CAPSTONE PROJECT 1

EDA ON HOTEL BOOKING ANALYSIS

By
Avishek Patra
Siddharth Ray
Kaushik Dey
Kushal Dixit
Kanha Ch. Pradhan

(Cohort Istanbul)

AlmaBetter

❖ Problem Statement:

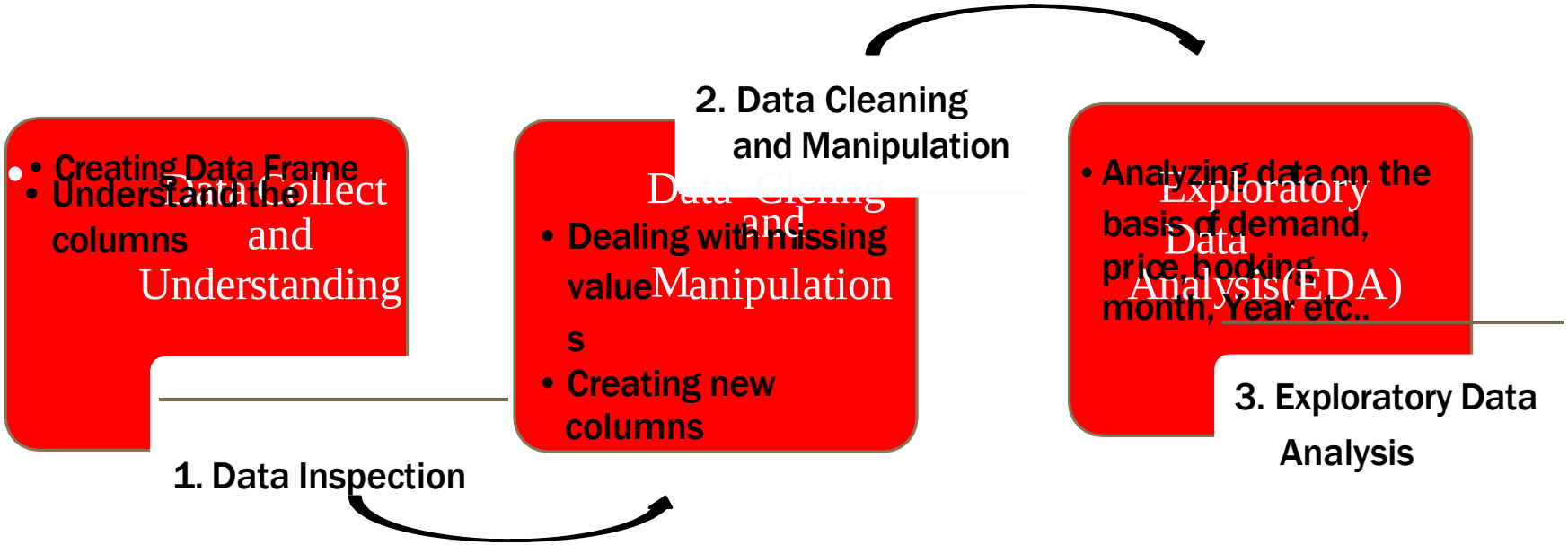
- In this project we will be analyzing Hotel Booking data. This data set contains booking information for a City hotel and a Resort hotel along with information on various booking criteria such as booking season, pricing data, length of stay, number of adults, children and babies, parking spaces, market 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 help the customers in deciding best period to visit places while availing low accommodation cost benefits.

Work Flow :

- We will divide our work flow into following 3 steps.



❖ Data Exploration and Inspection:

AI

- Libraries imported:
  - Data manipulation: numpy and pandas
  - Data visualization : matplotlib and seaborn

- Initial dataset size: 119390 rows and 32 columns.Data contains following features:
  - **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 thearrival 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.

Data Exploration and Inspection:

- **market_segment**: Market segment designation (TA/TO)
- **distribution_channel**: Booking distribution channel.(T/A/TO)
- **is_repeated_guest**: is a repeated guest (1) or not (0)
- **previous_cancellations**: Number of previous bookings that were cancelled prior to the current booking
- **previous_bookings_not_canceled**: Number of previous bookings not cancelled by the customer priorto 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
- **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 highfloor)
- **reservation_status**: Reservation last status.

Data Cleaning and Manipulation:

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

```
#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

```
#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 columns and rows

```
#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 Datetime 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

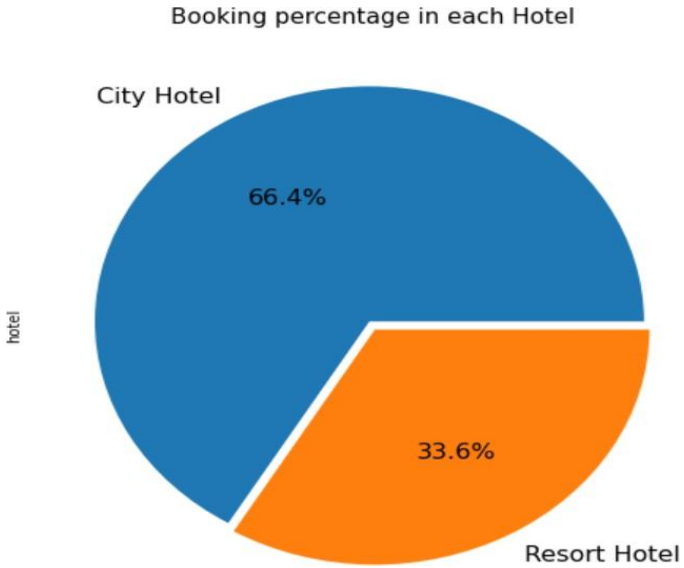
a. 'total_people' = total of adults, children and babies
b. 'total_stay' = total of weekend nights and weekdays nights

```
#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):

Booking percentage of different type of Hotels

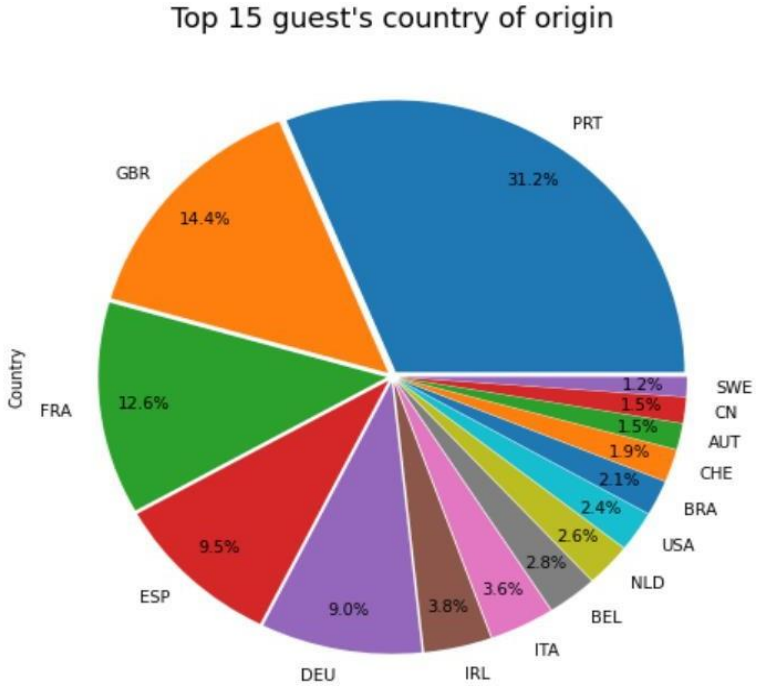


INFERENCE:

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

Exploratory Data Analysis (EDA):

Home country of majority of guests

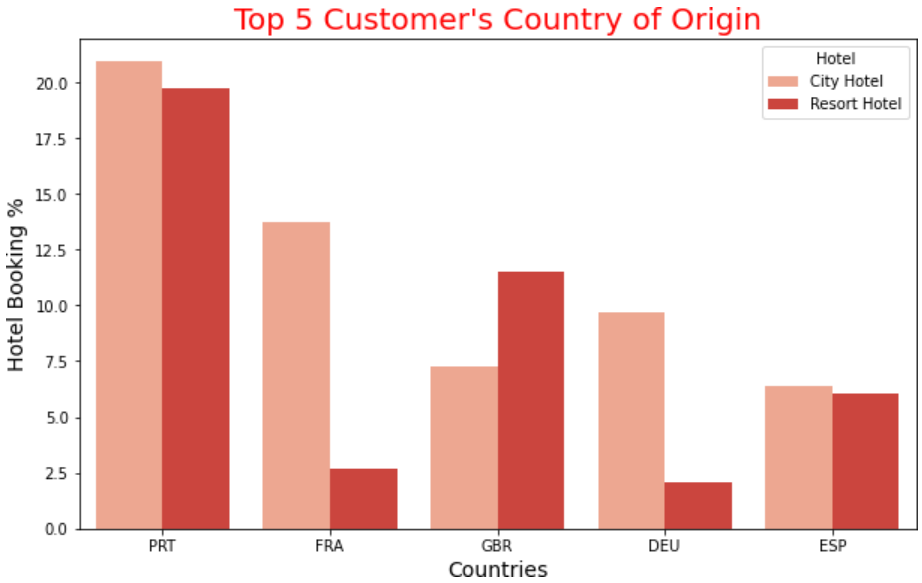


INFERENCE:

- From plotted pie chart, its evident that most of guest visiting these City hotels and Resort hotels are from Portugal and other European countries namely Britain, France, Spain and Germany.

Exploratory Data Analysis (EDA):

Hotel preference of guest from Top 5 Countries

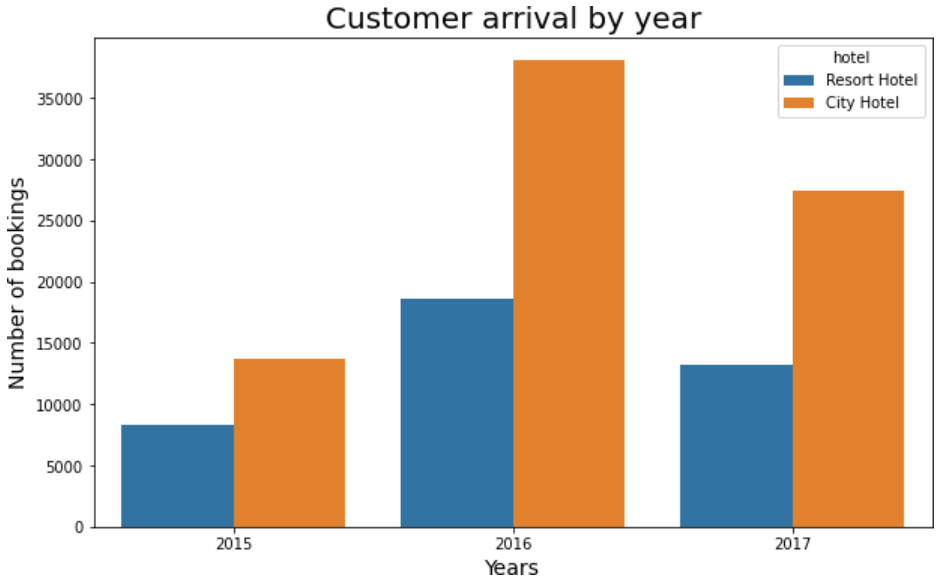


INFERENCE:

- Guest from southern European countries like Portugal and Spain prefer City Hotel and Resort Hotel equally
- Guest from northern European countries like France and Germany prefer City Hotel a lot more than Resort Hotel
- Guest from Britain prefers lavish Resort hotels

Exploratory Data Analysis (EDA):

Overview of guest's visit over different years



INFERENCE:

- As we can see that 2016 was the year where number of hotel booking was highest followed by total booking in 2017 and 2015

Exploratory Data Analysis (EDA):

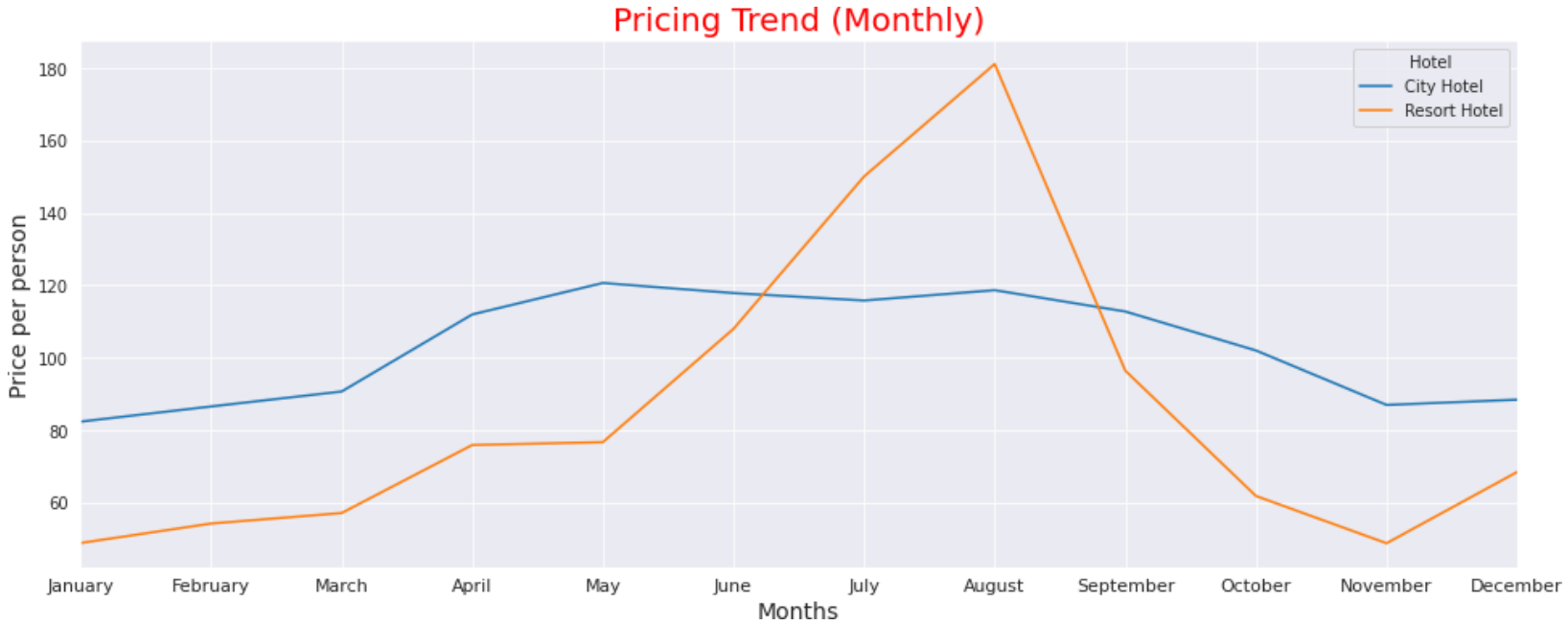
Booking trend round the year

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
- Guests can consider visiting these hotels during month of June and September to enjoy decent weather with almost full availability of hotels accommodation.

Exploratory Data Analysis (EDA):

Price trend round the year

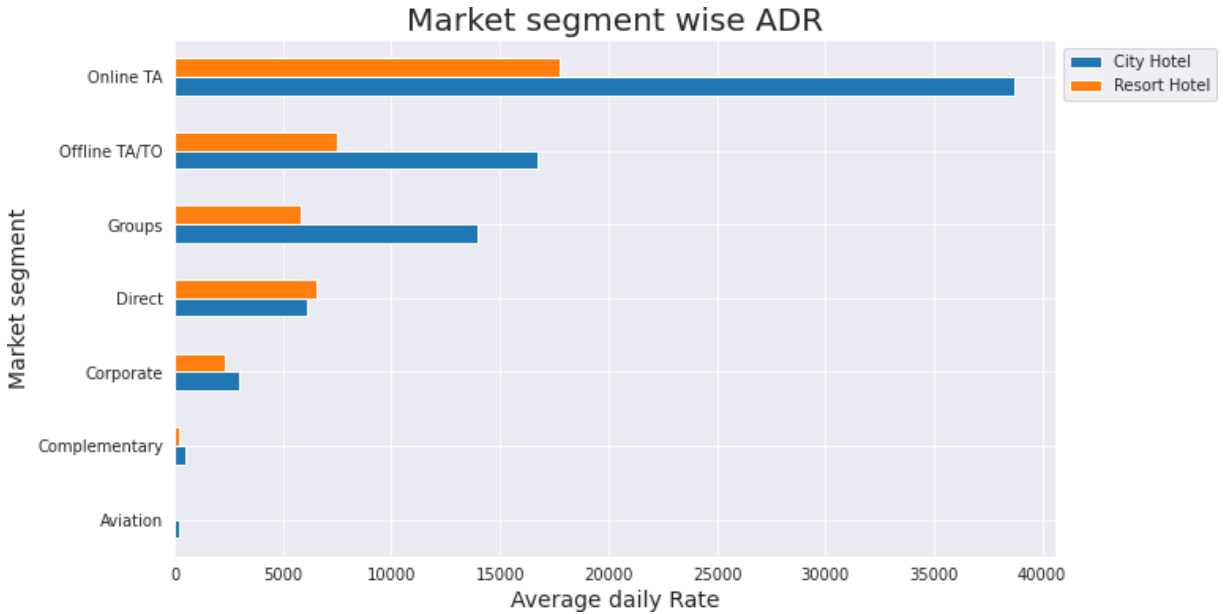


INFERENCE:

- Pricing trend is highly correlated with booking trend when it comes to Resort hotel
- Pricing trend here indicates that during Peak season, price for Resort hotels is triple compared to off-season
- Pricing trend for City hotels suggests almost same pricing throughout the year with bit of fluctuation during May to August

Exploratory Data Analysis (EDA):

Average booking rate of different Market Segments

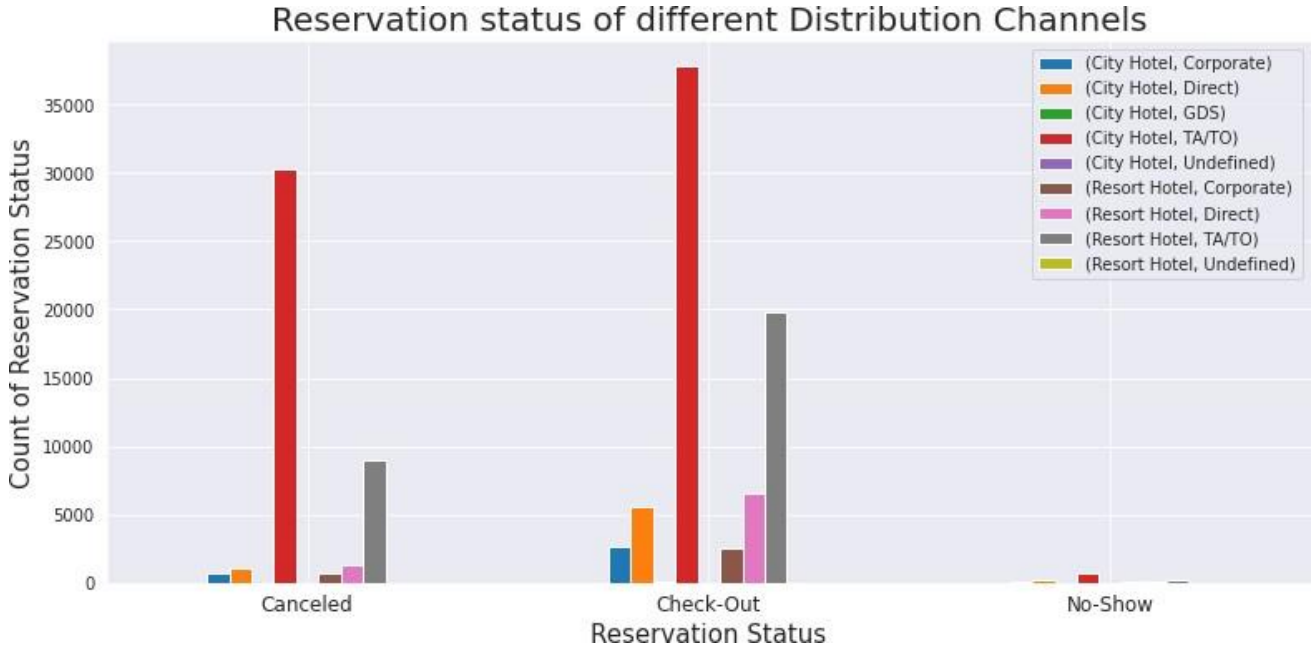


INFERENCE:

- As we can see, Resort hotel and City hotel are getting the most of booking from Online travel agency and may be in future it will be monopoly by them. Hence hotel owners should promote more in different market segments

Exploratory Data Analysis (EDA):

Reservation status from different Distribution Channels

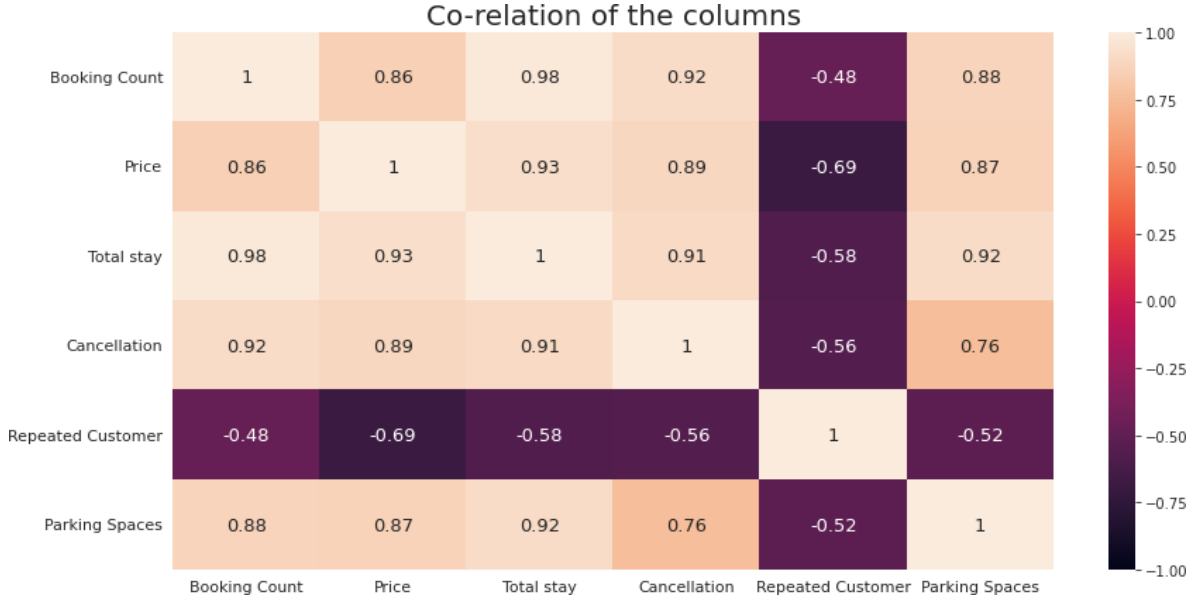


INFERENCE:

- 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):

Correlation between different booking criteria

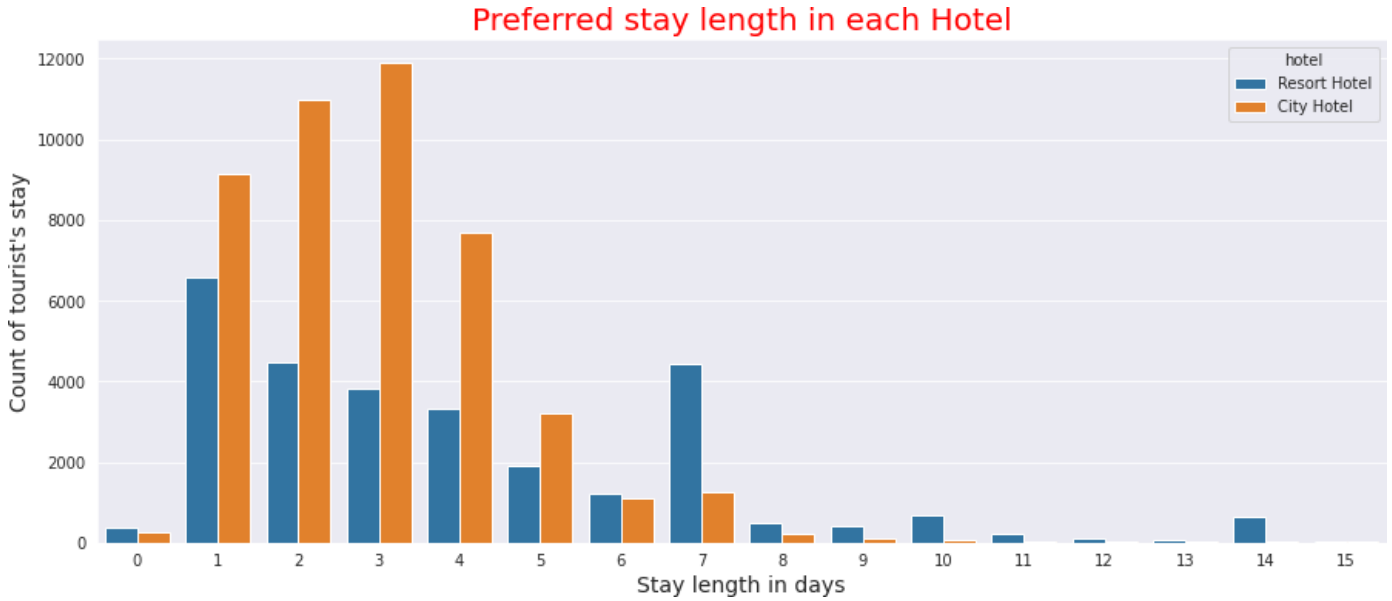


INFERENCE:

- There is high positive correlation between Booking, Pricing, Total Stay, Cancellations and Parking spaces where as negative correlation with Repeated guests
- With increase in Booking → Pricing, Total stay and Parking spaces occupation increases but increase in Pricing also 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.

Exploratory Data Analysis (EDA):

Guest's stay length



INFERENCE:

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

Conclusion:

- Majority (**66%**) 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.

- Guest from southern European countries like Portugal and Spain prefer both hotels equally. Guest from northern European countries like France and Germany prefer City hotel nearly **70%** more than Resort hotel. Guests from Britain prefers lavish Resort hotels nearly **25%** more than Resort hotel. This indicated that people from different region of Europe prefer different type of accommodations and comforts.

- 2016 observed the highest booking reservations. From Booking trend its 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.

- Pricing trend is highly correlated with booking trend indicating that price for Resort hotels during peak season hikes to nearly **300%** compared to off-season. Meanwhile, Pricing trend for City hotels suggests almost same pricing throughout the year with low fluctuation during busy period from May to August

Conclusion:

- Inspecting different market segments, it was concluded that Online travel agency holds monopoly as both hotels are getting the most of booking from Online 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 also 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

- 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.

Signing off...

THANK YOU