Capstone Project-1

EDA ON HOTEL BOOKING ANALYSIS

BY

TUSHAR.V.CHASKAR

(COHORT MADRID)

❖ Problem Statement :

- For this project we will be analyzing Hotel Booking data. This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces.

- The hotel industry is a particularly dynamic sector, and bookings depend on the aspects listed above as well as many more.

- The major goal of this project is to study and analyse data to identify significant elements that influence bookings and provide management with information so they can run various campaigns to increase revenue and performance.

❖ Work Flow :

- I have divided my work flow into 3 steps which are as follow :
  1. Data Collection And Understanding.
  2. Data Cleaning And Data Manipulation.
  3. Then Exploratory Data Analysis i.e EDA

- Exploratory Data Analysis is divided into following 3 analysis :

1.**Univariate Analysis :** Univariate analysis is basically the simplest form to analyze data. Uni means one and this means that the data has only one kind of variable. The analysis will take data, summarise it, and then find some pattern in the data.

2.**Bivariate Analysis :** Bivariate analysis refers to the analysis of two variables to determine relationships between them. Bivariate analyses are often reported in quality of life research.

3.**Multivariate Analysis :** The statistical study of data where multiple measurements are made on each experimental unit and where the relationships among multivariate measurements and their Structure are important.

❖ Data Collection And Understanding :

The Understanding of given data is very important for doing analysis On it. I have Data of Hotel Bookings which has total 1,19,390 rows and 32 columns in it, so let understand what this 32 columns are describing.

Data Description :

1. **Hotel :** Showing Type of Hotel i.e City Hotel or Resort Hotel.
2. **is_canceled :** Indicating the value for canceled(1) And Not Canceled(0).
3. **lead_time :** The number of days that passed between the booking date And the arrival date after it was entered into the PMS.
4. **arrival_date_year :** Showing Year of Arrival Date.
5. **arrival_date_month :** Showing Month of Arrival Date.
6. **arrival_date_week_number :** Showing the Week number of year for arrival date.
7. **arrival_date_day_of_month :** Showing the Day of arrival date.
8. **stays_in_weekend_nights :** Showing the Number of Saturday or Sunday nights the visitor spent or reserved to spend at the hotel.
9. **stays_in_week_nights :** The Number of weeknights (Mon to Fri) the visitor stayed or made a reservation at the hotel.

❖ Data Collection And Understanding :

10. **adults :** Showing the Number of Adults.
11. **children :** Shwing the Number of Childrens.
12. **babies :** Showing the Number of Babies.
13. **meal :** Showing the booked meal type. Category presentation are made in typical hospitality meal packages.
14. **country :** Showing the Country of Origin.
15. **market_segment :** Showing Names of a Market Segment, The term 'TA' And 'TO' stands for 'Travel Agents' And 'Tour Operators' respectively in categories.
16. **distribution_channel :** Showing Distribution method of Reservations. Travel Agents are referred as 'TA' And Tour Operators are refered as 'TO'.
17. **is_repeated_guest :** Showing value indicating if the booking name was from a repeated guest (1) or not (0).
18. **Previous_cancellations :** Showing the Number of past reservations that the consumer cancelled before the current reservation.
19. **previous_booking_not_cancelled :** Sowing the number of earlier reservations that the customer did not cancel before the current resevation.

❖ Data Collection And Understanding :

20.**reserved_room_type :** Showing the reserved room type code. To preserve anonymity, code is used in place of designation.
21.**assigned_room_type :** Showing the Code for the room type which is assigned to the booking.
22.**booking_changes :** Showing the Number Modification/Changes made to the reservations from the time it was entered on the PMS to check-in or cancellation.
23. **deposit_type :** Indicate whether the customer paid a deposit to secure the reservation.
24. **agent :** Showing the ID of the travel agency that made the booking.
25.**company :** Showing the ID of the company or other entity that made the reservation or is in charge of making the payment.
26. **days_in_waiting_list :** Showing how long the reservation was on wait before the consumer received the confirmation.
27.**customer_type :** Showing the type of booking, assuming one of four category
28. **ADR :** ADR is determined by dividing the total of all hotel transactions by the total number of resevation.

❖ Data Collection And Understanding :

29.**required_car_parking_spaces :** Showing the Number of parking spaces the consumer needs.

30.**total_of_special_requests :** Showing the Number of special requests made by the customer.

31. **reservation_status :** Reservation status, assuming one out of three categories.

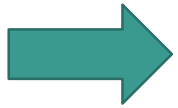
32.**reservation_status_data :** The most recent data the status was set. To determine when the reservation was cancelled or the customer checked out of the Hotel, use this variable in conjunsction with the Reservation status.

❖ Data Cleaning And Data Manipulation :

- **There is four columns which has missing values, i.e company, agent, country and children.**

```
# Checking the null values.
df1.isna().sum().sort_values(ascending=False)[:4]
```

	Columns	Null values
0	company	82137
1	agent	12193
2	country	452
3	children	4



```
# Filling/replacing null values with 0.
null_columns = ['company','agent','children']
for col in null_columns:
  df1[col].fillna(0,inplace=True)

# Filling/replacing null values with Others.
df1['country'].fillna('Others',inplace=True)
```

```
# Checking the null values.
df1.isna().sum().sort_values(ascending=False)[:4]
```

Columns	Null values
hotel	0
days_in_waiting_list	0
reserved_room_type	0
assigned_room_type	0

- **Handling Duplicates : In this data I found 31994 duplicates values so I dropped that duplicates from the data.**

```
# checking for the duplicate rows. # True Means Duplicated Rows.
x = df1.duplicated().value_counts()
```

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

❖ Data Cleaning And Data Manipulation :

➢ **Feature Engineering :**

I have created two coulums named as follows :

1.‘Total_No_of_Person’ = This column is created with the addition of children column, adults column, and babies column.

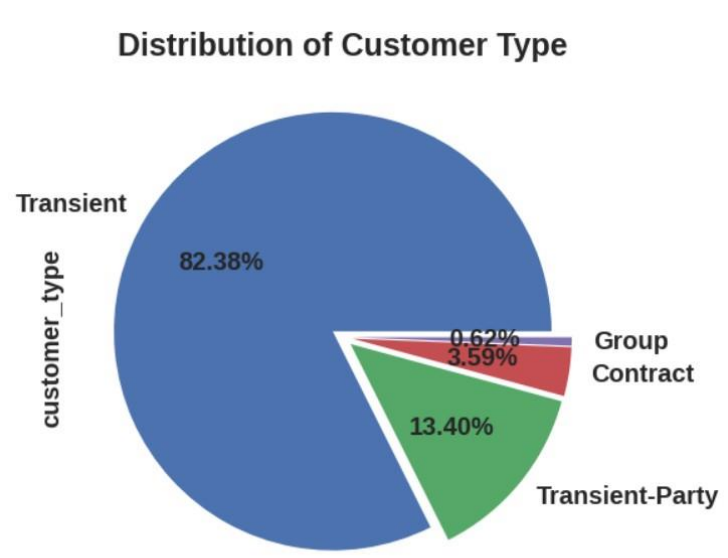
2.‘Total_Stay’ = This column is created with the addition of stays in week nights and stays in weekend nights.

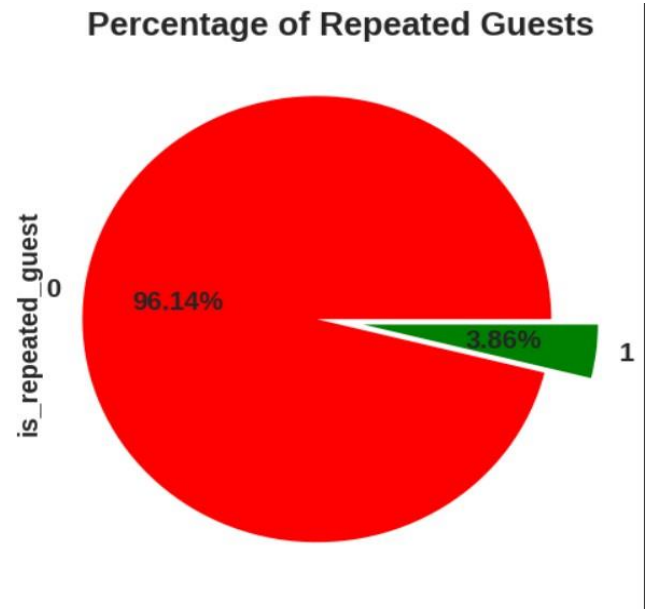
After that I have dropped the rows where total no of person is zero.

```
# Checking the unique values in categorical columns.
category_col = list(set(df1.drop(columns=['reservation_status_date','country','arrival_date_month'])) - set(df1.describe()))
for col in category_col:
  print(f'The Unique Value in {col} is {(df1[col].unique())}')
```

```
# lets add total stay column.
df1['Total_Stay'] = df1['stays_in_week_nights'] + df1['stays_in_weekend_nights']
```

❖ Exploratory Data Analysis (EDA) :





Conclusion :

- **The Most of customers is in type of Transient i.e 82.40 %, then the Transient Party includes 13.40 % customers, then 3.60 % customers belongs to Contract type and Remaining customers is in Group i.e 0.6 %.**
- **The Most preferred Hotel type by the customers are City Hotels, it means City Hotels is busy as compare to Resort Hotels.**
- **Out of all the bookings only 3.90 % customers are revisited to the Hotel rest 96.1 % was new customers it means the retention rate is very low.**

❖ Exploratory Data Analysis (EDA) :

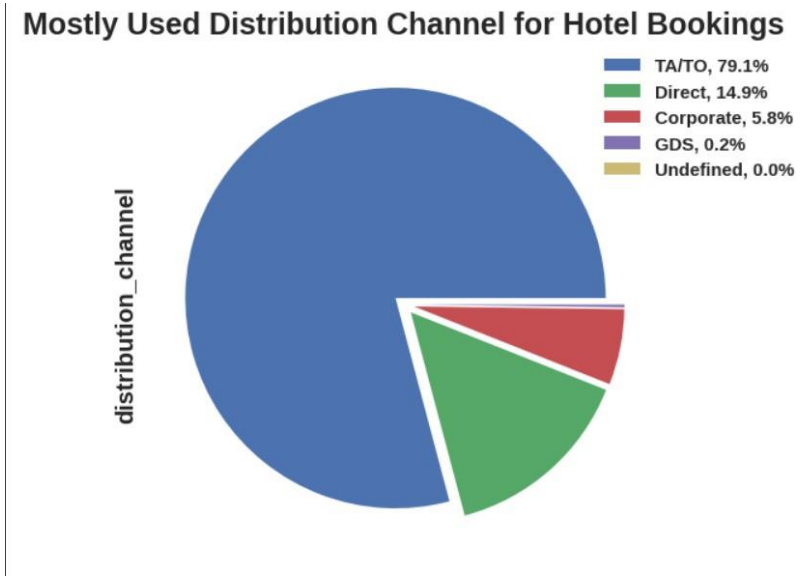
Conclusion :

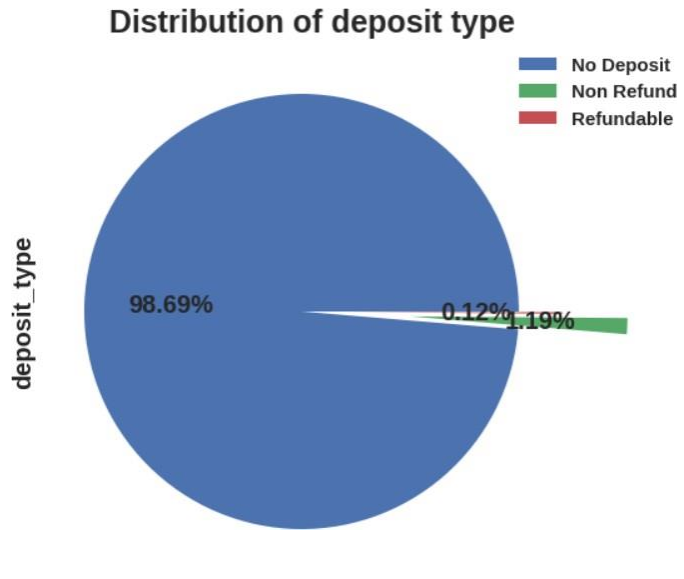
The percentage of zero booking changed by customers is more than 82 %. The percentage of two times booking changed is more than 10 % And The percentage of three and four times booking changed is less than 5 %.

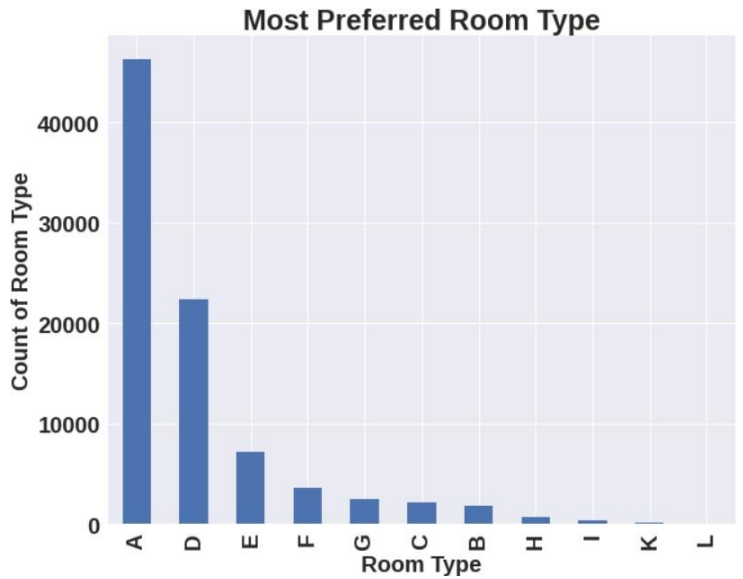
Conclusion :

The highest bookings made by the Agent id No - 9.0 which is more than 28000. Agent id No – 240.0 also made large amount of booking which is more than 13000.

❖ Exploratory Data Analysis (EDA) :



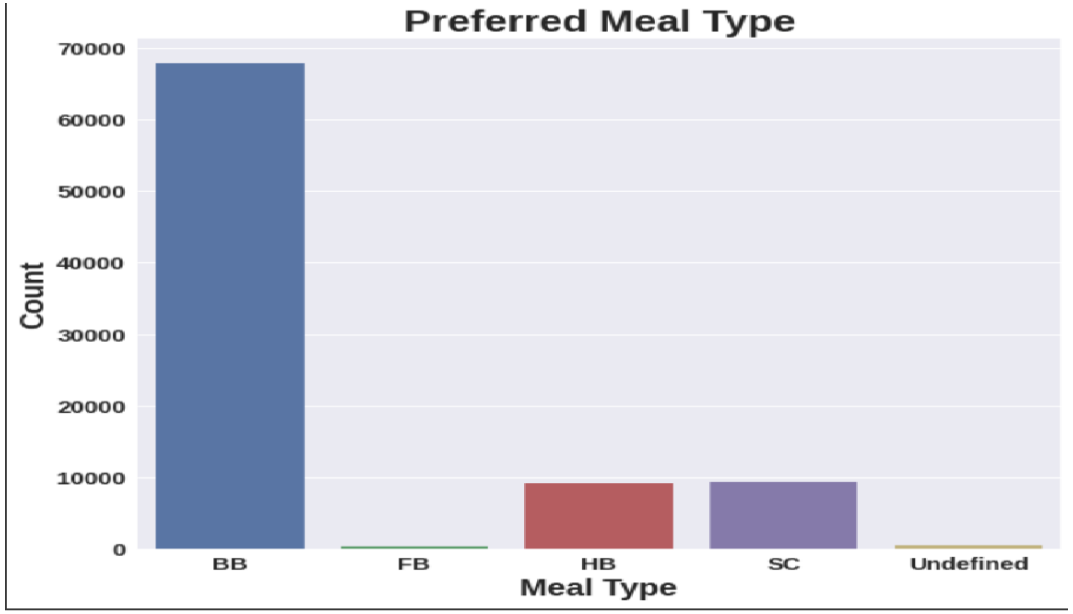




Conclusion :

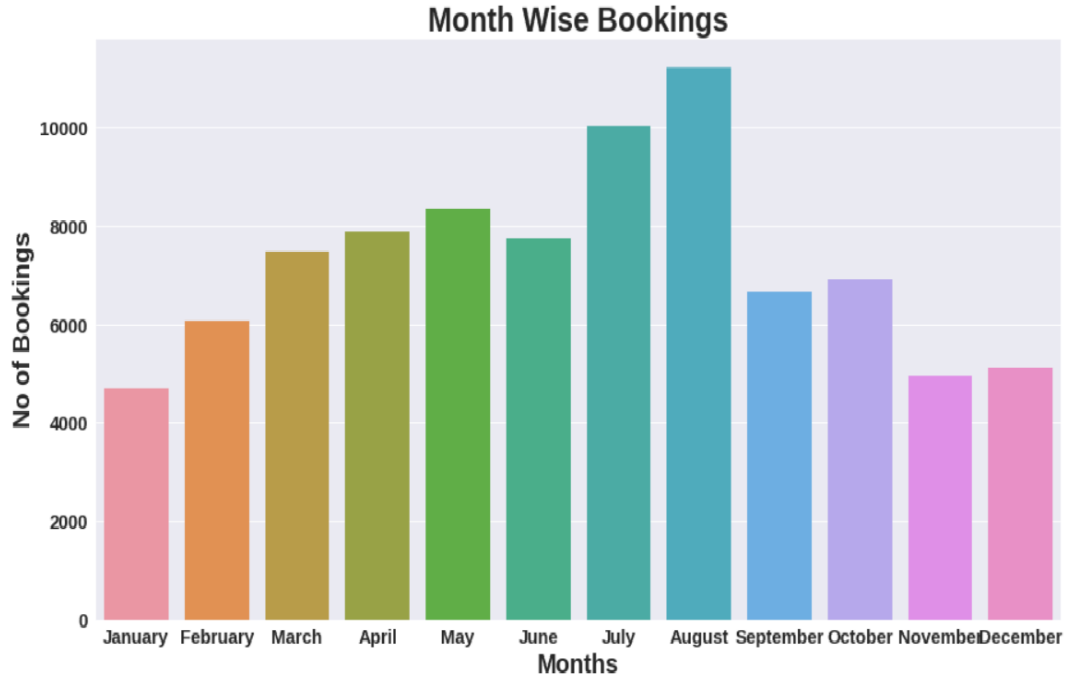
- **The highest bookings was made by the Travel Agents And Tour Operator (TA/TO) i.e 79.10 %. The second most highest bookings was made through Directly from Customers i.e 14.90 %.**
- **Out of all the bookings made by customers almost 98.70 % customers prefer 'No Deposit' type while doing bookings.**
- **The Most preferred room type by customers is 'Type A' And the second most preferred room type by customers is 'Type D'.**

❖ Exploratory Data Analysis (EDA) :



Conclusion :

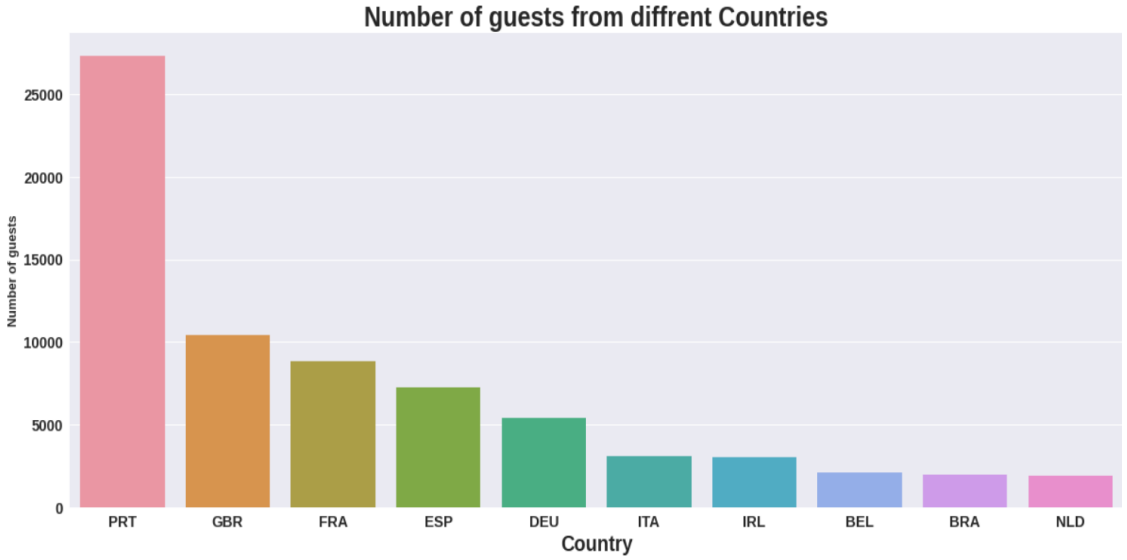
- The Most preferred meal type by customers is Bed And Breakfast (BB).
- The Least preferred meal type by customers is Full Board (FB).
- Half Board (HB) And Self Catering (SC) is average preferred meal type by customers.



Conclusion :

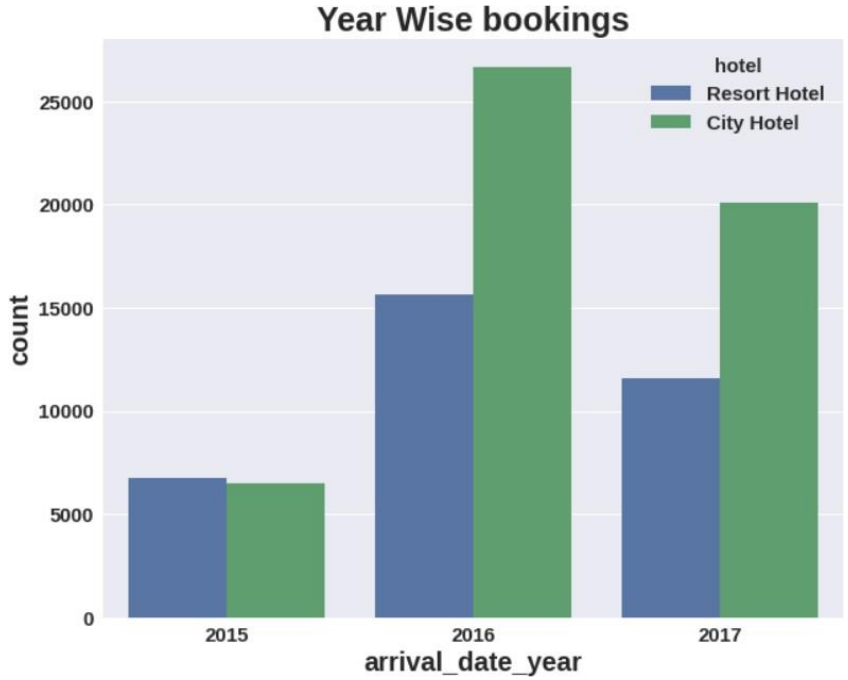
We can see the most of the bookings made is in July And August, maybe the reason is Rainy Season. After August month the bookings are less. In March, April, May And June there is second most bookings maybe the reason is Summer season.

❖ Exploratory Data Analysis (EDA) :



Conclusion :

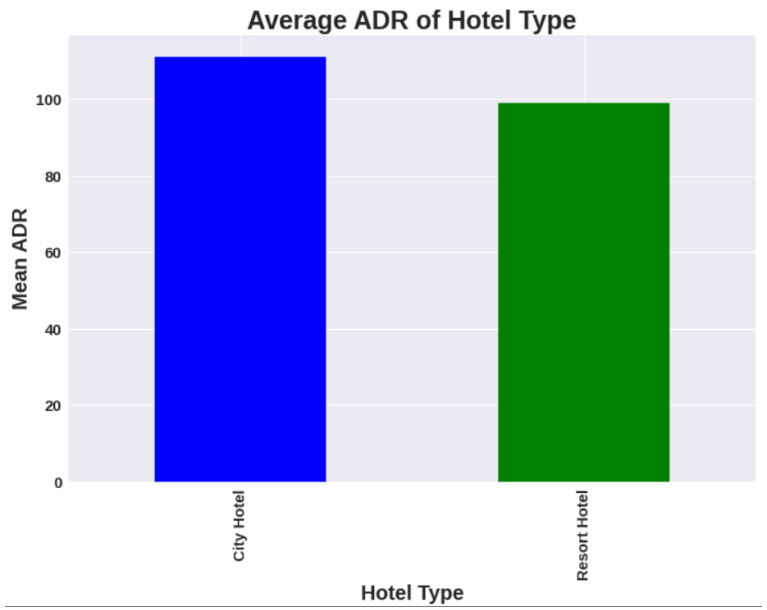
- The highest booking made by customers was from Portugal i.e more than 26000.
- Then After Portugal second most highest booking made by customers was from Great Brittan, France And Spain

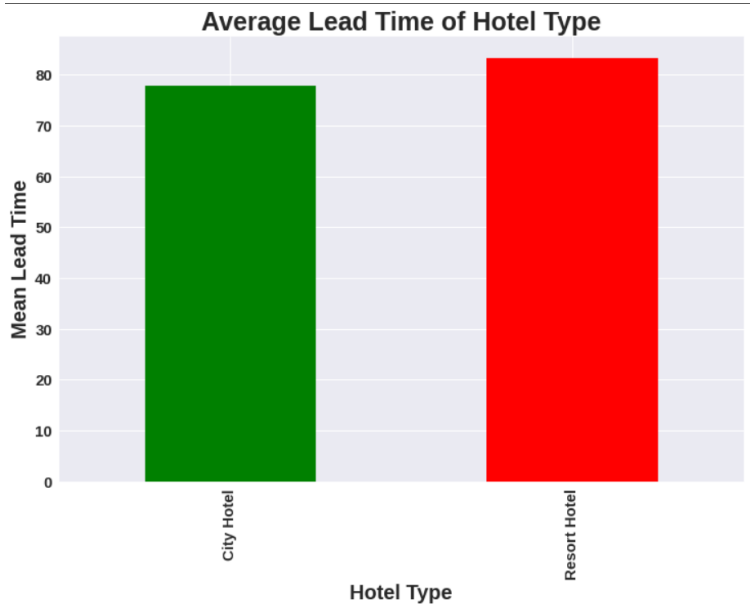


Conclusion :

As we can see in 2016 the most of bookings for City Hotels And Resort Hotels was done. In 2016 booking of City Hotels are more than Resort Hotel.

❖ Exploratory Data Analysis (EDA) :

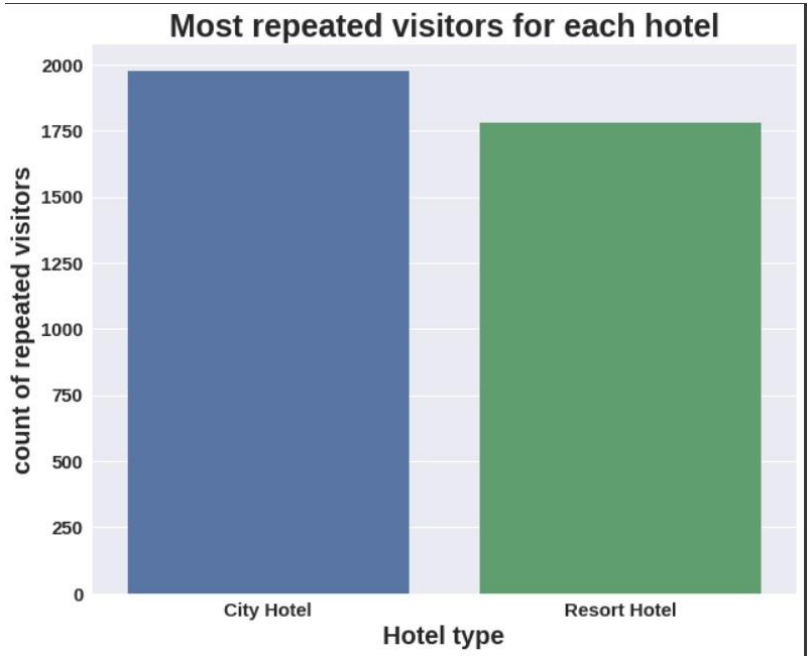


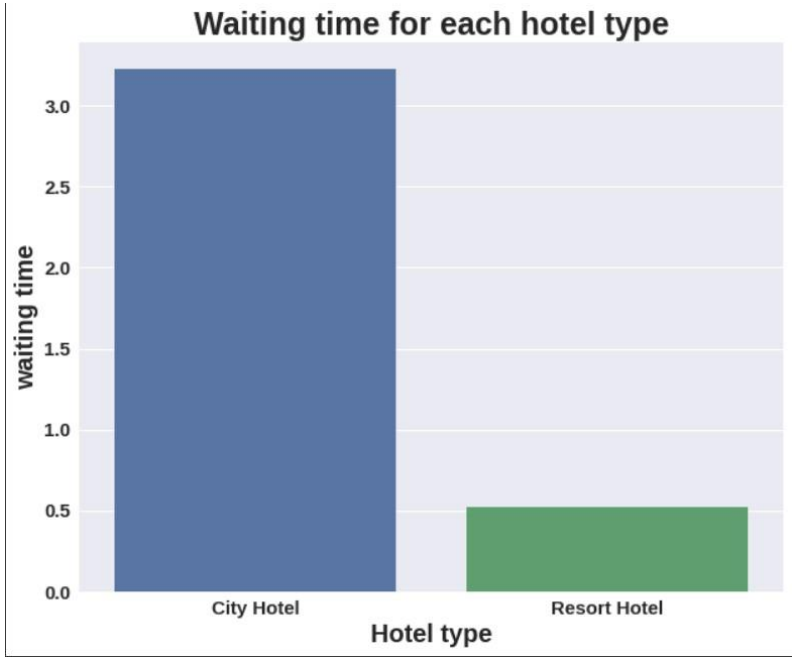


Conclusion :

- **The City Hotels has the high ADR as compared to Resort Hotels it means City Hotels are making more profit than the Resort Hotels.**
- **The average lead time is high for Resort Hotels as compared to City Hotels, it means people plan trips too early and preferred Resort Hotels for longer stay.**
- **The booking cancellation rate is high for City Hotels which is almost 30 %.**

❖ Exploratory Data Analysis (EDA) :

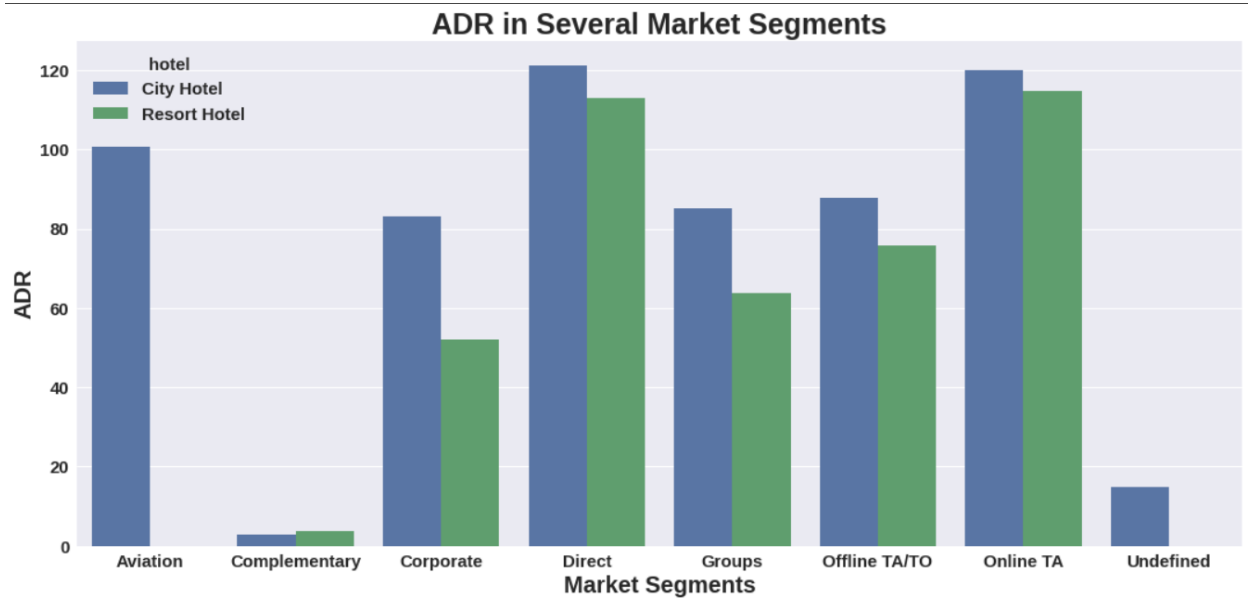


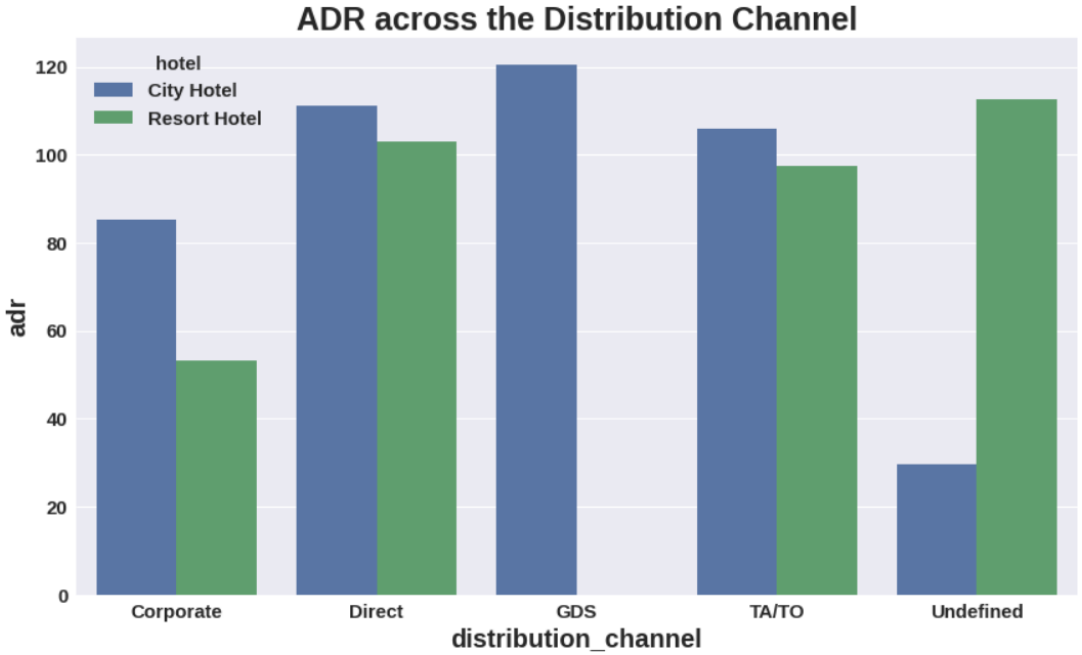


Conclusion :

- **The repeated customers are more in City Hotel as compare to Resort Hotel. As we see earlier the retention of repeated customers is very less so hotels need to give good services to increase retention of repeated customers.**
- **The City Hotel has more waiting time as compare to Resort Hotel so it means City Hotel is busiest than Resort Hotel.**

❖ Exploratory Data Analysis (EDA) :





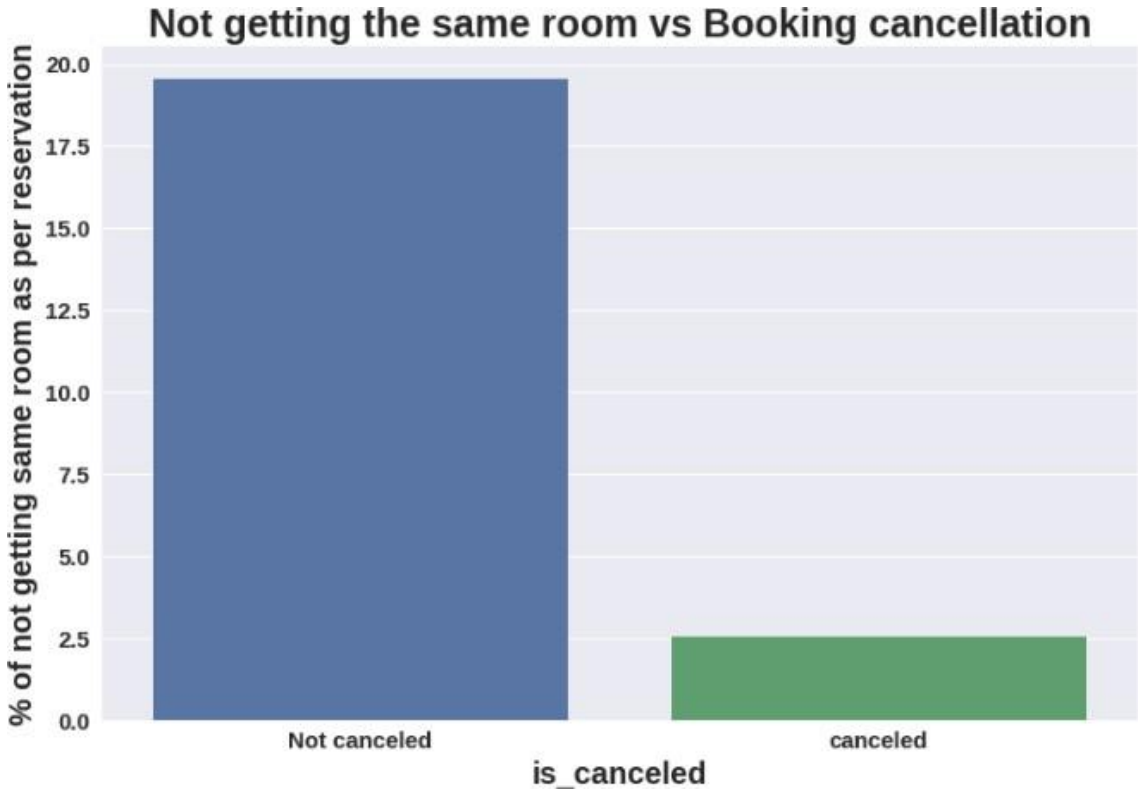
Conclusion :

- **The Highest ADR is in the ‘Direct’ And ‘Online Travel Agency’. It means ‘Direct’ And ‘Online Travel Agency’ is getting more revenue than others.**
- **The ‘GDS’ type has high ADR in City Hotel but need to increase ADR in Resort Hotel.**

Conclusion :

- **The Highest ADR is in the ‘Direct’ distribution type And ‘TA/TO’ Distribution type. It means ‘Direct’ And ‘TA/TO’ is getting more revenue than others.**
- **The ‘GDS’ type has high ADR in City Hotel but need to increase ADR in Resort Hotel.**

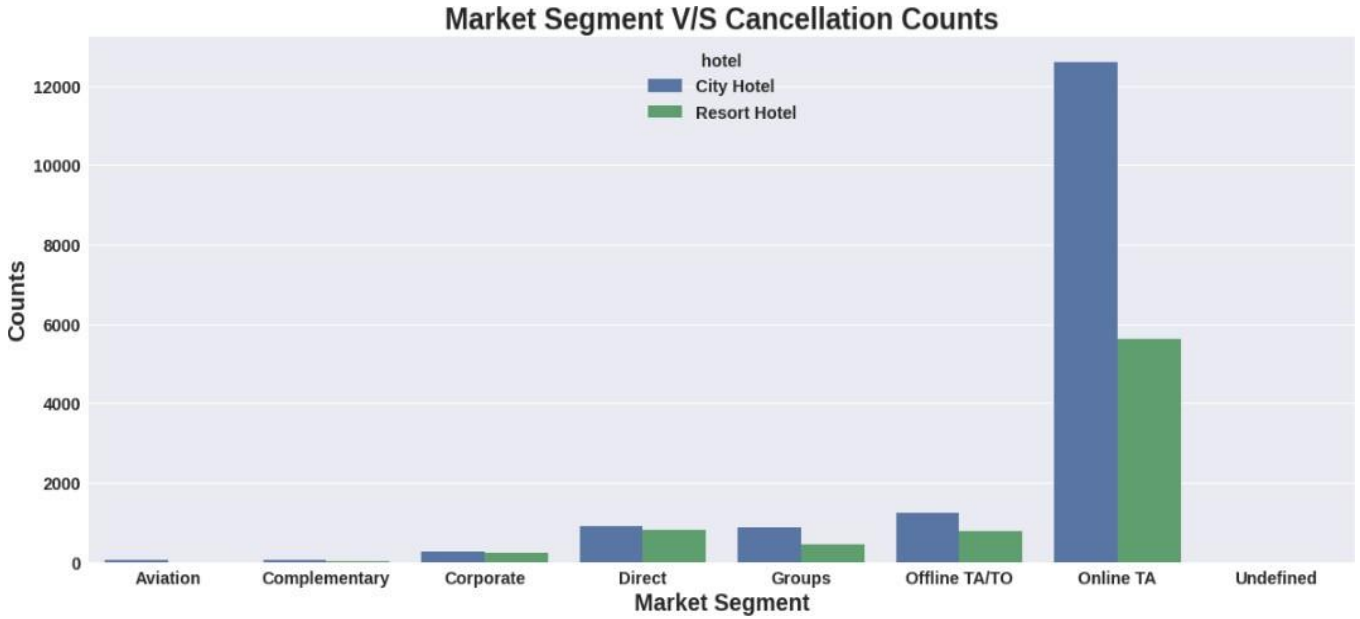
❖ Exploratory Data Analysis (EDA) :



Conclusion :

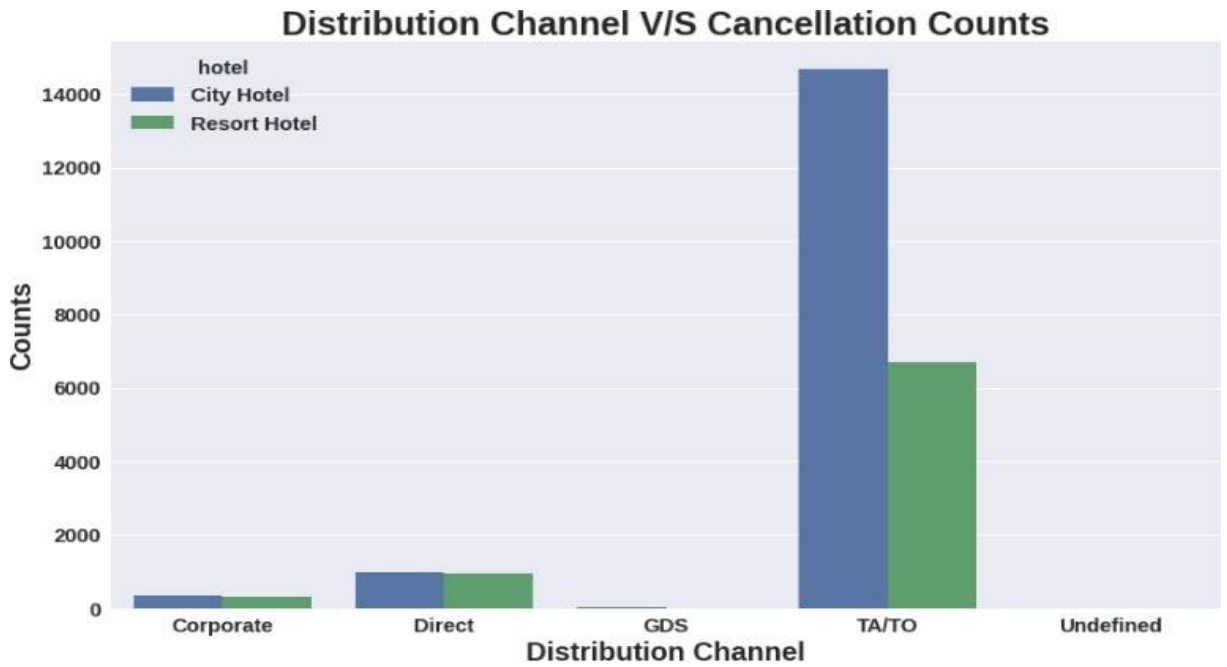
- **Only 2.5 % people cancelled bookings as they didn't getting same room which they reserve while booking else almost 19 % people didn't cancelled booking even after not getting same room which they reserve while booking.**
- **It means the not getting same room as per reserved room is not the reason for booking cancellations.**

❖ Exploratory Data Analysis (EDA) :



Conclusion :

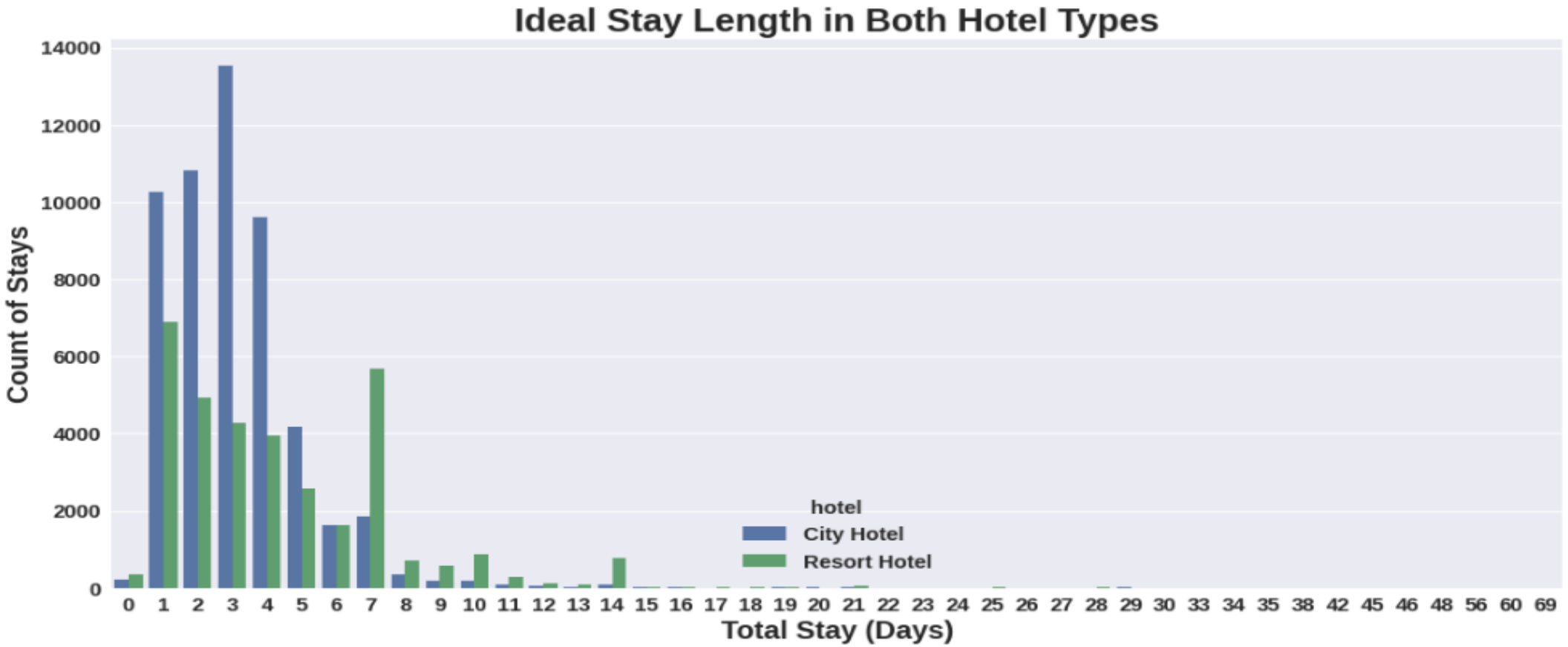
- **The highest cancellation rate is in 'Online TA' market segment for City Hotel as well as Resort Hotel.**
- **The 'Aviation' market segment has lowest cancellation count in both City Hotel and Resort Hotel.**



Conclusion :

- **The highest cancellation rate is in 'TA/TO' channel which is more than 14000 in City Hotel and more than 6000 in Resort Hotel.**

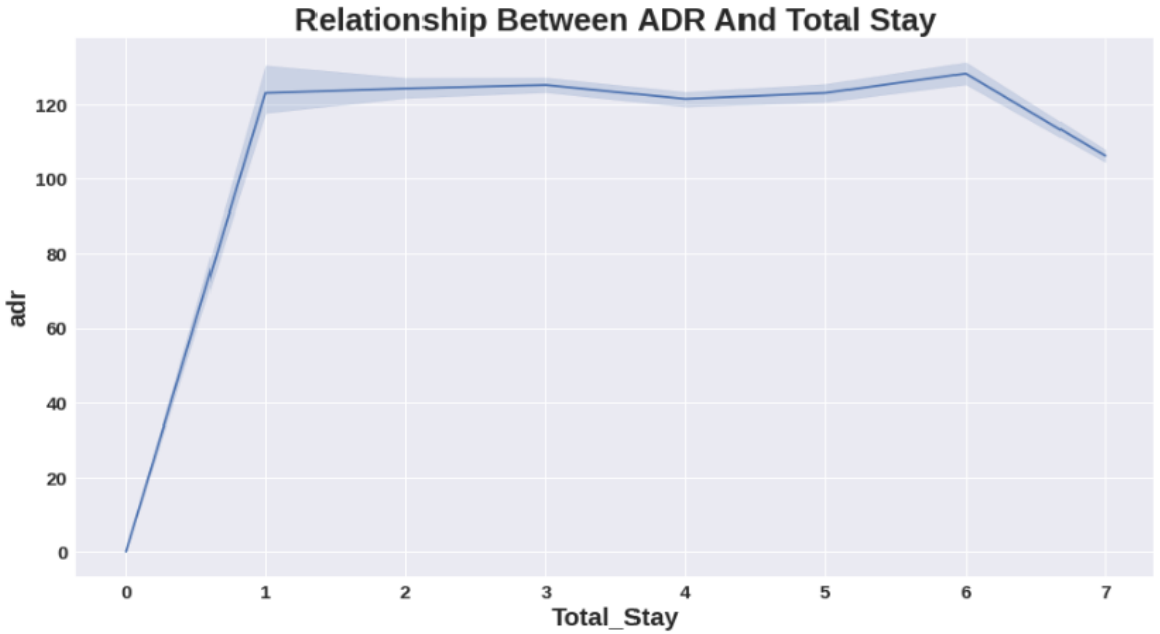
❖ Exploratory Data Analysis (EDA) :



Conclusion :

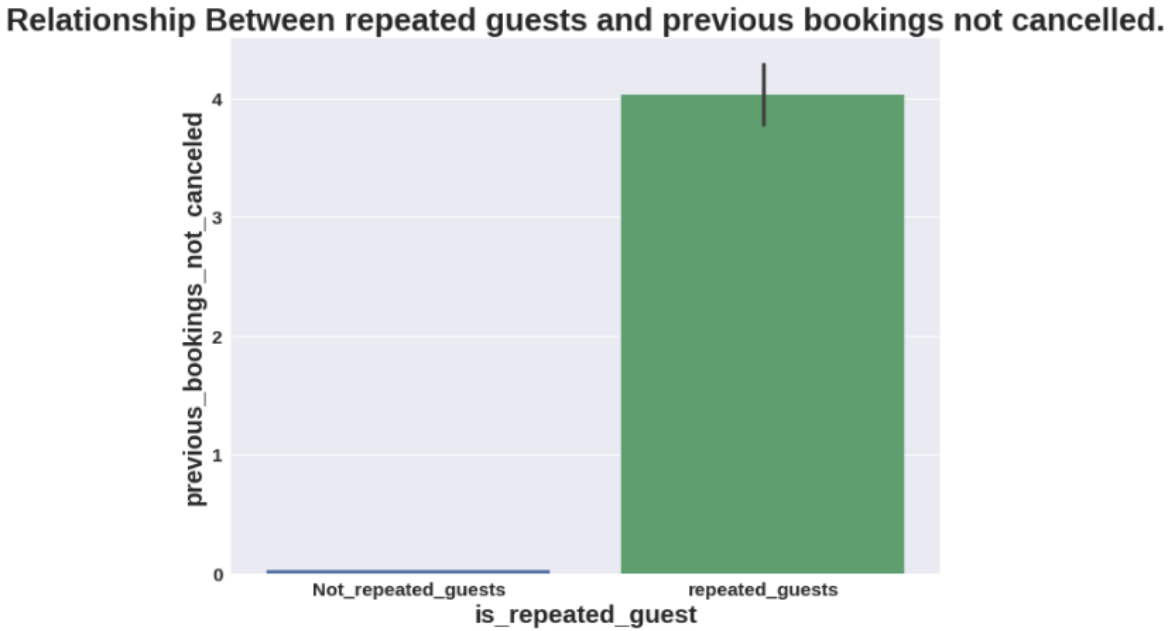
- As we can see above bar plot most of people stays less than 7 days in City Hotel as well as Resort Hotel.
- To stay more than 7 days most of people preferred Resort Hotel.

❖ Exploratory Data Analysis (EDA) :



Conclusion :

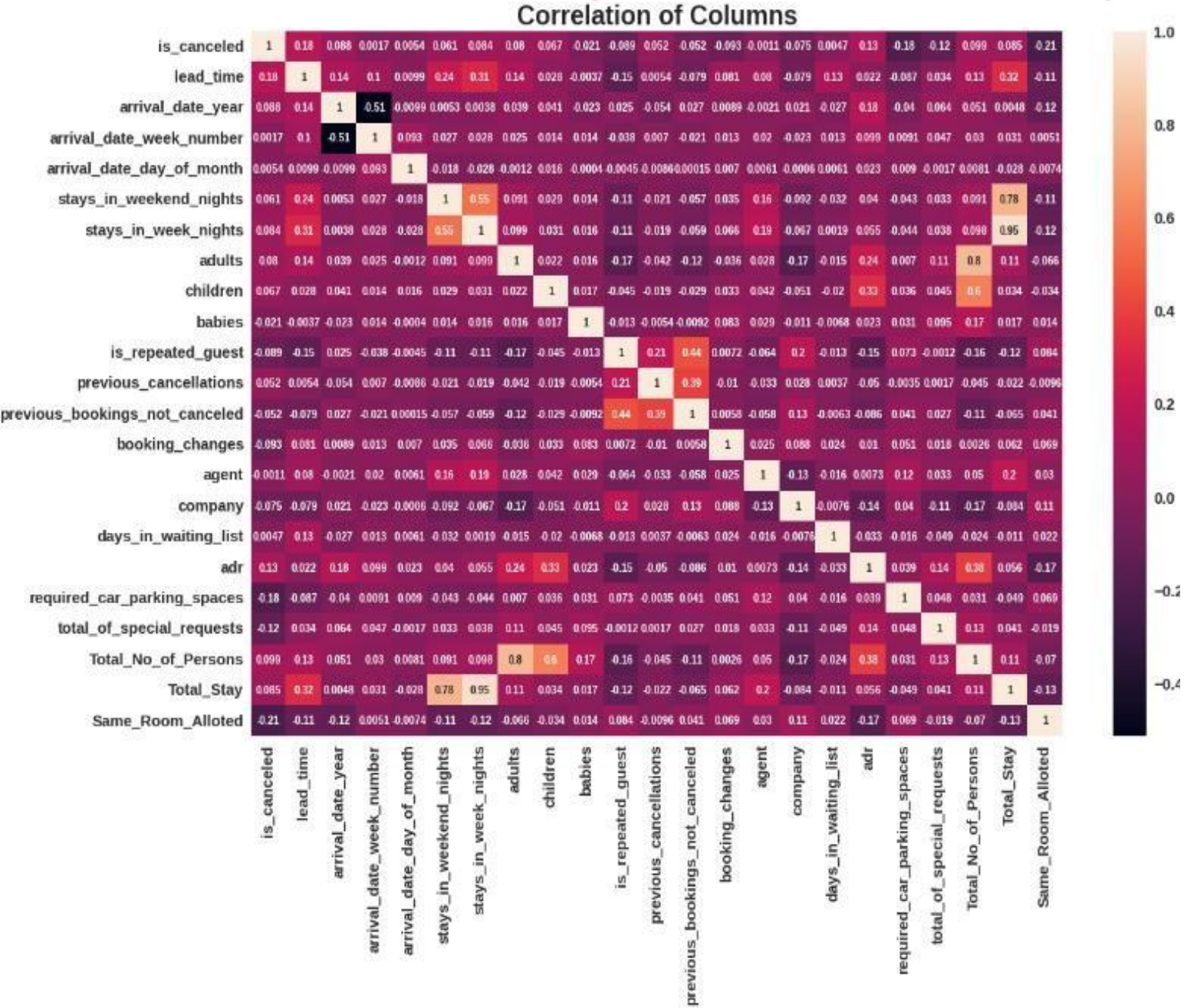
- As we can see line chart there is positive correlation in Total Stay and ADR, As the Total Stay increases the is also increasing.



Conclusion :

- As we can see bar plot most of repeated customers are from group of previous bookings not cancelled.
- It means the customers are satisfying with the service given by both Hotel types.

❖ Exploratory Data Analysis (EDA) :



Conclusion :

- **Lead Time And Total Stay is Positive correlated with each other it means more the total stay then more the lead time.**
- **Total No of persons And ADR are highly correlated means more the people more will be revenue.**
- **is_canceled And same_room_allotted_or_not are negatively correlated with each other it means same room not allotted as per reserved room is not the reason to cancellation.**

Thank You…