Capstone Project

Exploratory Data Analysis
on
Hotel Booking Analysis

By
PREMSINGH RATHOD

Problem statements

- In the project hotel booking analysis, the problem's occurrences are the booking cancellation, double-booked room or mishandled reservation is a tough but sometimes unavoidable part of the job, and it's still a nightmare of any hotel manager, are the current problems hotels are facing.
- So,implementing some changes in hotel system like giving your guests accurate and helpful information will go a long way to smoothing out their booking process, developing an eye-catching website, user-friendly hotel website and by using a commission-free booking engine are ideas over the above problem solving, but most importantly offering loyalty to customer counts more.
- The main objective 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.

WorkFlow:

- **We will divide our workflow in 3 steps.**

- **EDA will be divided into following 3 analysis**

Data collection and understanding

After collecting data, it's very important to understand data.So,we had hotel booking analysis data with 119390 rows and 32 columns. lets us understand first about 32 columns.

- **hotel** : *Hotels(Resort Hotel or City Hotel)*
- **is_canceled** : *Value indicating cancellation of booking, if the booking was canceled (1) or not (0)*
- **lead_time** :* Number of days that elapsed between the entering date of the reservation booking into the PMS and the arrival date*
- **arrival_date_year** : *Year of arrival date of guest to hotel*
- **arrival_date_month** : *Month of arrival date of guest to hotel*
- **arrival_date_week_number** : *Week number of year of guest arrival date*
- **arrival_date_day_of_month** : *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** : *Type of meal booked. Categories are presented in standard hospitality meal packages:*

Data collection and Understanding

- **country** : *Country of origin.`*
- **market_segment** : *Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”*
- **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** : *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. Code is presented instead of designation for anonymity reasons.*
- **assigned_room_type** : *Code for the type of room assigned to the booking.*
- **booking_changes** : *Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation*

Data collection and Understanding

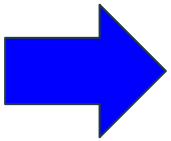
- **deposit_type** : *Indication on if the customer made a deposit to guarantee the booking.*
- **agent** : *ID of the travel agency that made the booking*
- **company** : *ID of the company/entity that made the booking or responsible for paying the booking.*
- **days_in_waiting_list** : *Number of days the booking was in the waiting list before it was confirmed to the customer*
- **customer_type** : *Type of booking, assuming one of four categories*
- **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, assuming one of three categories*

Data Cleaning and Manipulation

- **Handling Duplicates:Data had 31994 rows of duplicate values.So, we dropped it from data:True are Duplicated rows**

```
# Dataset Duplicate Value Count
duplicated_rows=df.duplicated().value_counts()
duplicated_rows
```

```
False    87396
True     31994
dtype: int64
```



```
[ ]  df= df.drop_duplicates()
     df.shape
```

```
(87396, 32)
```

- **Handling Null Values:There are 4 columns company,agent,country and children with null values.**

Identifying the values and replacing it with 0's and other

```
[12] missing_values =   df.isna().sum().sort_values(ascending = False)[:4]
     missing_values
```

```
company     82137
agent       12193
country       452
children        4
dtype: int64
```

```
# Filling/replacing null values of columns agent children and company with 0.
null_columns=['agent','children','company']
for col in null_columns:
  df[col].fillna(0,inplace=True)
# Replacing NA values of country column with 'others'.
df['country'].fillna('others',inplace=True)
```

Data Cleaning and Manipulation

- **Feature Engineering:We created two columns**

1. ‘Total_stay’ = added Weekend nights and Weekdays night.

```
df["Total_stay"] = df["stays_in_weekend_nights"] + df["stays_in_week_nights"]
```

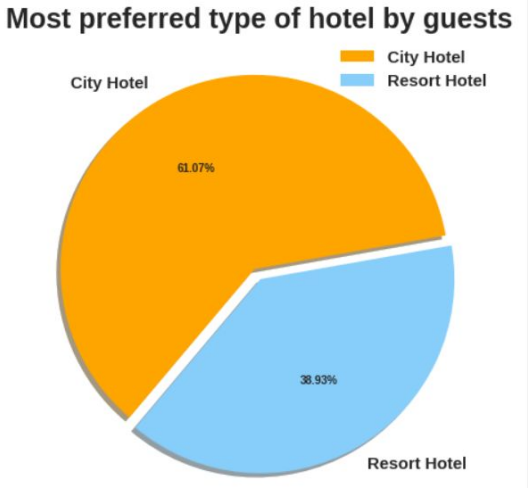
2. ‘’Total_people’ = Added three columns children,adult and Babies.

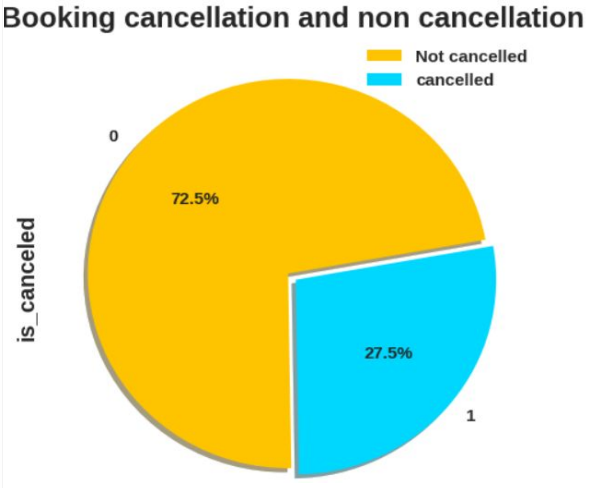
```
'''droppping all 166 those rows in which addtion of of adlults ,children and babies is 0.
That simply means  no bookings were made.'''
df["total_people"] = df["adults"] + df["children"]+ df["babies"]
df.drop(df[df["total_people"] == 0].index,inplace =True)
df.shape
```

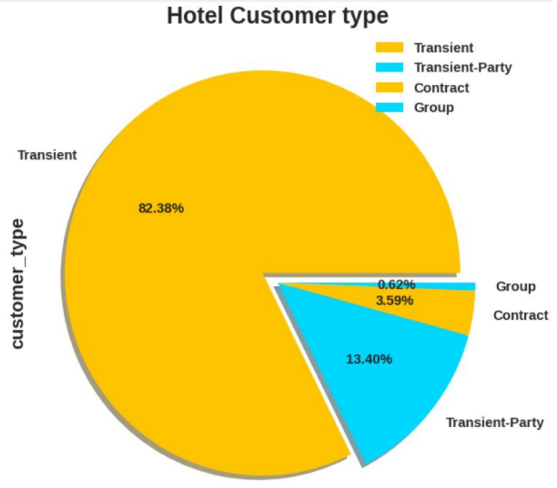
➔ 166 rows dropped that after adding three columns to total people gets 0 , all 0 rows were dropped.From rows 87396 to 87229

```
(87229, 35)
```

Exploratory Data Analysis(EDA)







Conclusion:

- ★ City hotel is most preferred hotel type by guests with 61.1% . So, we can say city hotel is busiest hotel .
- ★ Hotel has 27.5% of cancellation of booking.
- ★ Most of the customers/guests were Transient type(82.38%), followed by transient-party(13.40%) and contract with 3.59% remaining goes to group.

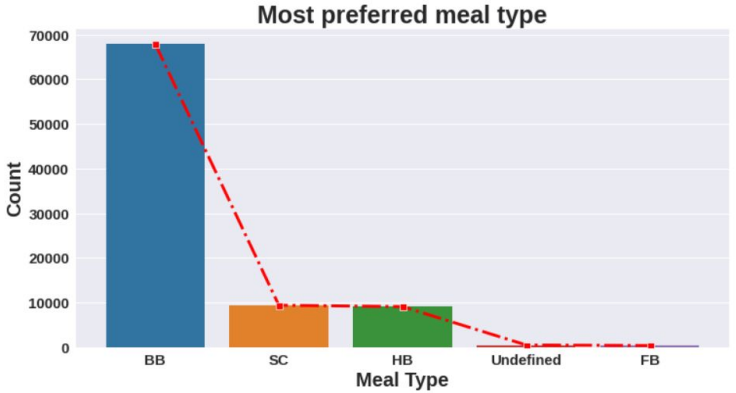
Exploratory Data Analysis(EDA)

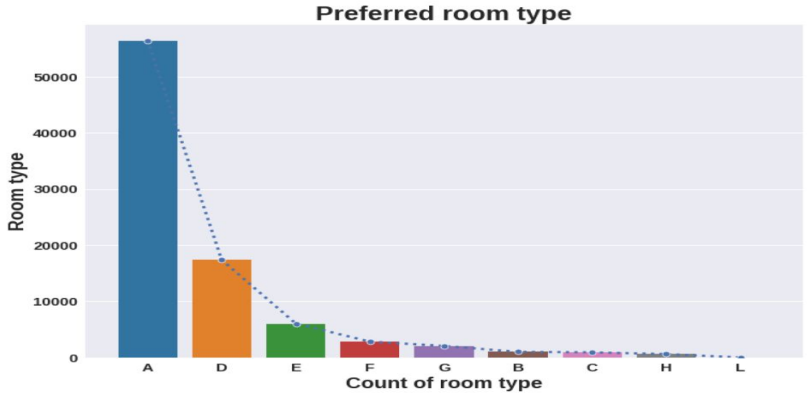
Conclusion:

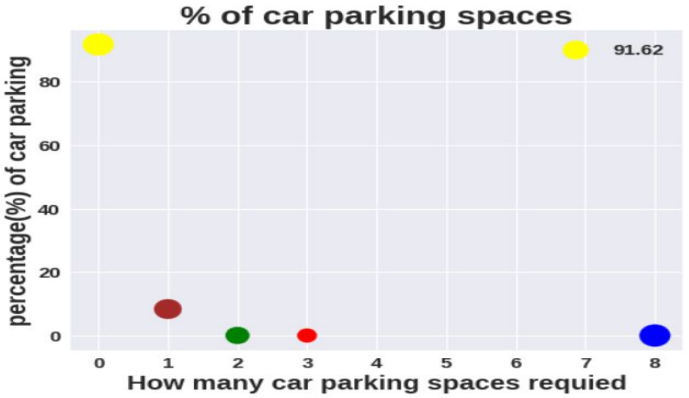
- ★ **Almost 82% of bookings were not changed by guests and followed by 1 time changes made above 10% .And highest changes made was 18 by one guest.**
- ★ **June and August months had the rapid growth of bookings. So, summer vacation can be reason for bookings.**

Exploratory Data Analysis(EDA)

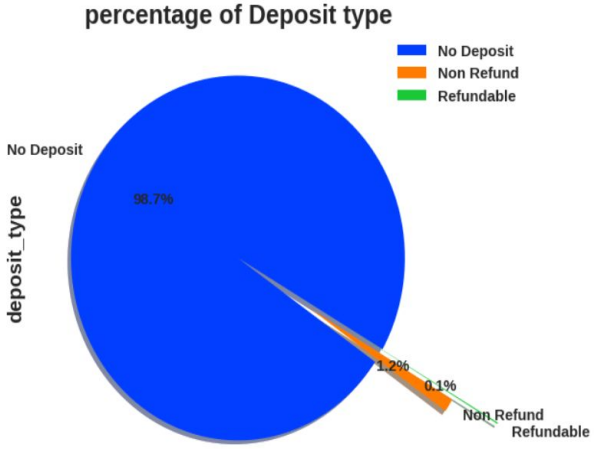
- ★ BB type food is most mentioned food while reservation compared to SC and HB
- ★ So the most preferred Room type is "A". Followed by D and E
- ★ 0 has the highest value with 91.6 % that doesn't mentioned special request about car parking spaces.





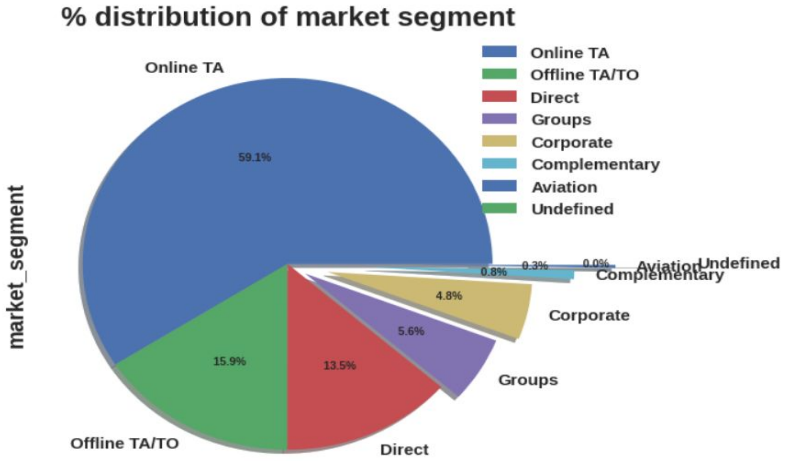


Exploratory Data Analysis(EDA)

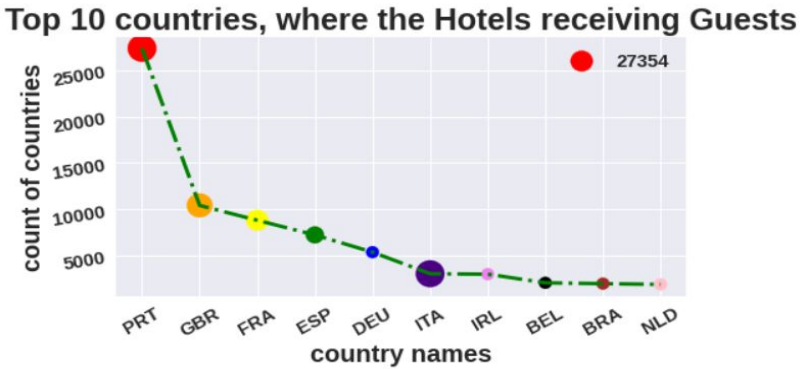


No deposit" guest type has more than 98% of arrivals, where as booking associated with Refundable has lowest with 0.1%.

'Online TA' is mostly(59.1%) used for booking hotels. 'Offline TA/TO' and 'Direct' segment holds more than 29%.

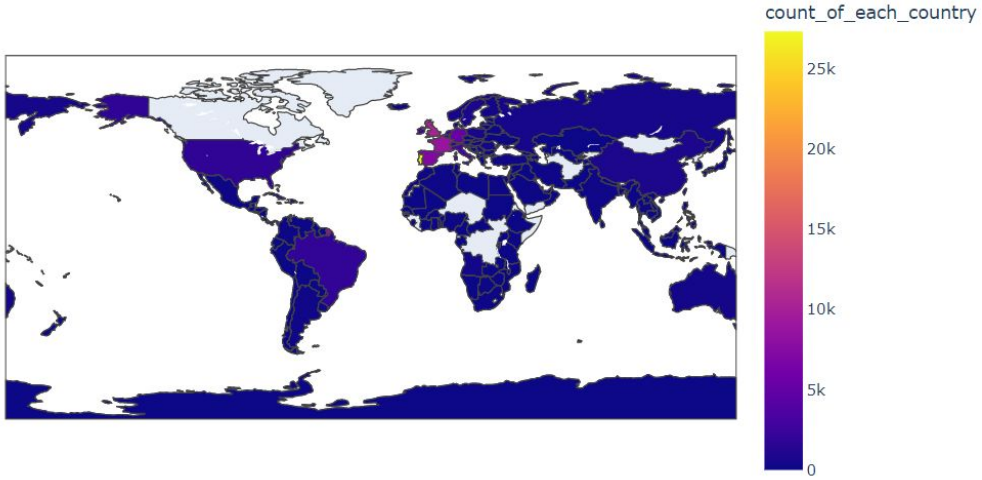


Exploratory Data Analysis(EDA)

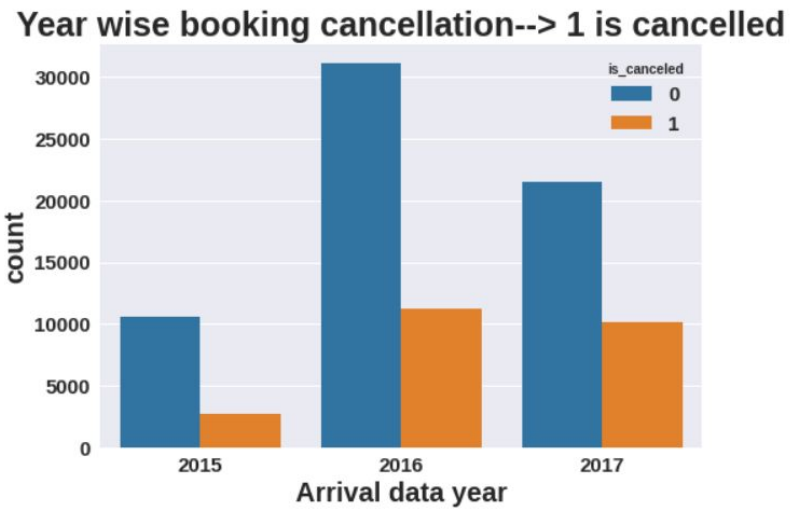


Guests are from total 178 countries among that most of the guests are coming from portugal, that is more than 27000 guests are from portugal.
Followed by United kingdom and France.

1. PRT- Portugal
2. GBR- United Kingdom
3. FRA- France
4. ESP- Spain
5. DEU - Germany
6. ITA -Italy
7. IRL - Ireland
8. BEL -Belgium
9. BRA -Brazil
10. NLD-Netherlands



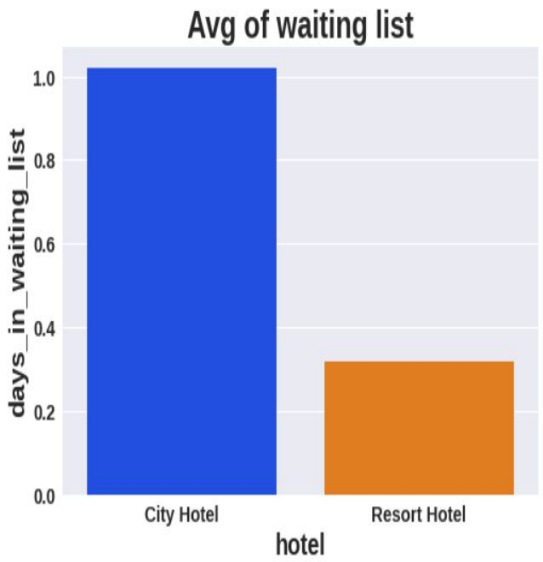
Exploratory Data Analysis(EDA)

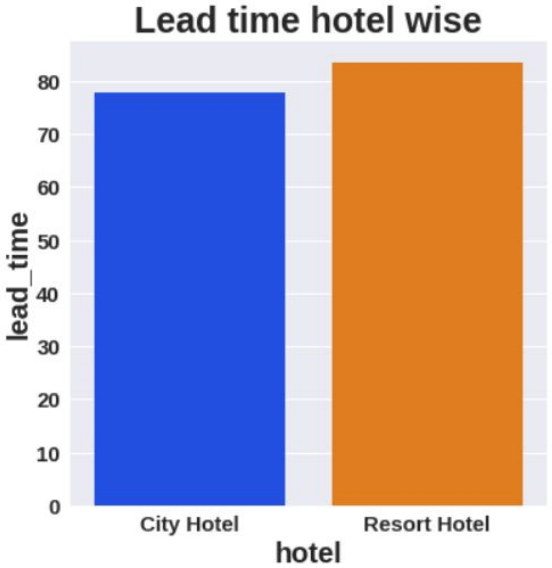


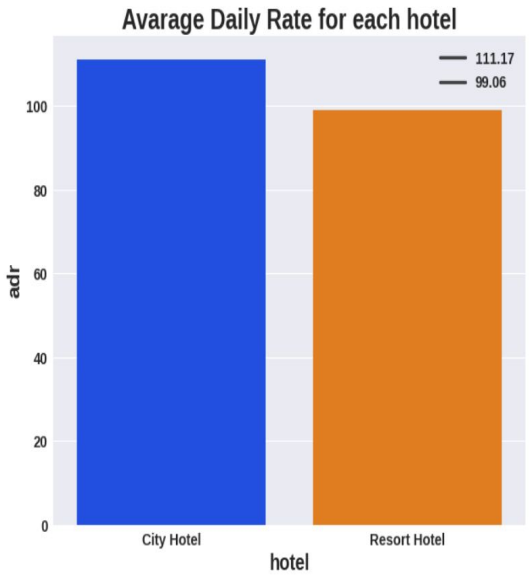
★ In 2016 the cancellation of booking was more than 10000 later next year 2017 down fall to margin of 10000.

★ Except 2015 other two years city hotel has more booking as compared to resort hotel.In 2016 has more bookings for both hotels as compared to 2015 ands 2017

Exploratory Data Analysis(EDA)

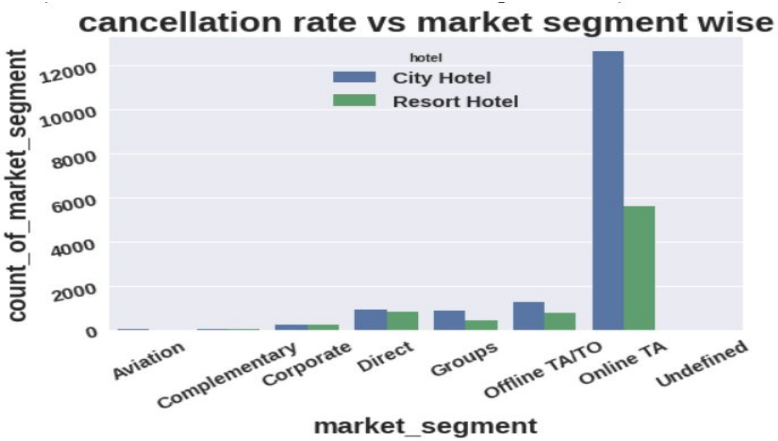




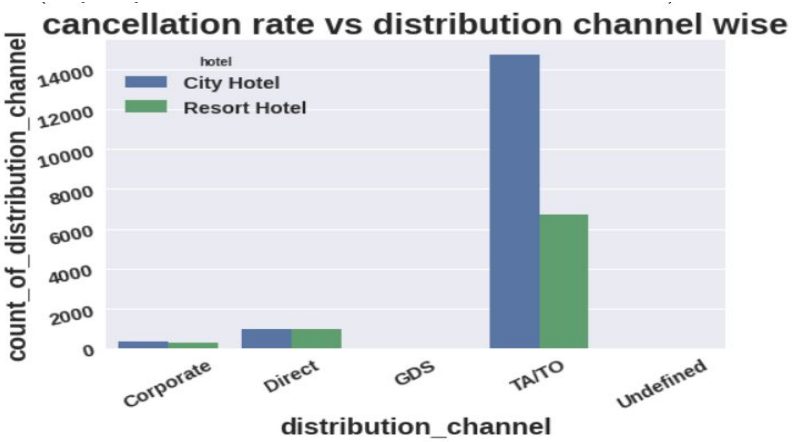


- City hotel has highest in average of waiting list of days more than 1 day compared to resort hotel.
- Resort hotels has slightly high avg lead time.
- City hotel has higher Average Daily Rate of 111.27 as compared to resort hotel.

Exploratory Data Analysis(EDA)

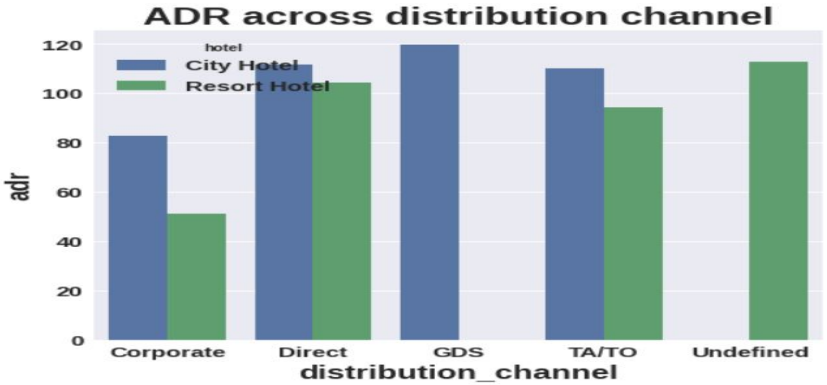


Online TA/TO market segment has the highest booking cancellation in both the hotel types, city hotel with more than 12000 and resort hotel with more than 5000.



TA/TO distribution channel has the highest booking cancellation in both the hotel types, city hotel with more than 14000 and resort hotel with more than 60(

Exploratory Data Analysis(EDA)



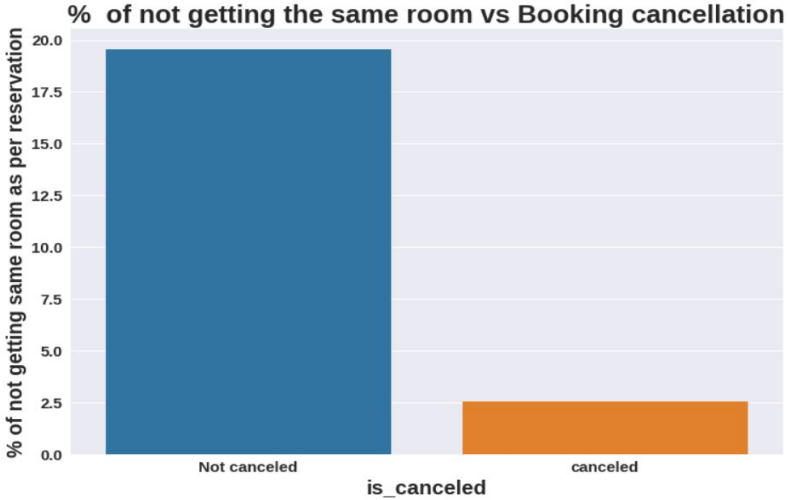
★ 'Direct' and 'Online TA' are contributing the most in both types of hotels.

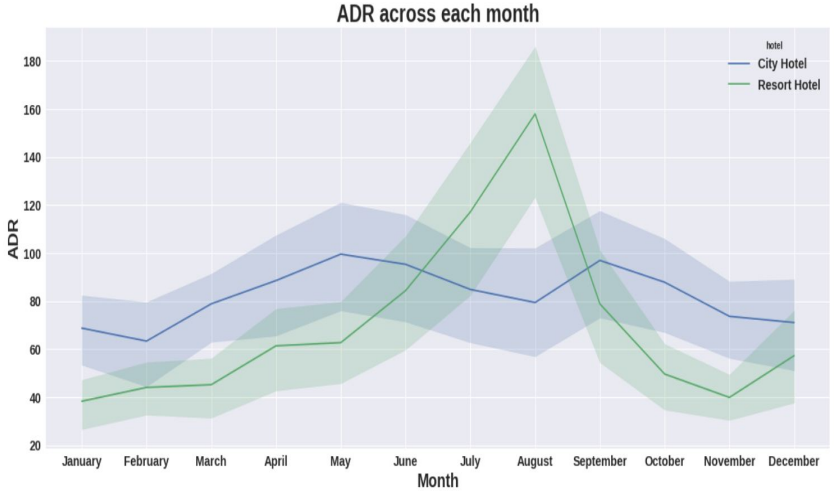
TA is Travel agent

TO is Tour Operator

★ Global distribution system(GDS) channel has the highest adr for city hotel.for resort hotel GDS need to start reservation service.And at some undefined channel need to focus on city hotel reservation too.

Exploratory Data Analysis(EDA)

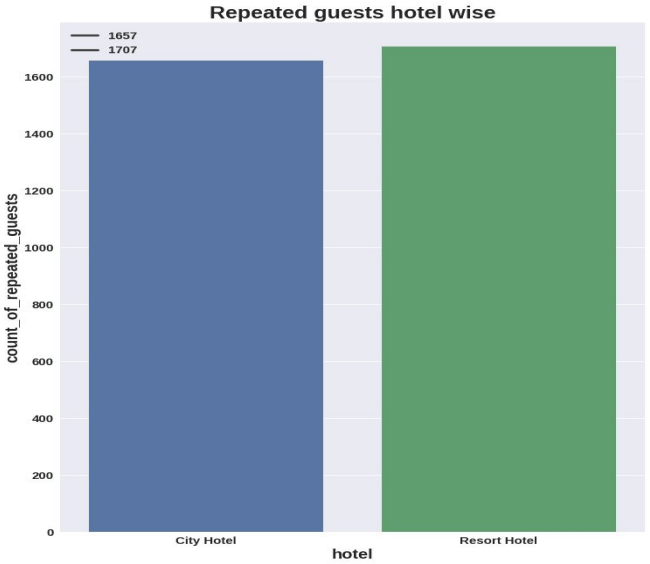




- ★ Booking cancelled percentage is (2.5%) and not cancelled is more than (19%), no much effect on cancellation of the bookings even if the guests are not assigned with rooms which they reserved during booking but with 2.5%.

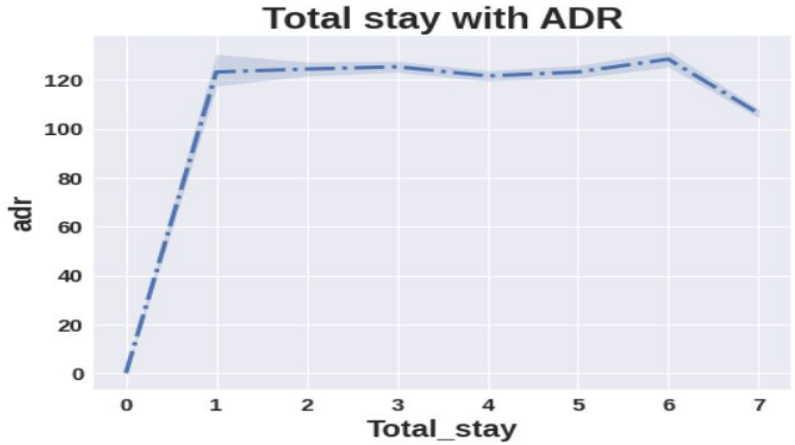
- ★ For Resort hotel is ADR is high in the months June,July,August as compared to City Hotels.City hotel has avg ADR difference is 40 range for every months.

Exploratory Data Analysis(EDA)



- ★ Resort hotel has slightly more repeated guests of 1707 as compared to city hotel.And difference of 50 repeated guests.
- ★ No much effect on cancellation of the bookings even if the guests are not assigned with rooms which they reserved during booking but with 2.5%.

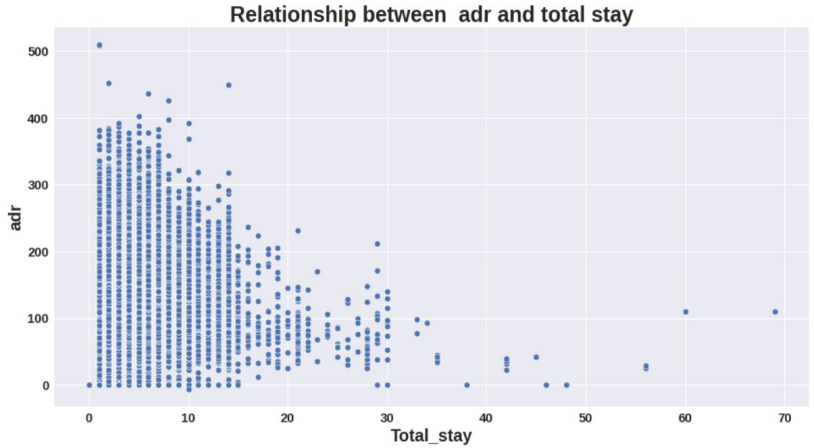
Exploratory Data Analysis(EDA)



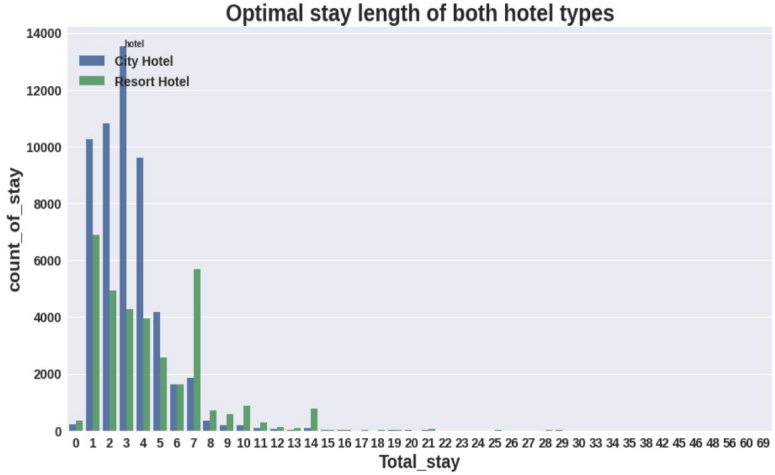
ADR has approximately same to total stay till 6 stays.

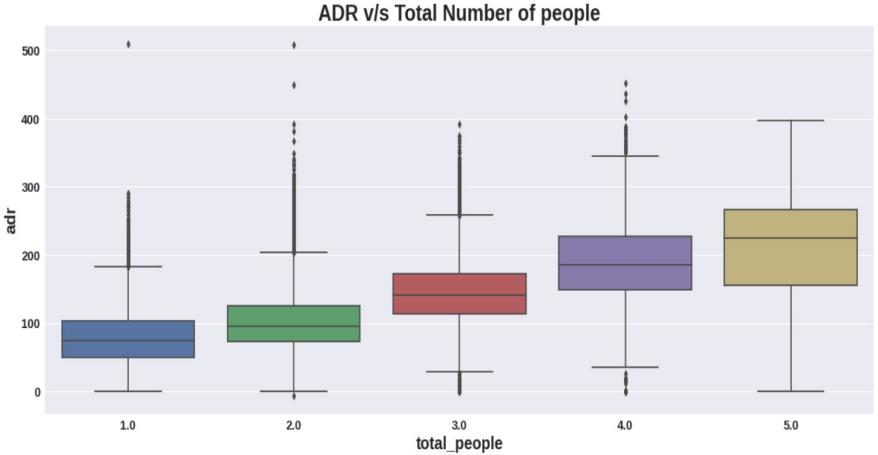
As the total stay increases the adr also increases slightly.

From above scatter we can say that as the stay increases adr is decreasing. Thus for longer stays customer can get good adr



Exploratory Data Analysis(EDA)



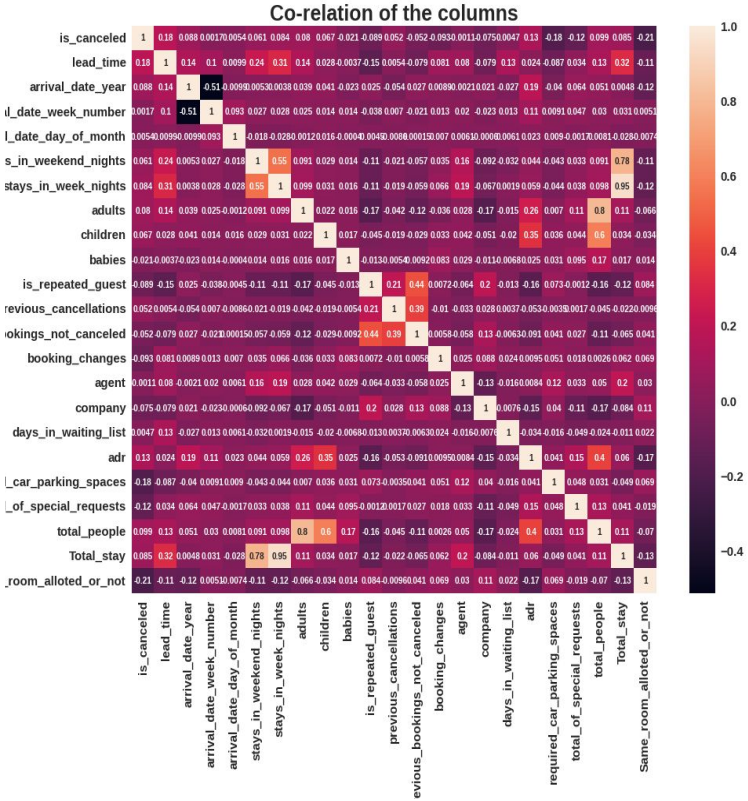


★ **Total stay raised for three days than we can see down fall gradually.Three days stays has highest mark above 13000 for city hotel.**

★ **Adr and total people are directly proportional to each other.Mean while outlier of adr is more for some total_people values. so, property management system need to rectify such errors.**

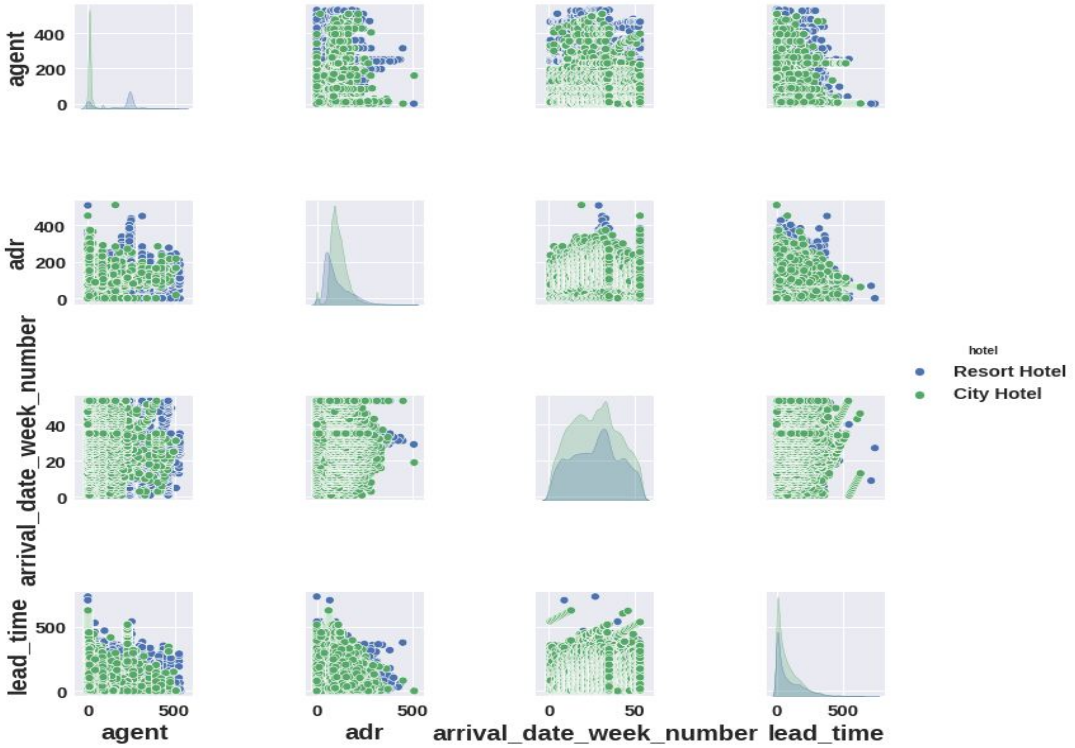
Exploratory Data Analysis-Correlation

- ★ lead_time and total_stay is positively correlated. That means more in the total stay of customer will relate more in lead time.
- ★ adults,childrens and babies are positive correlated to adr. That means more the people more will be adr.
- ★ is_repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.
- ★ is_canceled and same_room_alloted_or_not are negatively correlated. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.



Exploratory Data Analysis(EDA)-Pairing

- ★ Diagonal charts of pair plot are univariate variable and rest of 12 chart are bivariate variable
- ★ City hotel shows trend in all the univariate variable compared to resort hotel which shows chart across diagonal of pair plot.
- ★ Other than diagonal chart,12 charts are bivariate shows trends between two variables like arrival date week number. Agent,adr and lead time.
- ★ Agent and lead time shows resort hotel has greater lead time than city hotel.
- ★ lead time and adr shows adr higher for resort hotel.
- ★ Arrival date week number and lead time shows highest lead time records by Resort hotel.



Solution to Business Objective

- ★ Offering their guests the highest levels of luxury through personalized services, will helps in generating repeated guests
- ★ Assigning the same room type as reserved could help in decreasing the cancellation.
- ★ Offering the special discounts on non holidays periods by applying strategies.
- ★ Market segment and distributional channel should focus on improvements, overbooking by applying individual strategies.
- ★ PMS property management system need to rectify post technical glitches, as guests should not face issues while bookings.
- ★ Providing delicious food and minimising the waiting list time helps out for business development.

Thank You!!