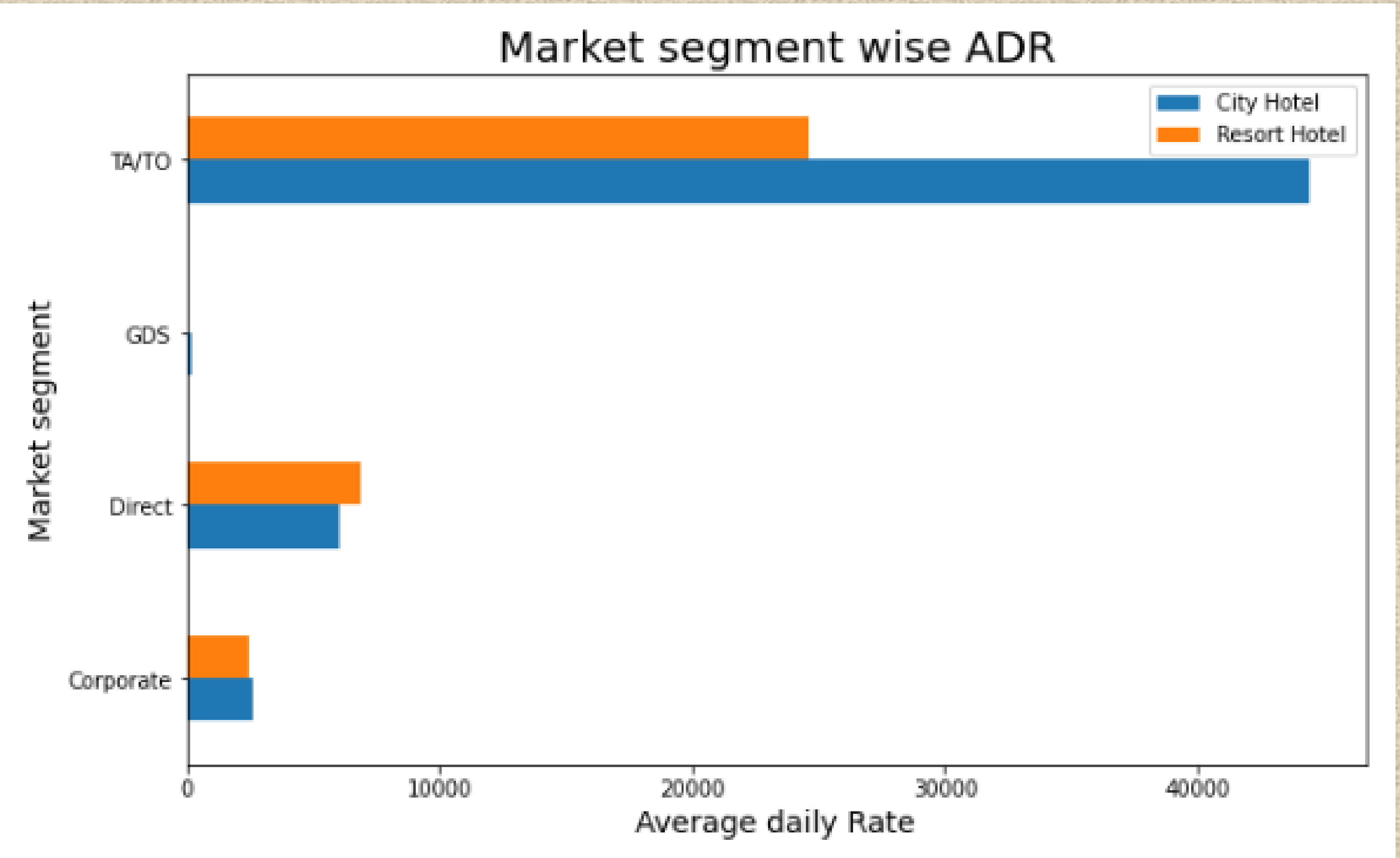


INFERENCE:

- Guest prefer 1-4 days when staying in City Hotels.
- Guest prefer 1-4 days when staying in Resort Hotels as well, while 7 days stay is also a popular choice among guests.

❑ Exploratory Data Analysis (EDA):

Which distribution channel contributed more to ADR in order to increase the income ?



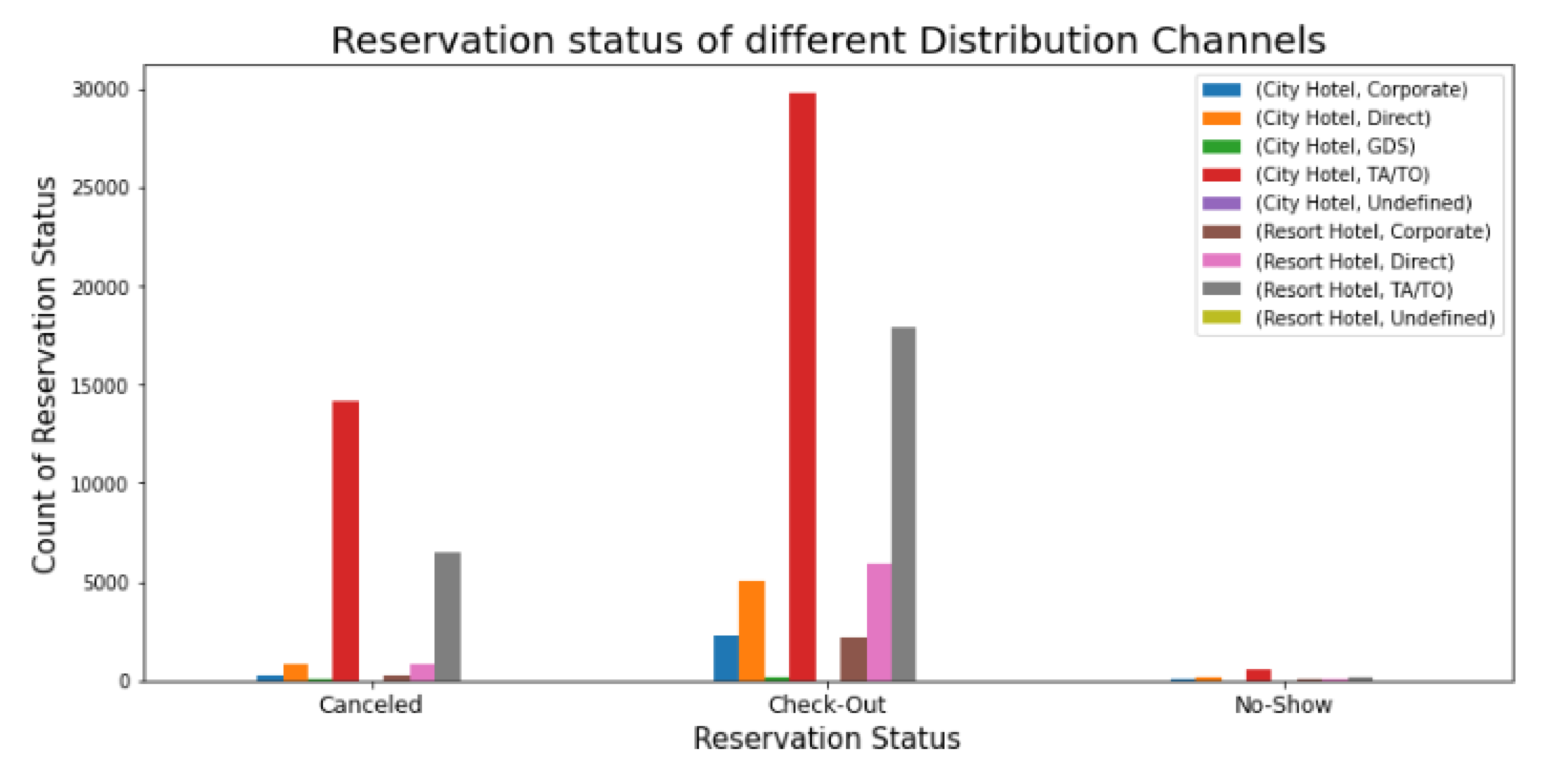
INFERENCE:

From the chart is clear that :

- 'Direct' and 'TA/TO' channel has almost equally contributed in ADR in both type of hotels i.e. 'City Hotel' and 'Resort Hotel' while 'TA/TO' dominates the chart in terms of total ADR.
- GDS has slightly contributed in ADR in 'City Hotel' type.
- Corporate booking channel has also contributed to the ADR.

❑ Exploratory Data Analysis (EDA):

What is the reservation status from different distribution Channels ?

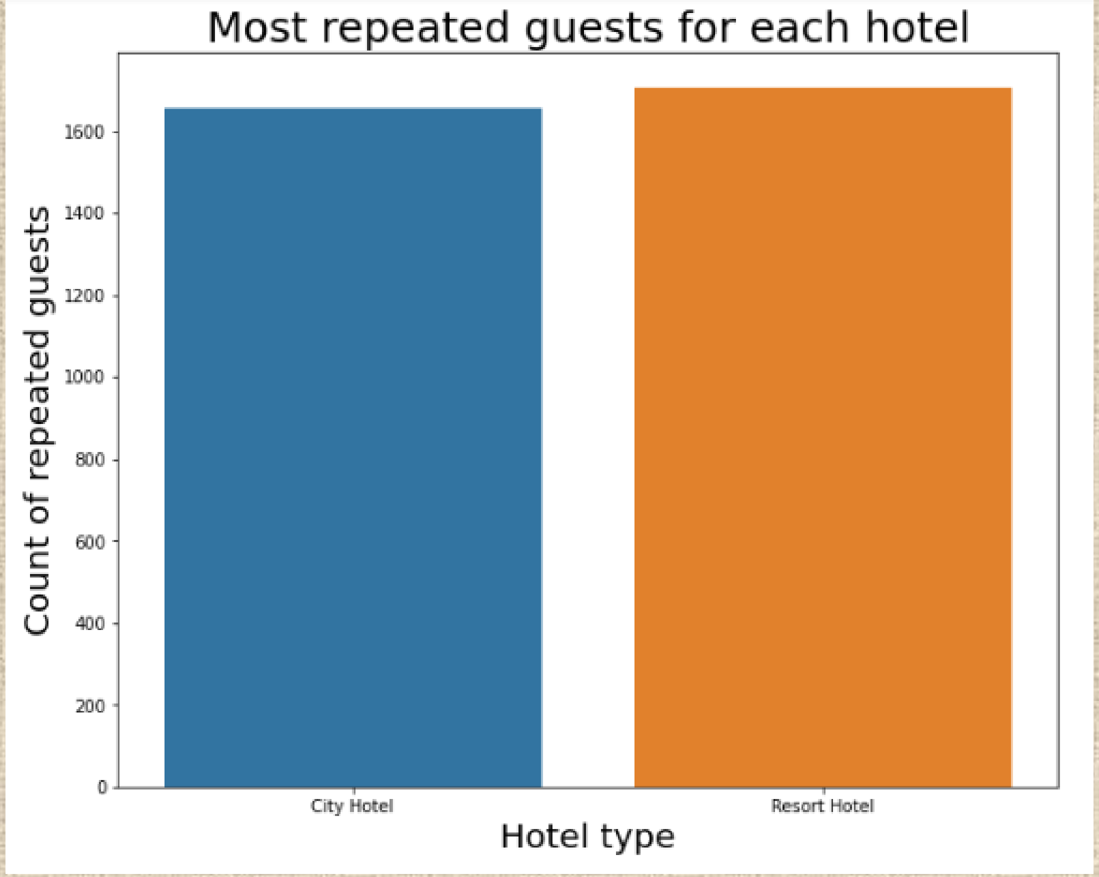


INFERENCE:

- We can infer from above graph that Bookings and Cancellations from both Hotels are more from Travel agency (TA/TO).
- Guest visiting both hotels, Directly and via Corporate are less likely to cancel their booking.
- We can notice a very small proportion of guest booking via Travel agency not showing up at Hotel.

❑ Exploratory Data Analysis (EDA):

Which Hotels has the most repeated guests?

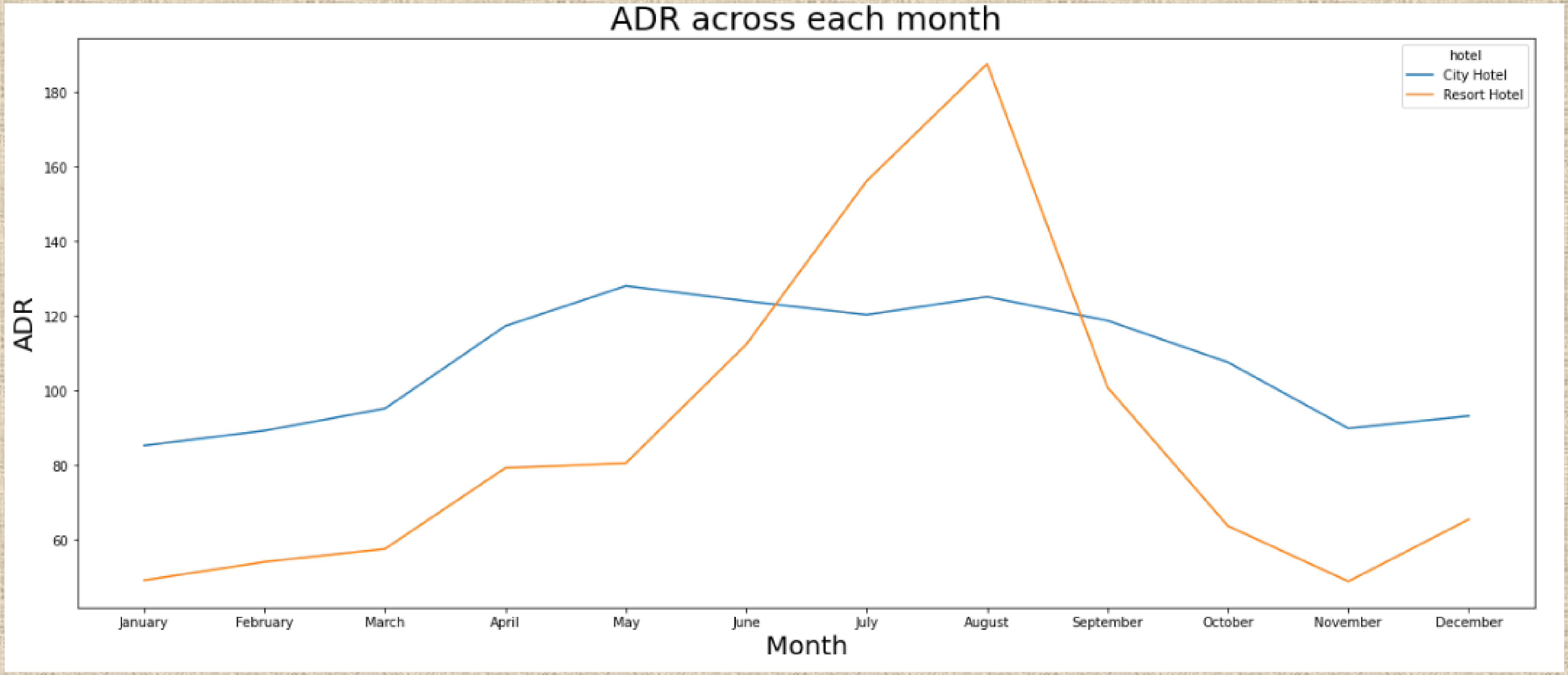


INFERENCE:

- Resort Hotel has slightly more repeated guests than the City Hotel. It is almost similar for both hotels.

❑ Exploratory Data Analysis (EDA):

What is the ADR across different months?

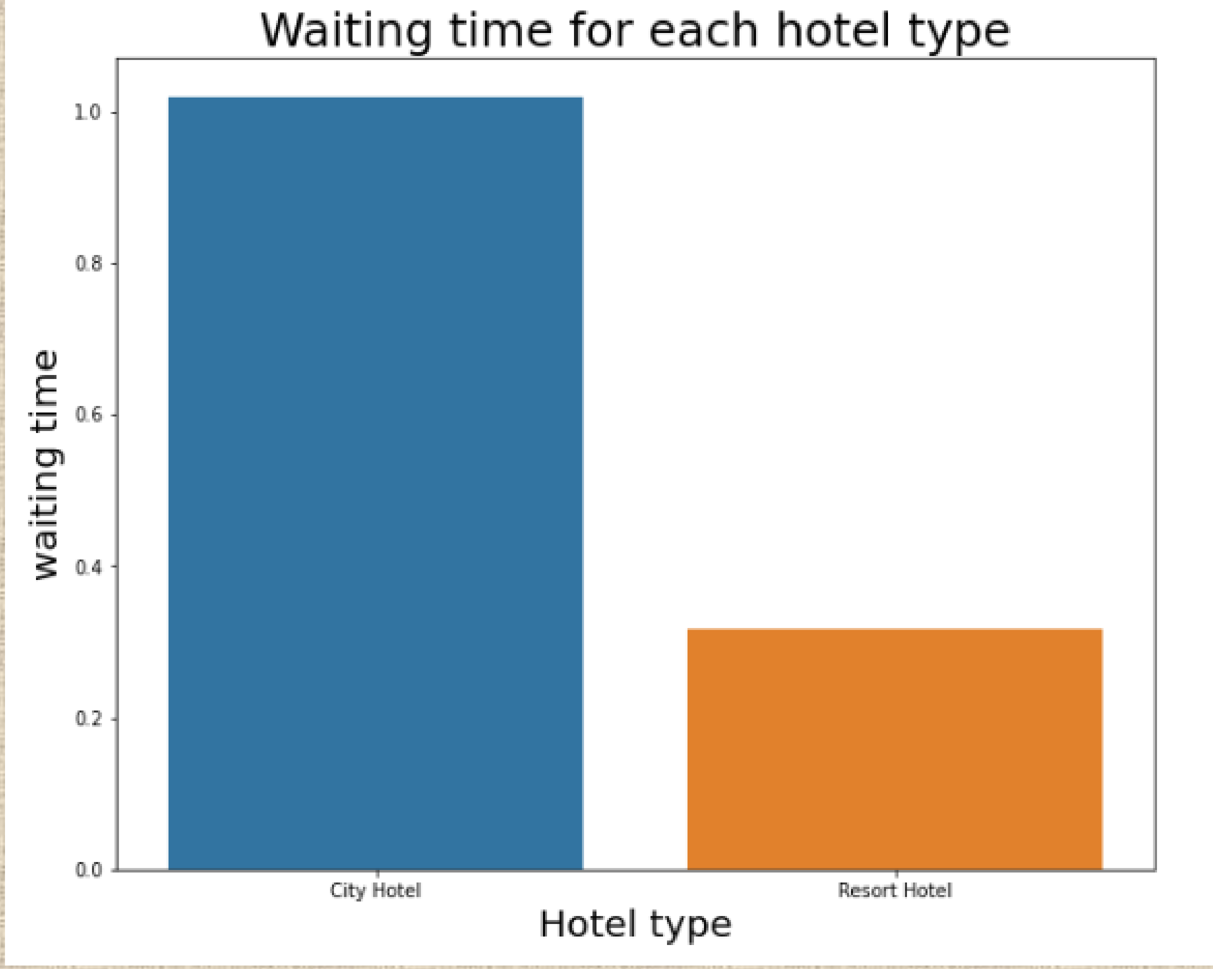


INFERENCE:

- For Resort Hotels the ADR is high in the months of June, July and August as compared to City Hotels. May be Customers wants to spend their Summer vacation in Resorts Hotels.
- The best time for guests to visit Resort or City hotels is January, February, March, April, October, November and December as the average daily rate in these months is very low.

❑ Exploratory Data Analysis (EDA):

Which hotel has the longer waiting time?

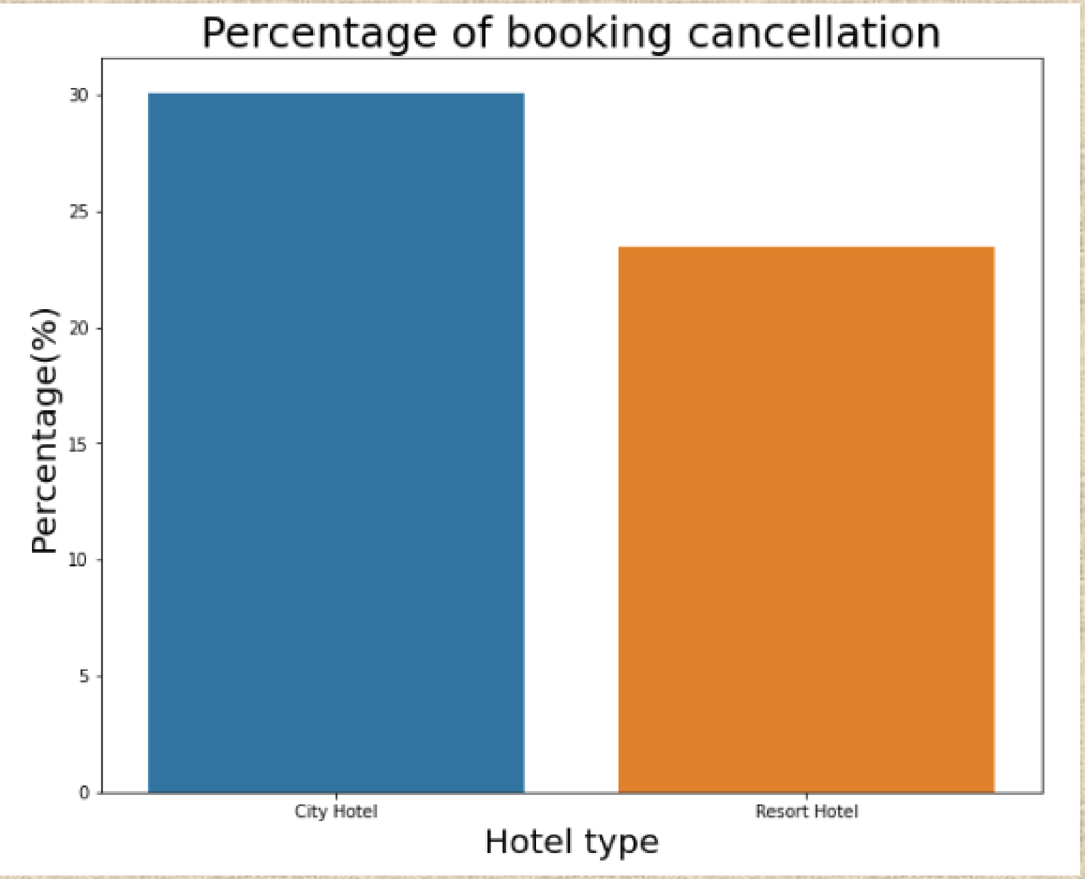


INFERENCE:

- The City Hotels have a longer waiting period than the Resort Hotels. Thus we can say that City Hotels are much busier than the Resort Hotels.

❑ Exploratory Data Analysis (EDA):

Which hotel has the highest percentage of booking cancellation?

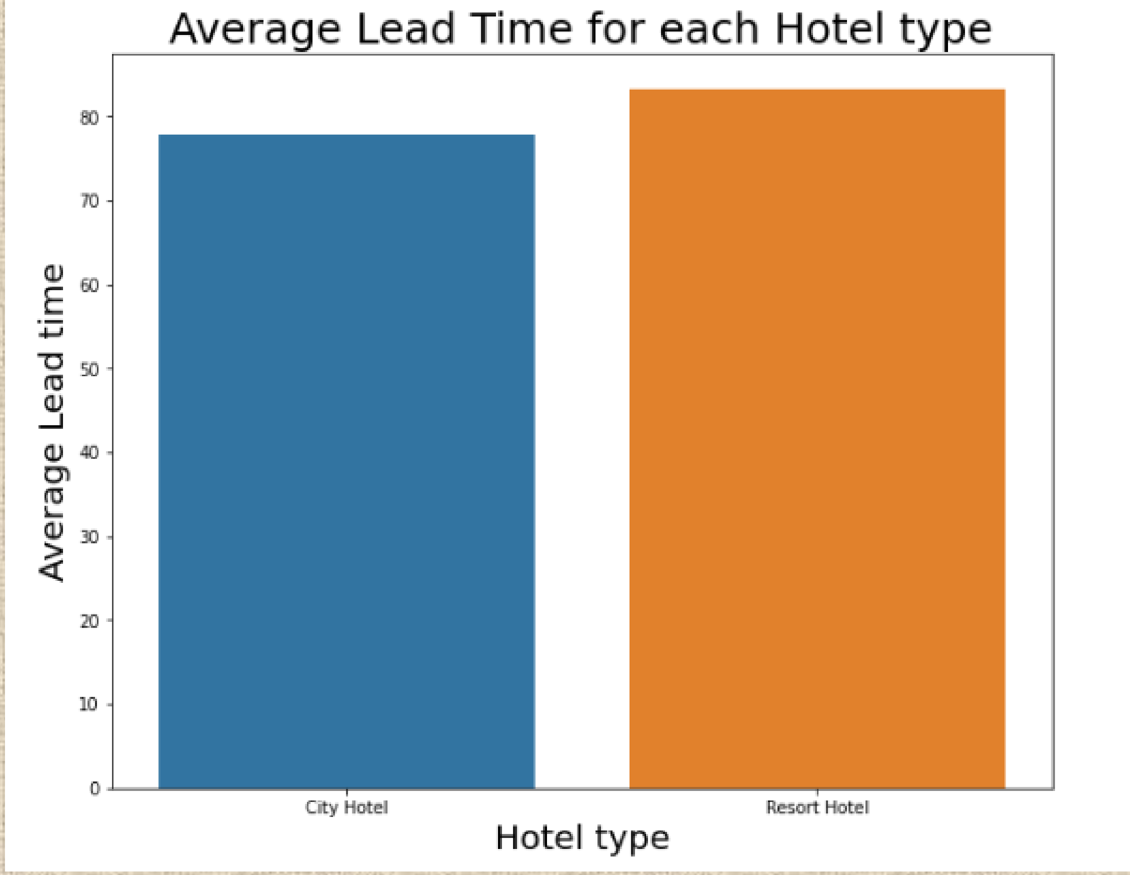


INFERENCE:

- City hotels have got a cancellation percentage of 30% whereas Resort hotels have got a cancellation percentage of around 24%

❑ Exploratory Data Analysis (EDA):

Which hotel type has the most lead time?

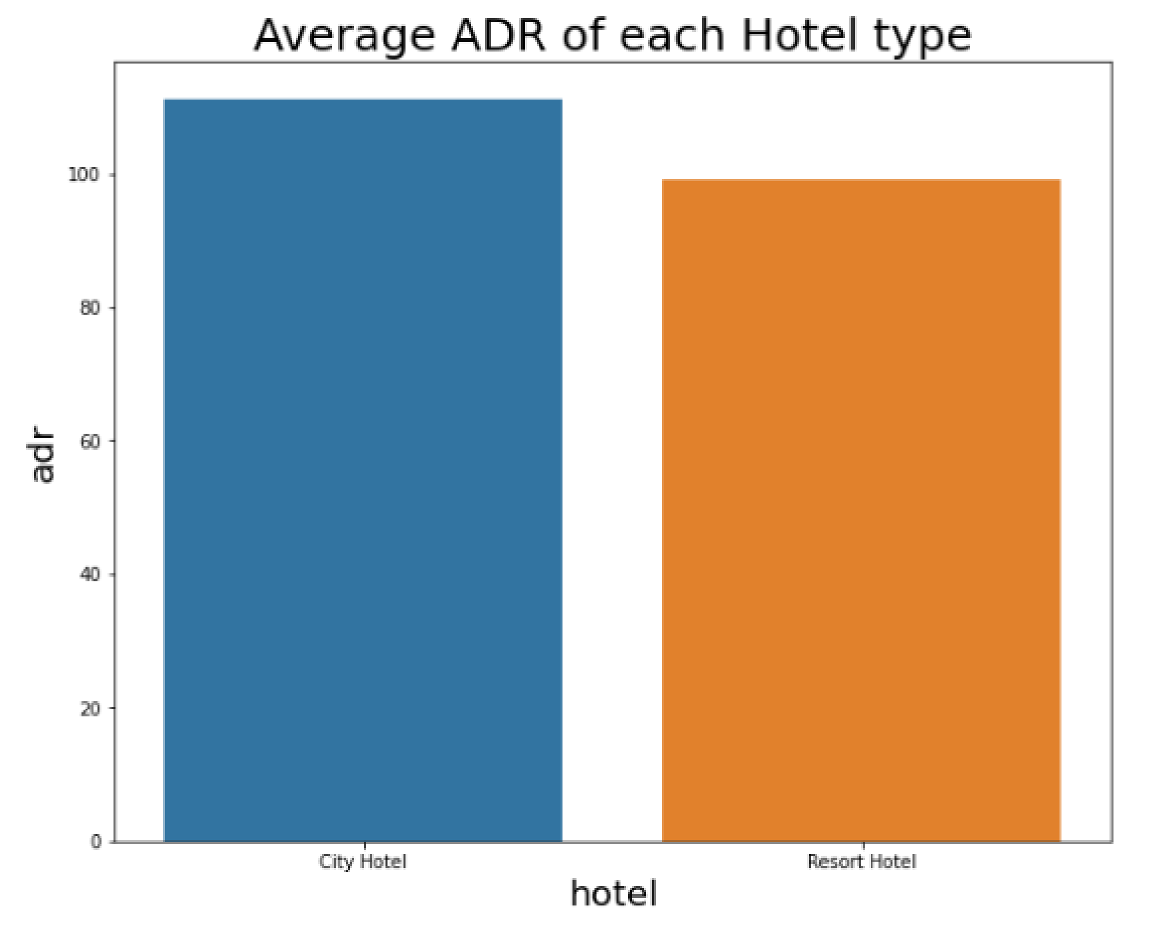


INFERENCE:

- Resort hotels have a slightly higher average lead time. Which means customers plan their trips way early than their day of check in.

❑ Exploratory Data Analysis (EDA):

Which Hotel type has the highest average ADR?

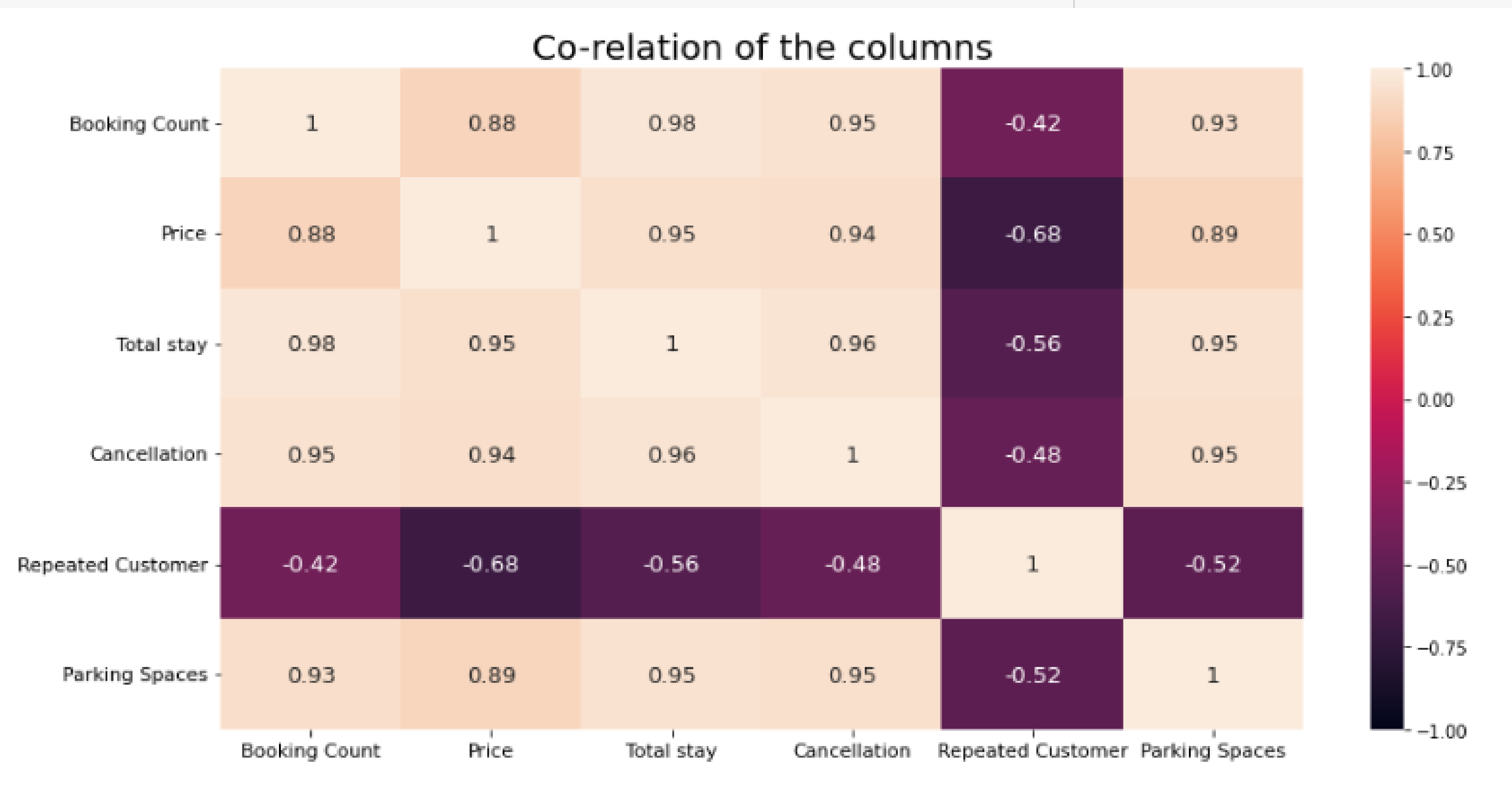


INFERENCE:

- City hotel has the highest average ADR. Which means city hotels are generating more revenues than the resort hotels. More the ADR, more is the revenue.

❑ Exploratory Data Analysis (EDA):

What is the Correlation between Booking, Pricing, Stay, Cancellation, Parking and Guest Revisiting?



Co-relation of the columns

	Booking Count	Price	Total stay	Cancellation	Repeated Customer	Parking Spaces
Booking Count	1	0.88	0.98	0.95	-0.42	0.93
Price	0.88	1	0.95	0.94	-0.68	0.89
Total stay	0.98	0.95	1	0.96	-0.56	0.95
Cancellation	0.95	0.94	0.96	1	-0.48	0.95
Repeated Customer	-0.42	-0.68	-0.56	-0.48	1	-0.52
Parking Spaces	0.93	0.89	0.95	0.95	-0.52	1

INFERENCE:

- There is a high positive correlation between Booking and Pricing, Total Stay, Cancellations and Parking spaces, where as negative correlation with Repeated guests.
- Increase in Pricing leads to repeated Customers not visiting again.
- There is firm correlation between Parking space and Cancellation inferring that people are more likely to cancel their booking if Parking space is not available.

Conclusion:

- Majority (61%) of the guests prefer City Hotel over Resort Hotel. Most of guest visiting these hotels are from European countries namely Portugal, Britain, France, Spain and Germany totaling to 75% of total booking count.
- 2016 observed the highest booking reservations. From Booking trend it can be inferred that Peak visiting season is from mid June to August because of summer breaks in Europe while November to February is off season because of freezing cold weather throughout Europe.
- Around 11.5% of total reservations throughout year are coming from August whereas January has the least reservation of mere 5%. Guests can consider visiting these hotels during month of June and September to enjoy decent weather with almost full availability of hotels accommodation.
- Inspecting different market segments, it was concluded that travel agency holds monopoly as both hotels are getting the most of booking from travel agency (around 79%). Hotel owners should consider promoting their hotels more in different market segments to penetrate market more.
- Interestingly, most of the Cancellations for both Hotels are from Travel agency (TA/TO) segment inferring that it is volatile market segment. Also, a very small proportion of guest booking via Travel agency do not showing up at Hotel. Guest visiting both Hotels directly and via Corporate are less likely to cancel their booking.

Conclusion:

- There is high positive correlation between Booking, Pricing, Total Stay, Cancellations and Parking spaces whereas negative correlation with Repeated guests. With increase in Booking > Pricing, Total stay and Parking spaces occupation increases but increase in Pricing leads to repeated Customers not visiting again.
- There is firm correlation between Parking space and Cancellation inferring that people are more likely to cancel their booking if Parking space is not available.
- Ideally guest prefer to stay 1-4 days in both hotels but 7 days stay at Resort hotel is also a popular choice among guests.
- Only 3.9 % people revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Booking cancellation rate is high for City hotels which is almost 30 %.
- Average lead time for resort hotel is high. This shows that customer make booking way in advance.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.

Thank
You!