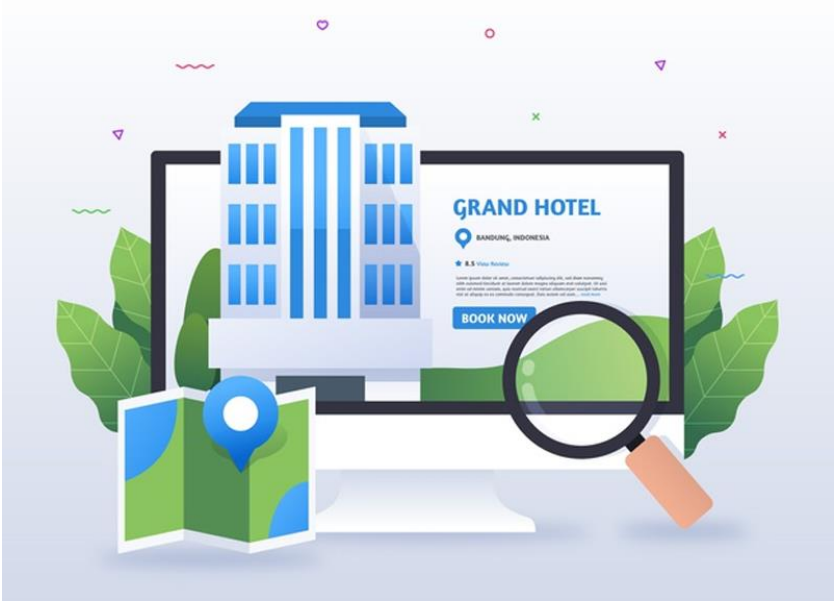
Capstone Project-1

Hotel Booking EDA Analysis

Aashruti Agrawal
(Individual)

Flow of Presentation

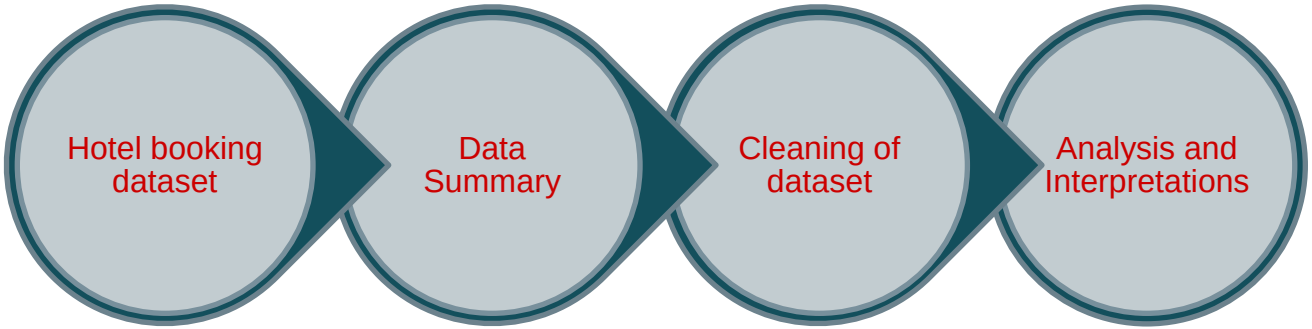
- Agenda
- Data Summary
- Cleaning of Dataset
- Data Visualization
- Inferences
- Conclusion



Agenda

Agenda is to discuss the given problem statement i.e. **Analysis of various covariates governing the bookings of hotel** in the given hotel booking dataset.

APPROACH:



Data Summary

hotel

is_canceled

lead_time

arrival_date_year

arrival_date_month

arrival_date_week_number

arrival_date_day_of_month

stays_in_weekend_nights

stays_in_week_nights

adults

children

babies

meal

country

market_segment

distribution_channel

is_repeated_guest

previous_cancellations

previous_bookings_not_canceled

reserved_room_type

assigned_room_type

booking_changes

deposit_type

agent

company

days_in_waiting_list

customer_type

adr

required_car_parking_spaces

total_of_special_requests

reservation_status

reservation_status_date

Data Summary

hotel: category of hotel; resort hotel or city hotel

is_canceled: categorical column indicating 0 as booking not cancelled

lead_time: time between reservation and actual arrival of guest

stays_in_weekend_nights: number of weekend nights stayed by guest

meal: meal preference of the guest

market_segment: indicates the purpose of reservation Ex. Corporate, TA for Travel Agency

distribution_channel: platform of booking Ex. Direct, corporate, travel agency

is_repeated_guest: indicates the guest are previous customers or not. 0 indicates the customer is a new customer.

Cleaning of dataset

Missing Values

In [8]:
```
# getting summary of missing values present in the dataset.

for column in raw_df:
    if raw_df[column].isnull().any():
       #print('{0} has {1} null values {2} '.format(column, df[column].isnull().sum(), df[column].isnull().sum() * 100 /df.shape[
        print('{0} has {1} missing values, which are {2} % of total column '.format(column, raw_df[column].isnull().sum(), round(
```

```
children has 4 missing values, which are 0.0061 % of total column
country has 482 missing values, which are 0.73548 % of total column
agent has 9991 missing values, which are 15.24529 % of total column
company has 61836 missing values, which are 94.35569 % of total column
```

Checking for duplicates

In [9]:
```
# creating copy of dataset

working_df = raw_df.copy()

working_df[working_df.duplicated()].shape   # Show no. of rows of duplicate rows duplicate rows
```

Out[9]:
```
(17617, 32)
```

In [10]:
```
# Droping duplicate values
working_df.drop_duplicates(inplace = True)
```

In [11]:
```
working_df.shape
```

Out[11]:
```
(47918, 32)
```

Cleaning of dataset

In [63]:
```
#lets group some of the columns which can be useful in analysis as grouped element

working_df['kids'] = working_df['children'] + working_df['babies']

working_df['Total_Guests'] = working_df['adults'] + working_df['kids']
```

In [64]:
```
df= working_df.copy()

#dropping the column which are not required for the analysis

df=working_df.drop(['company', 'babies'],axis=1)
```

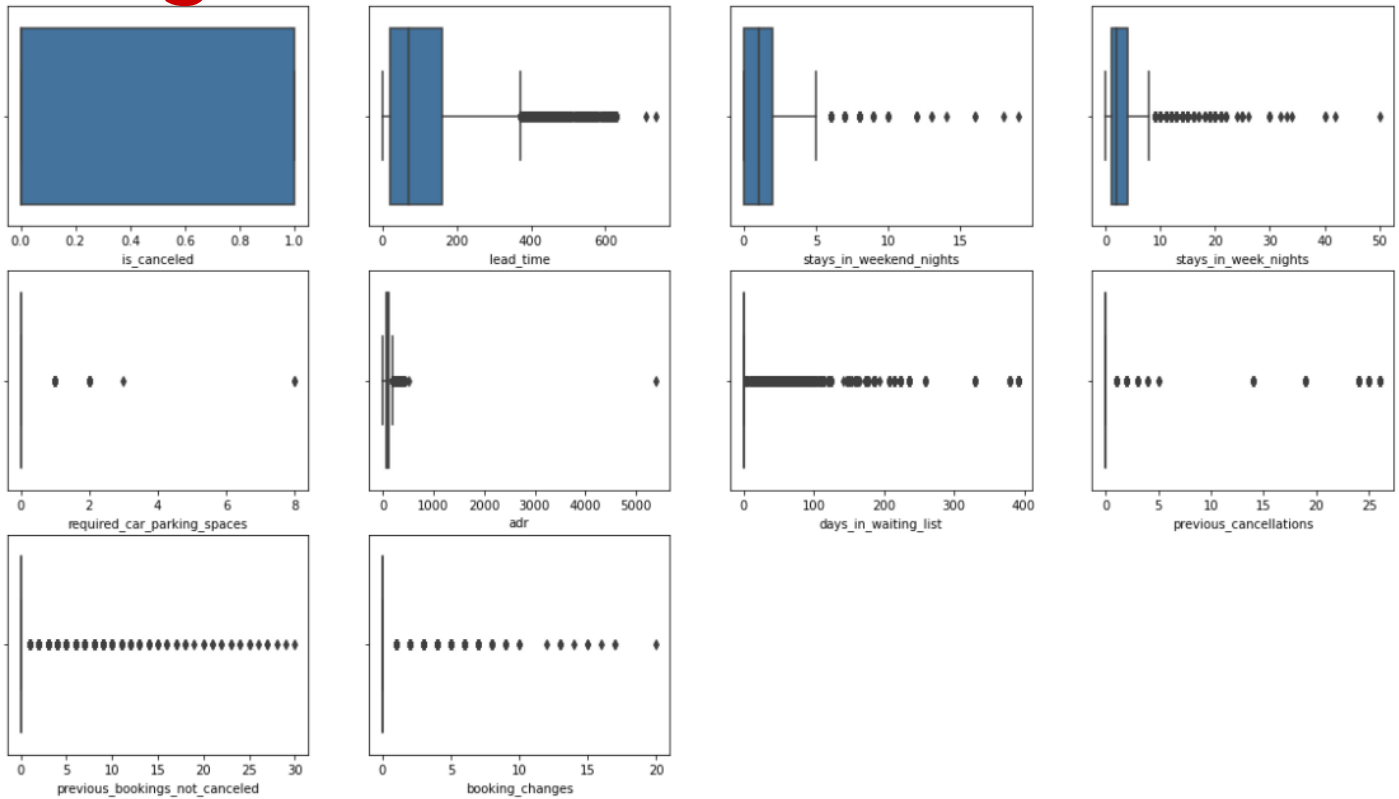
Now, our dataset is clean and ready-to-use for analysis and visualization. But, before moving towards analysis part , let us first try to estimate the reliablity of the dataset using Statistical techniques. As of now, we are limiting this till the outlier detection. Based on outliers, we will try to assume reliabilty of the data. This can also be done by observing the column features, if the data is normally distributed by plotting a histogram, the data will be more reliable.

In [65]:
```
# statistical summary of the numerical columns of the dataset
df.describe()
```

Out[65]:

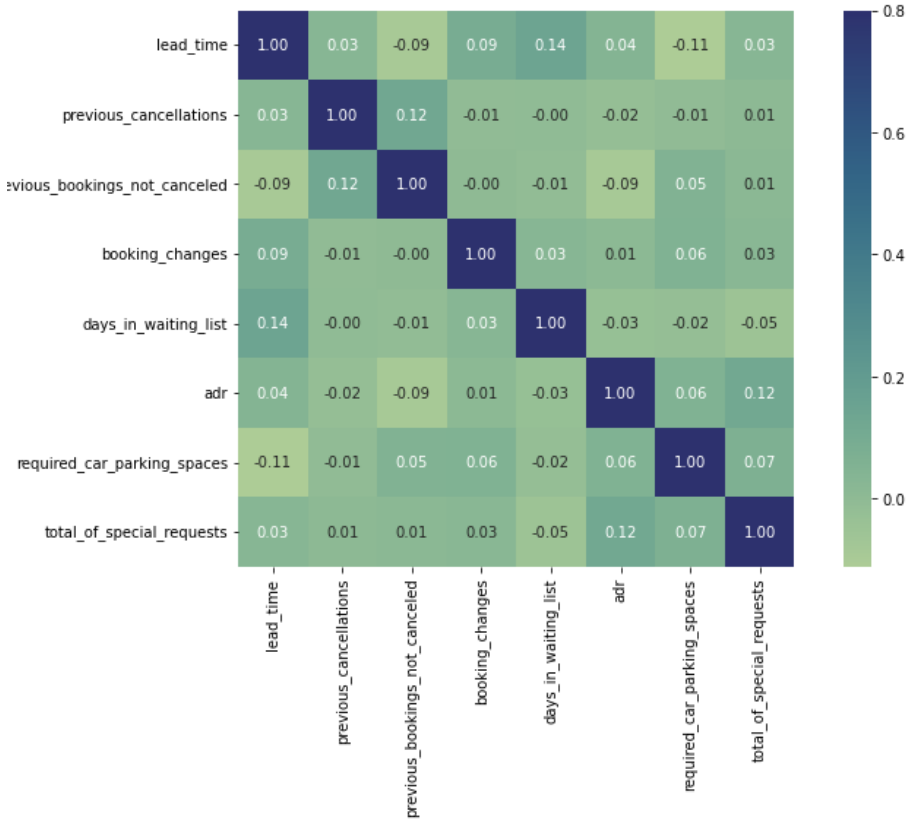
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	
count	65535.000000	65535.000000	65535.000000	65535.000000	65535.000000	65535.000000	65535.000000	655
mean	0.470802	104.106188	2016.030137	27.700633	15.727932	1.032685	2.807355	
std	0.499151	107.736740	0.690716	14.194923	8.837104	1.085659	2.197926	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	
25%	0.000000	19.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	

Cleaning of dataset



As we can see, this dataset has many outliers and thus we can infer that this dataset is not very reliable.

Data Visualization



Data Visualization

Following insights are pulled out from this analysis:

1. Which hotel is more preferred? Were the guests repeated?
2. How many bookings were cancelled? Which type of meal offered by hotels?
3. Which market and distribution channel is dominant?
4. Which room type was reserved by guests and were they assigned the same room types?
5. Which country has made the highest booking?
6. What were the most active business month?
7. What were the most active business week?
8. What were the most active business day?
9. Which country has cancelled more booking?
10. Which month has most booking cancellation?

Data Visualization

1. Most revenue is produced by which hotel?
2. Which room type is produces more revenue?
3. What is the length of stay of guests?
4. Do guests say more at week days or weekends?
5. Price trend of hotels?
6. Relation between ADR and length of stay
7. Effect of lead time on booking cancellation

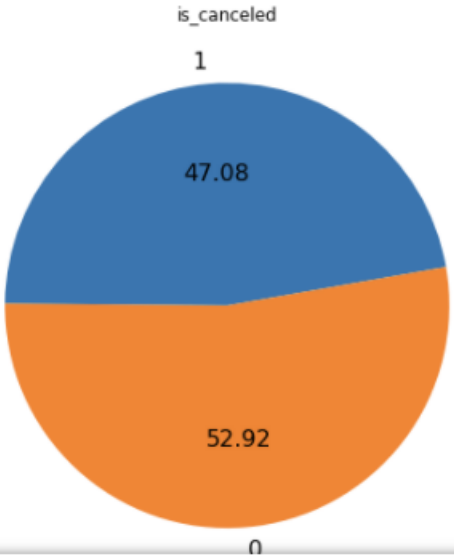
Data Visualization

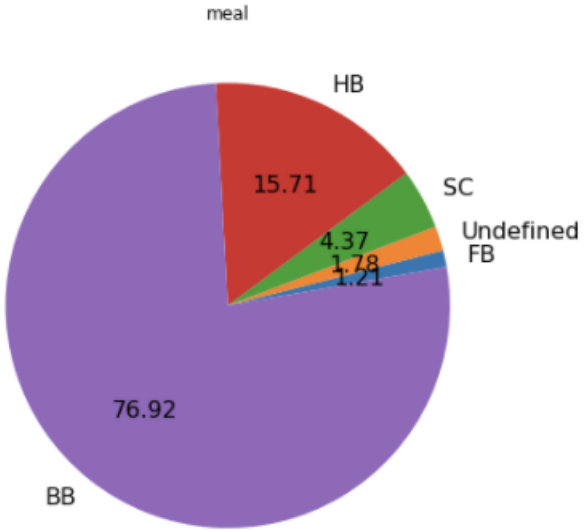
Which hotel is more preferred? Were the guests repeated?

Resort hotels were the preferred choice by customer with 61.13% bookings. It could be attribute to good customer facility.
Most of the **guests were new**, only 2.71% guests were repeated.

Data Visualization

How many bookings were cancelled? Which type of meal offered by hotels?



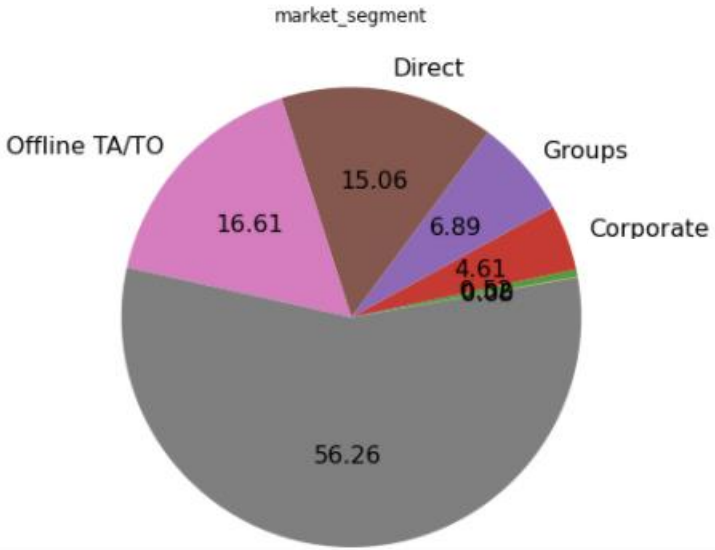


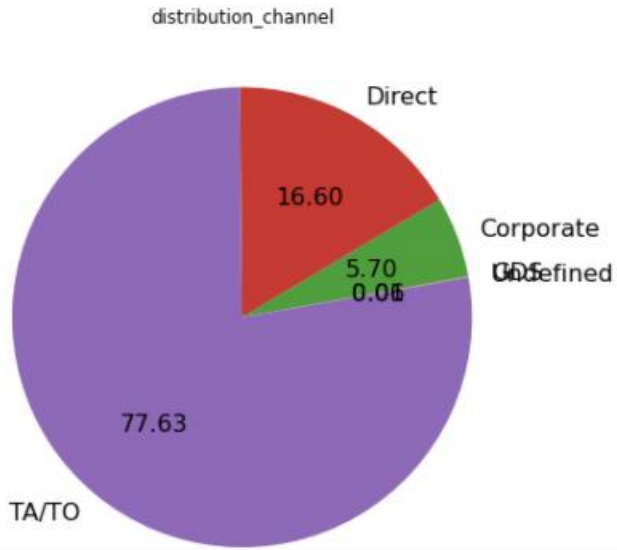
47.08% customers have **cancelled** their booking. Now, further exploration needs to be done to understand the possible reason of cancellation.

76.92% hotel provide **breakfast**, which may help hotel in getting good rating from customer. This could be a potential co-variate for booking status.

Data Visualization

Which market and distribution channel is dominant?

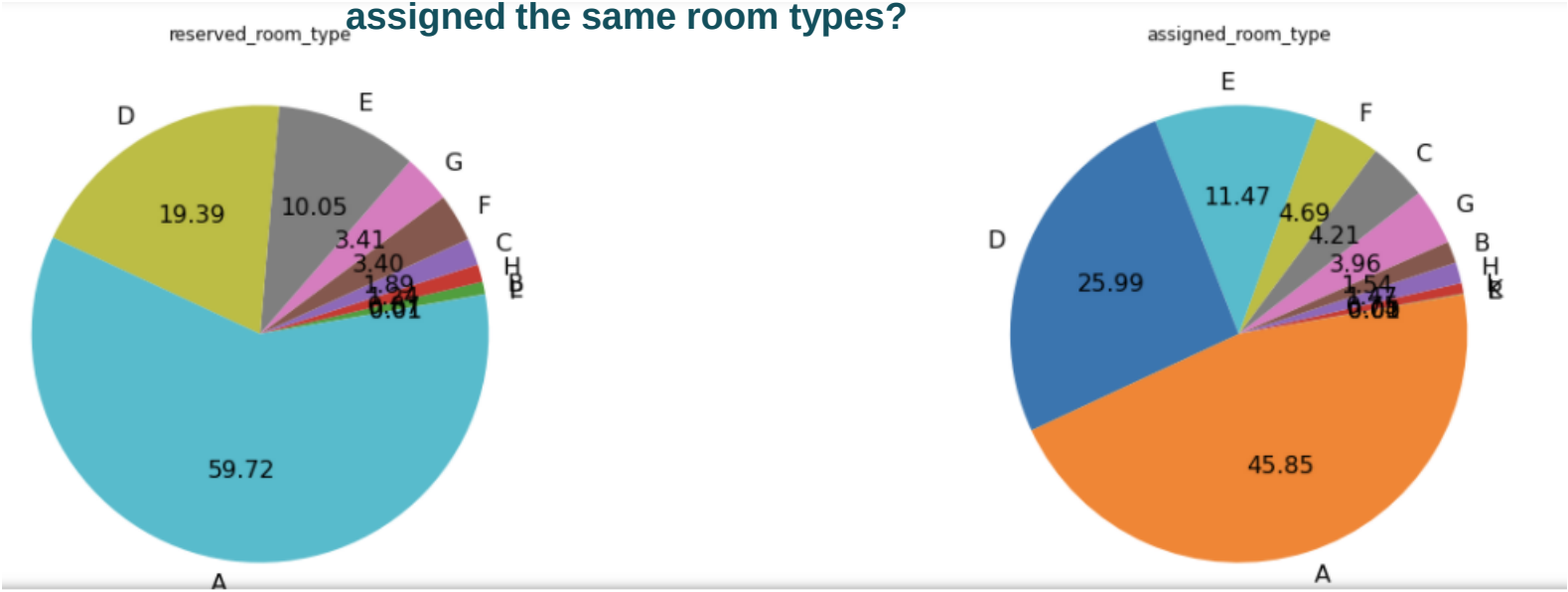




Market was widely captured by distribution channel like**: TA or TA/TO** especially by "Online TA" with 44.81% market capture.

Data Visualization

Which room type was reserved by guests and were they assigned the same room types?

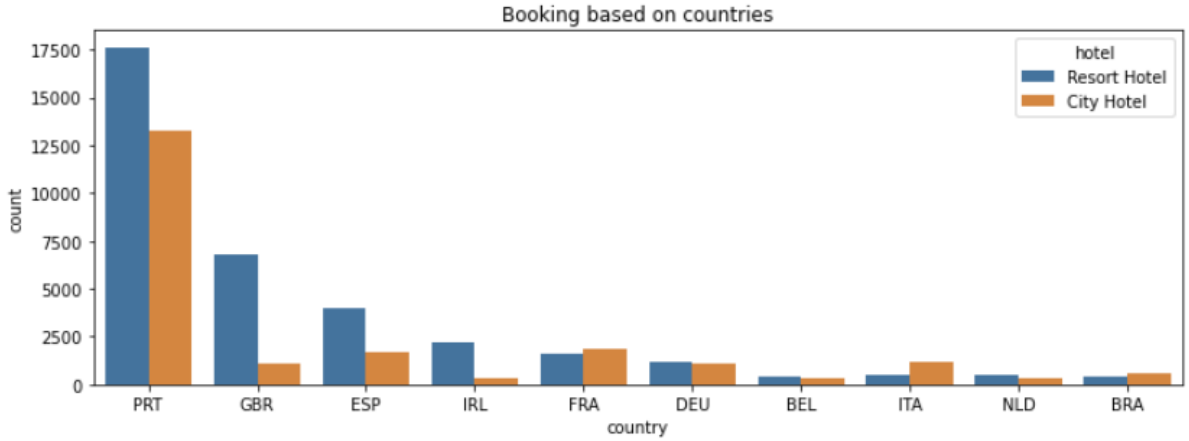


We can observe a similar trend in the assigned room type and the reserved room type. This is indicative that there has been **less modification in the bookings**. Around 56% customers were assigned **A room type** and 68% has reserved it. This also indicates the fact that room type A is on higher demand.

Data Visualization

Which country has made the highest booking?

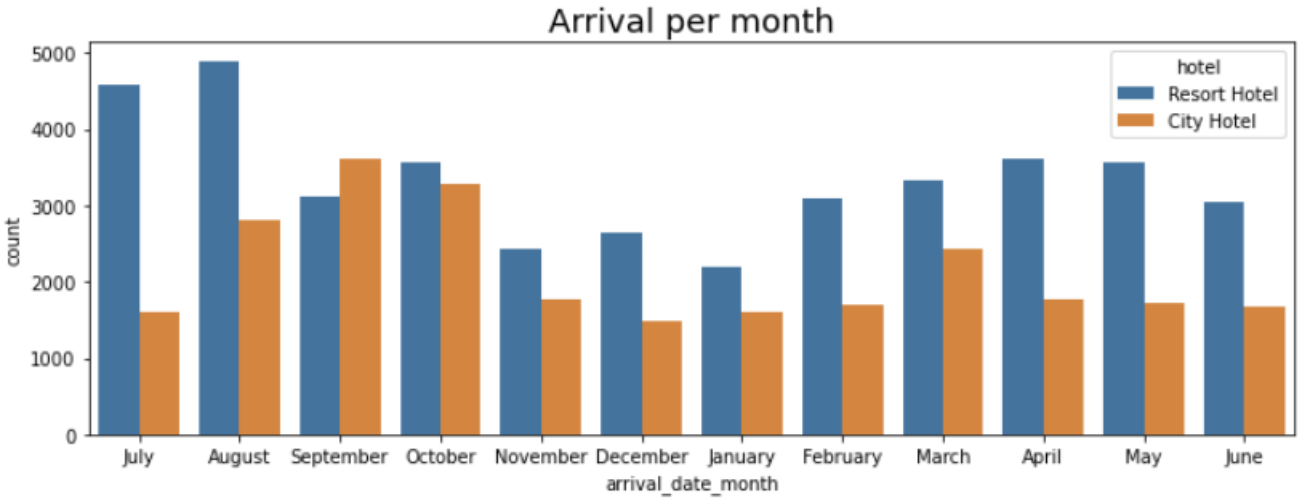
	country	number_of_bookings	percentage
0	PRT	12270	40.973753
1	GBR	6131	20.473519
2	ESP	3759	12.552595
3	FRA	2123	7.089428
4	IRL	1783	5.954051
5	DEU	1493	4.985641
6	ITA	655	2.187270
7	CN	652	2.177252
8	NLD	556	1.856675
9	BEL	524	1.749816



We can infer, that most number of confirmed booking was done by **Portugal** with around 41%. Out of them most of the booking were of resort hotel.

Data Visualization

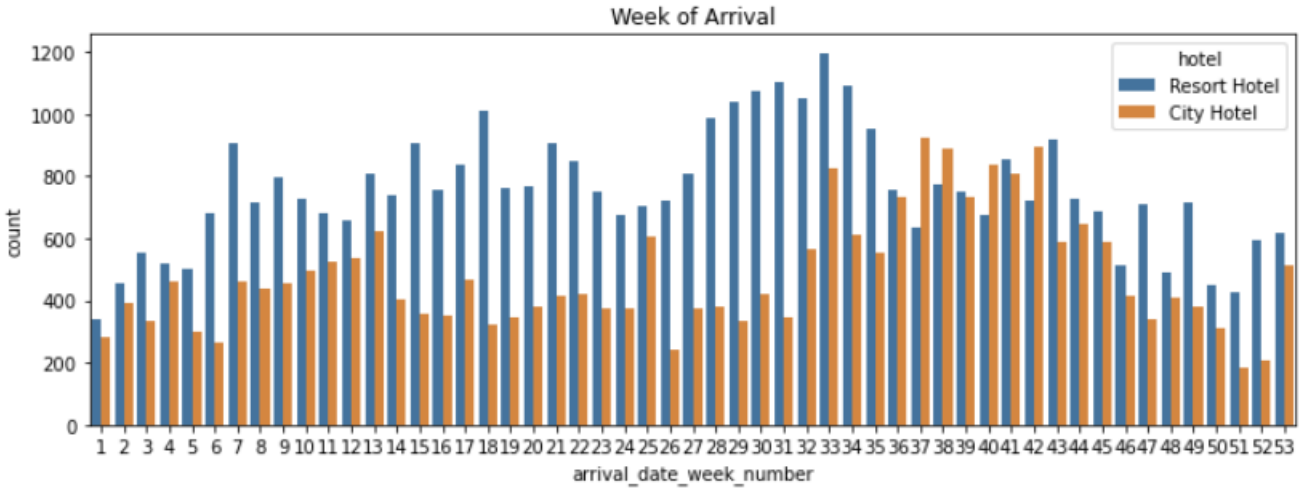
What were the most active business month?



Most active business month in terms of booking is **August** for resorts and **September** for city hotels. Resorts were more preferred option in Summer season

Data Visualization

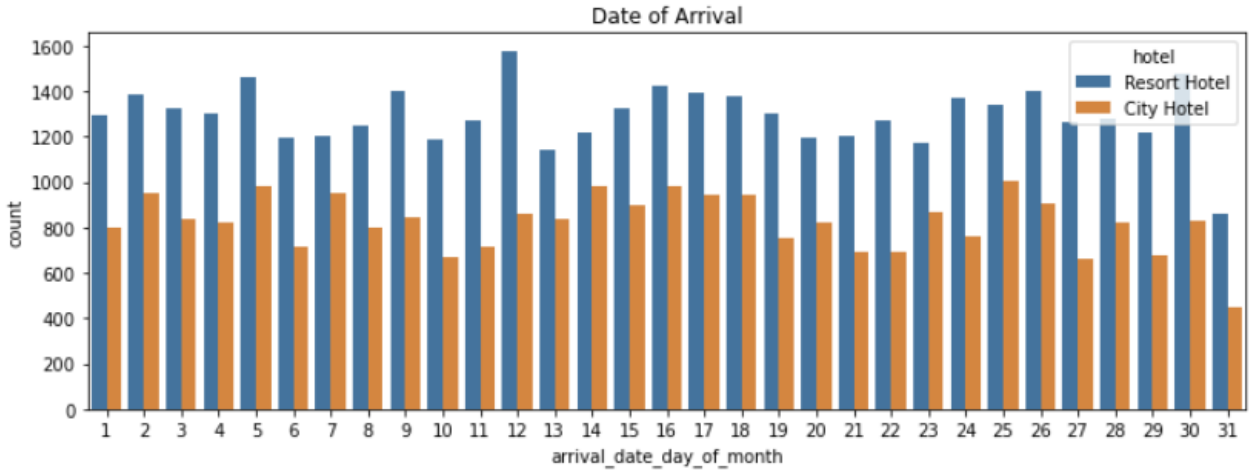
What were the most active business week?



As aligning with the month, **week 27 -34** collected more resort bookings while during **week 35-42** more city hotels were booked.

Data Visualization

What were the most active business date?



There is a similar pattern in booking of both the hotel. However, most of the bookings were done at the **middle of the month**. i.e. between 13 -20 date of the month.

Data Visualization

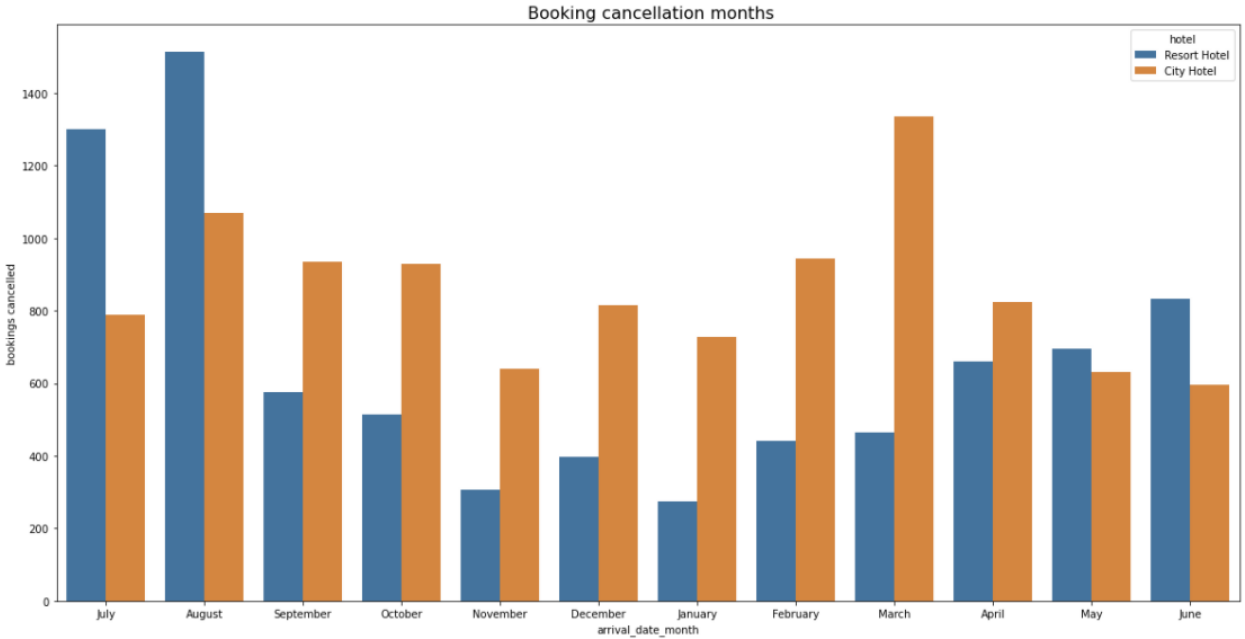
Which country has cancelled more booking?

Although **Portugal** has done the highest booking but its cancellation rate is also very high as compared to other countries.

Data Visualization

Which month has most booking cancellation?

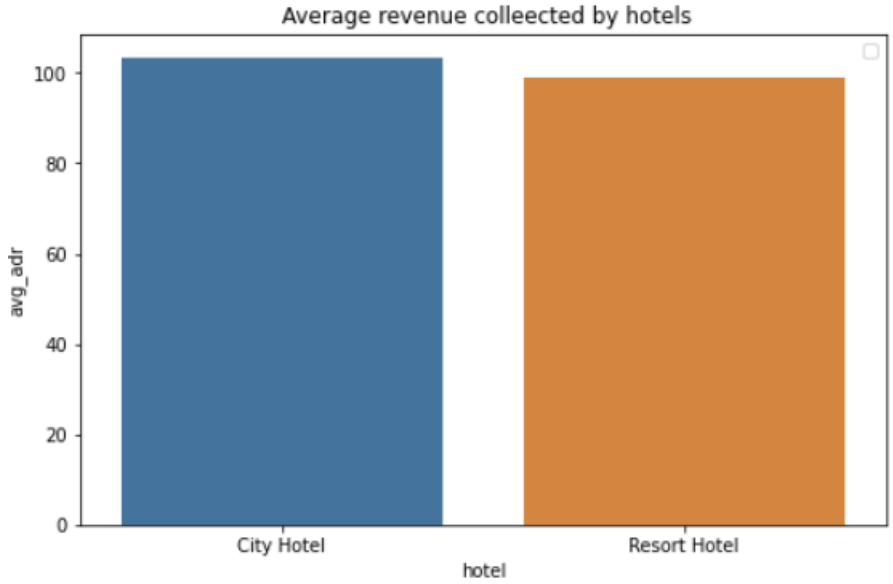
```
October      3349
August       3297
September    3129
March        2962
July         2704
June         2620
April        2589
May          2383
February     2329
November     1937
December     1862
January      1626
Name: arrival_date_month, dtype: int64
```



Most of the bookings were cancelled in the month of **March** **for City hotels**, while Resort hotels were cancelled in the **month of August**. Moreover, similar cancellation of the booking were seen in August month for both the hotel type.

Data Visualization

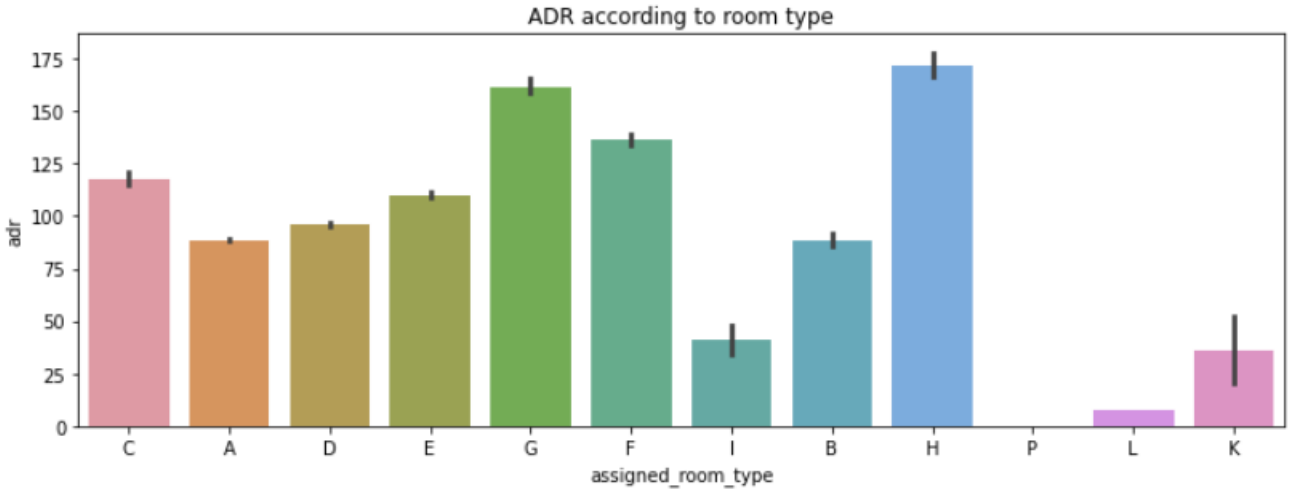
Most revenue is produced by which hotel?



Although resorts are the preferred hotel by guest but **city** **hotel** produced more revenue comparatively.

Data Visualization

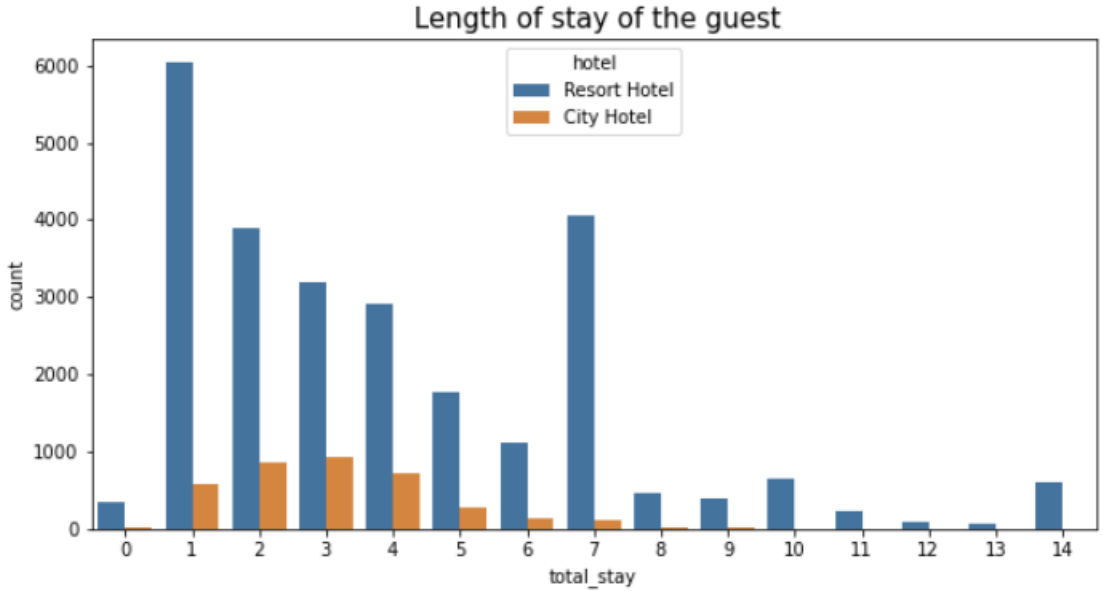
Which room type is produces more revenue?



Although room type A was on demand and most booked room. Highest revenue was produced by room type **H followed by G**.

Data Visualization

What is the length of stay of guests?

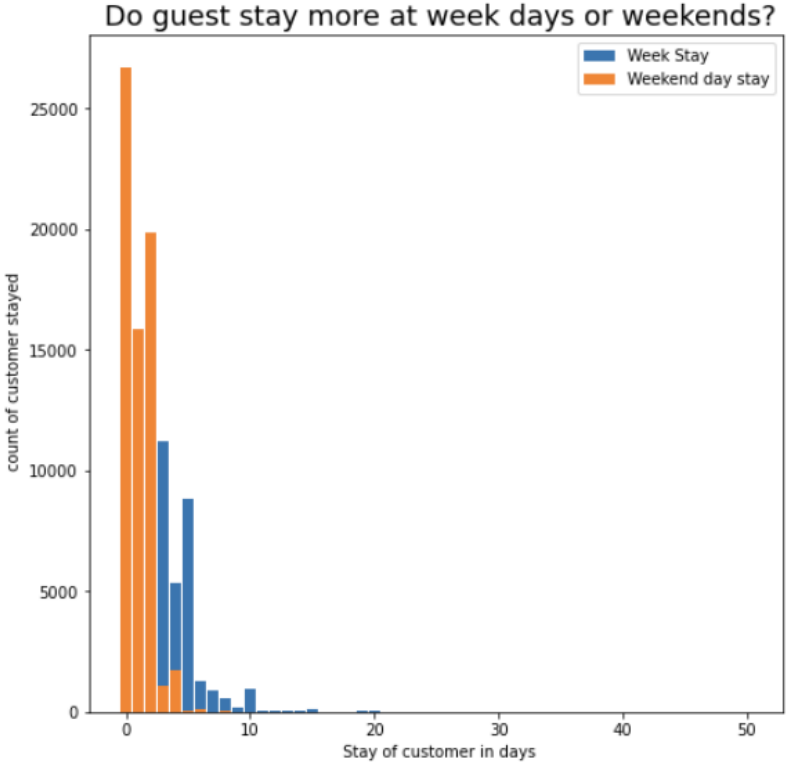


In Resorts, most of the **guest stayed for 1 day**, while for city hotel most of the guest stayed in range of **1-7 days**.

Data Visualization

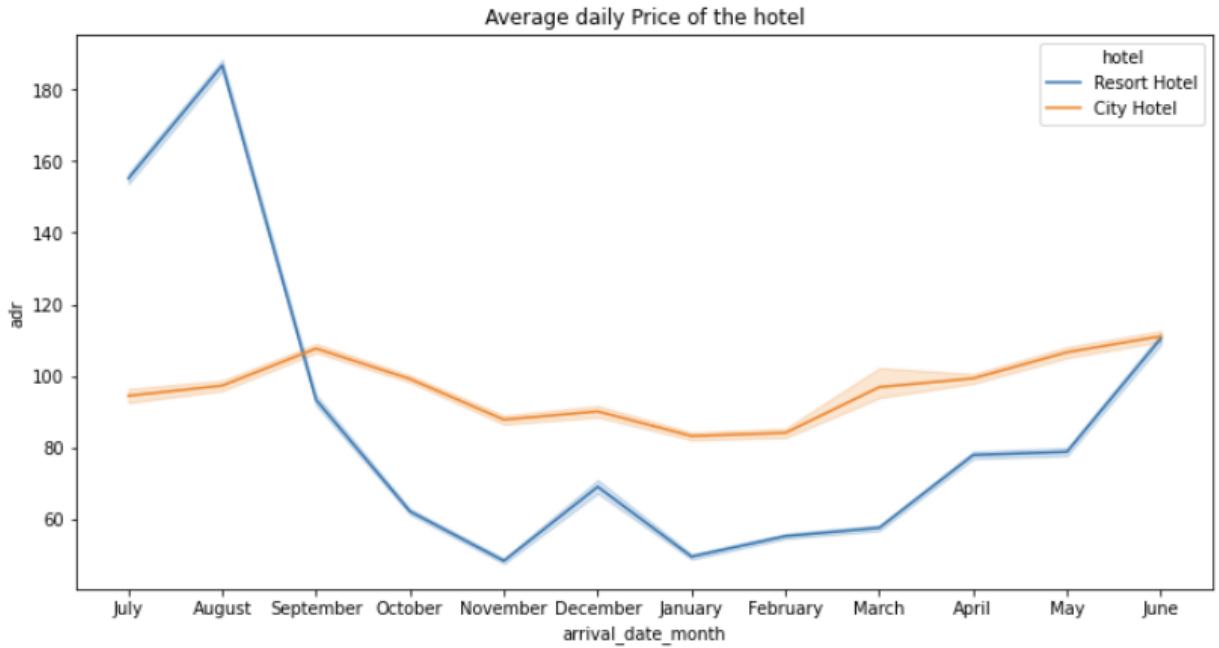
Do guests stay more at week days or weekends?

Most of the guests who stayed for more than 2 days stayed in week days. Thus, it can be inferred most of the guests **prefers weekends over weekdays**.



Data Visualization

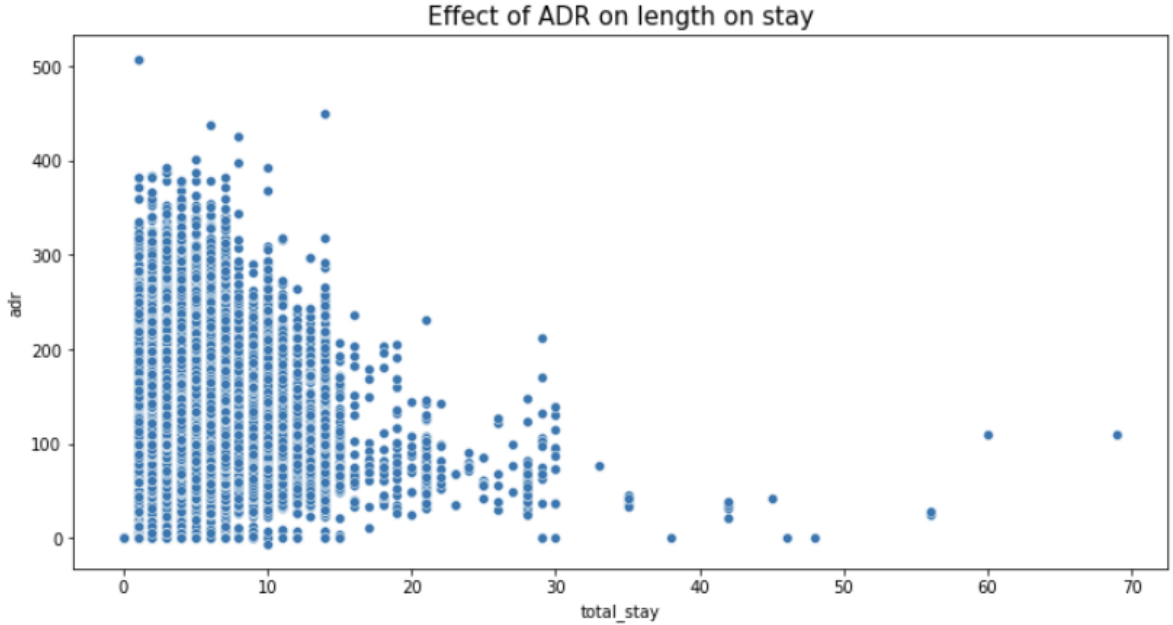
Price trend of hotels?



Average price of **resorts hiked remarkably in August** which drops to lowest in the month of November. **While city hotel prices increased on the month of October** which dropped to lowest in the month of February.

Data Visualization

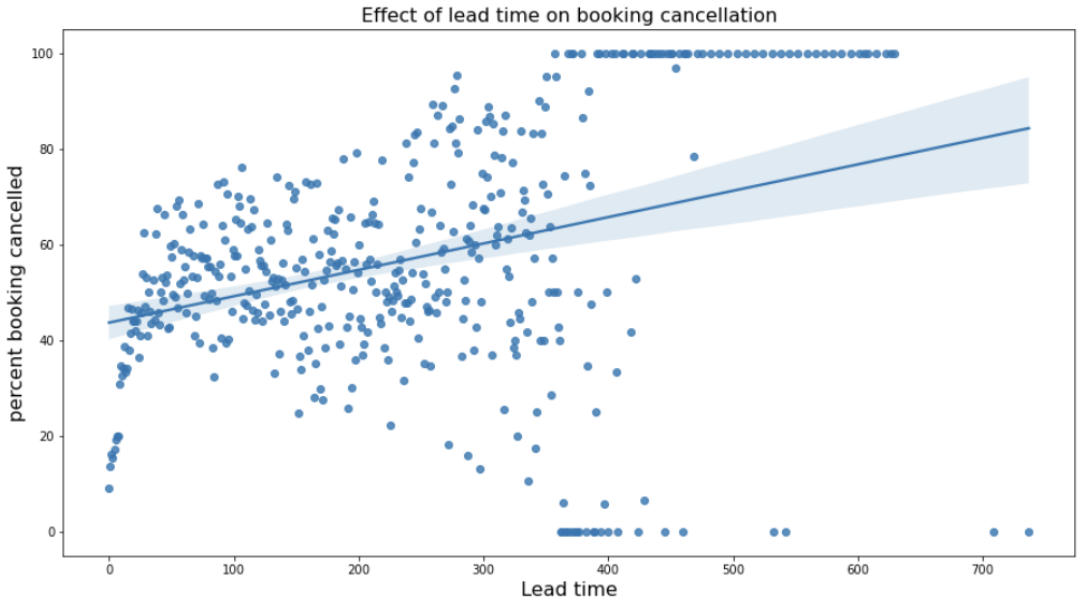
Relation between ADR and length of stay



We can observe, the length of total stay increases as the average daily price decrease. They show **inverse relation**.

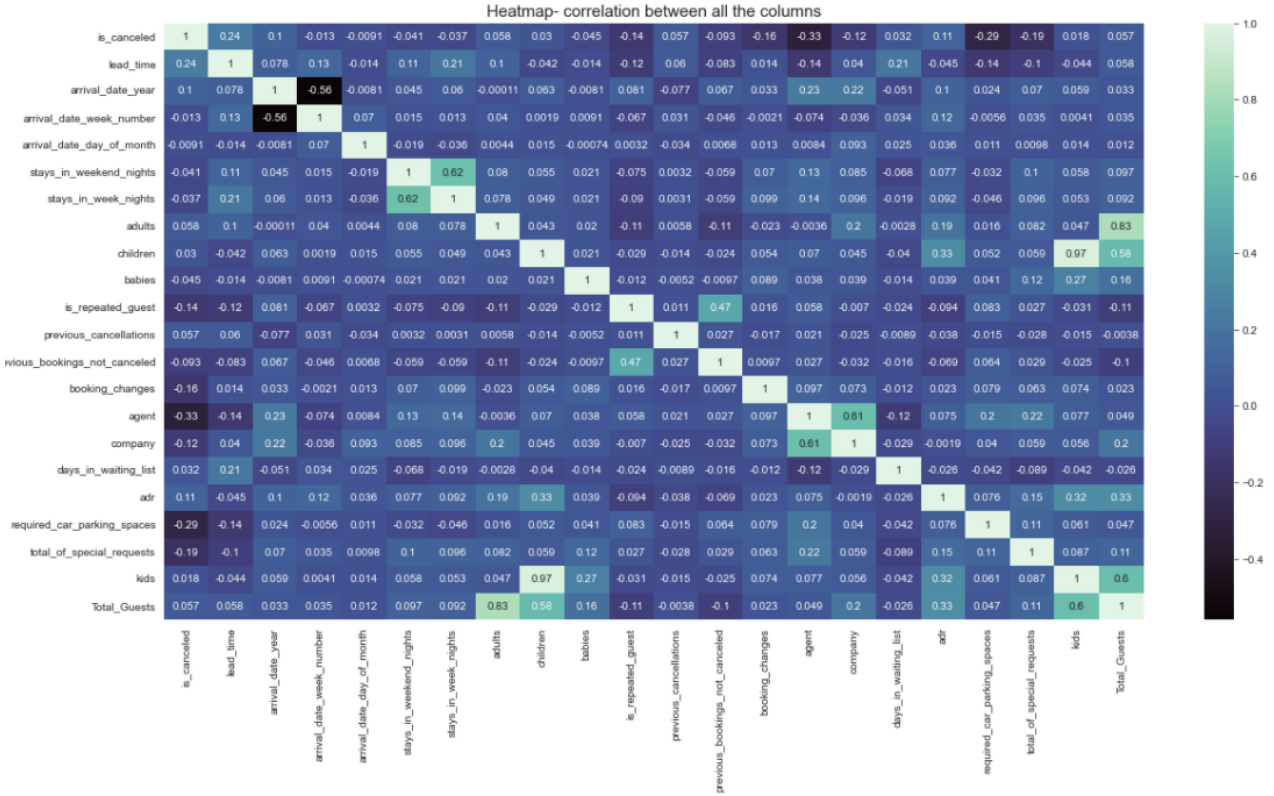
Data Visualization

Effect of lead time on booking cancellation



As we can observe, the plots is scattered i.e. there is likelihood of lesser correlation between lead time and booking cancellation. More lead time is **not related** to cancellation of booking.

Data Visualization



Inferences

a. Most preferred hotel type is 'Resort hotel' which is preferred by 61.13% of total visitors

b. 97.29% guest were not repeated guest. Now, we can further explore based on hotel type.

c. 47.08% customers have cancelled their booking. Now, further exploration needs to be done to understand the possible reason of cancellation.

d. 76.92% hotel provide breakfast, which may help hotel in getting good rating from customer. This could be a potential co-variate for booking status.

e. Market was widely captured by distribution channel like: TA or TA/TO especially by "Online TA" with 44.81% market capture.

f. Around 56% customers were assigned A room type and 68% has reserved it. This also indicates the fact that room type A is on higher demand.

g. As expected from the above insight, nearly 85% customers has not made any changes to their booking.

Inferences

h. nearly 75% of the customers are Transient customer type, with very few customers had booked as group.

i. Most of the hotel does not have any security deposit for booking (86.47%), this could increase the number of booking of those specific hotels.

j. Most of the customers who have booked hotel are 2 in number that is majorly adults have booked the hotel.

k. majority of the hotels (around 91%) had not provided the car parking spaces. Providing additional facilities to customer may increase the booking,

Conclusion

1. Hotel analysis dataset was loaded, cleaned and utilized for exploratory data analysis for the factors that drives the booking of the hotels.
2. Libraries like pandas, matplotlib, seaborn were used to clean, manipulate and visualize the data.
3. Most of the guests preferred resorts over city hotels. However, city hotels were producing more revenue.
4. Portugal was the country which perform most of the booking as well it, cancellation rates were also higher for this country.
5. Many of the potential covariates could be explored more if this project would further extended.

Thank You..

Happy Visiting!