CAPSTONE PROJECT - 1

EDA On HOTEL BOOKING ANALYSIS

By

VIKASKUMAR SHARMA

(Cohort Tosh)

★ Problem Statement

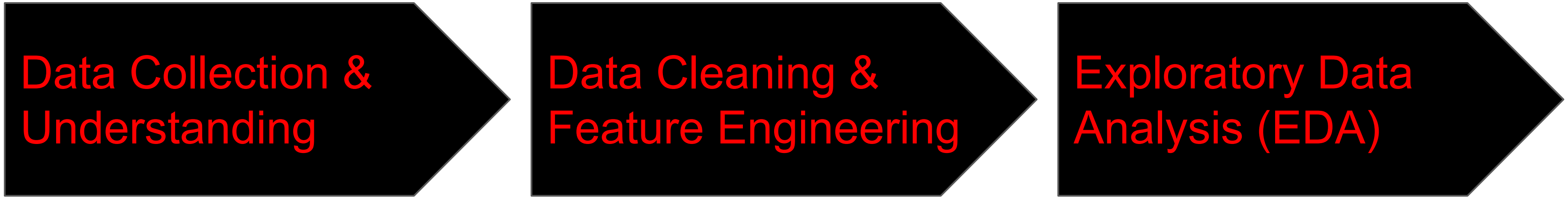
Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions!

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, among other things. All personally identifying information has been removed from the data.

We Will Explore and analyze the data to discover important factors that govern the bookings.

★ Work Flow

The Project is divided into 3 Main Steps:-



Data Collection & Understanding

Data Cleaning & Feature Engineering

Exploratory Data Analysis (EDA)

The EDA will be divided into 4 major Analysis:-

1)**Hotel Wise Analysis**

2)**Distribution Channel Wise Analysis**

3)**Booking Cancellation**

4)**Time Wise Analysis**

★ Data Description

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 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.

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.(T/A/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_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.

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.

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

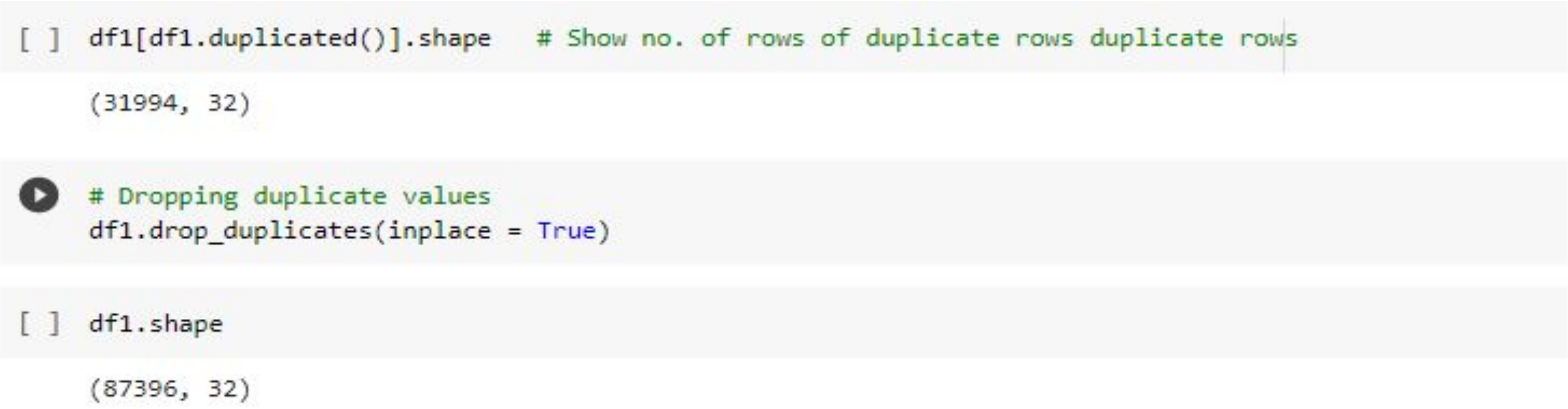
required_car_parking_spaces : Number of car parking spaces required by the customer t

total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)

reservation_status : Reservation last status.

❖ Data Cleaning & Manipulation

➢ Handling Duplicate Rows : Data had 31994 duplicate rows .So we dropped it.



```
[ ]  df1[df1.duplicated()].shape   # Show no. of rows of duplicate rows duplicate rows

     (31994, 32)
```

```
     # Dropping duplicate values
     df1.drop_duplicates(inplace = True)
```

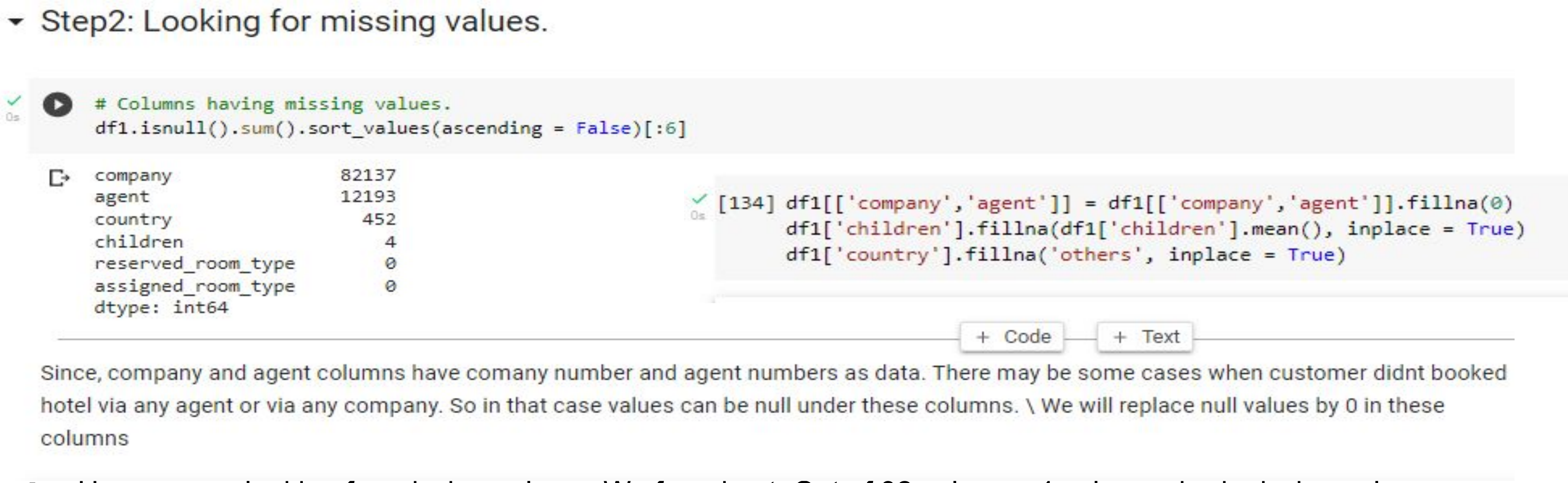
```
[ ]  df1.shape

     (87396, 32)
```

➔ The Dataframe Initially had 119390 rows and 32 columns.

➔ We checked for Duplicate rows and found there are 31994 rows duplicate .

➔ So, we just simply dropped the duplicate rows.

➔ Now, the dataframe is left with 87396 rows and 32 columns.



Step2: Looking for missing values.

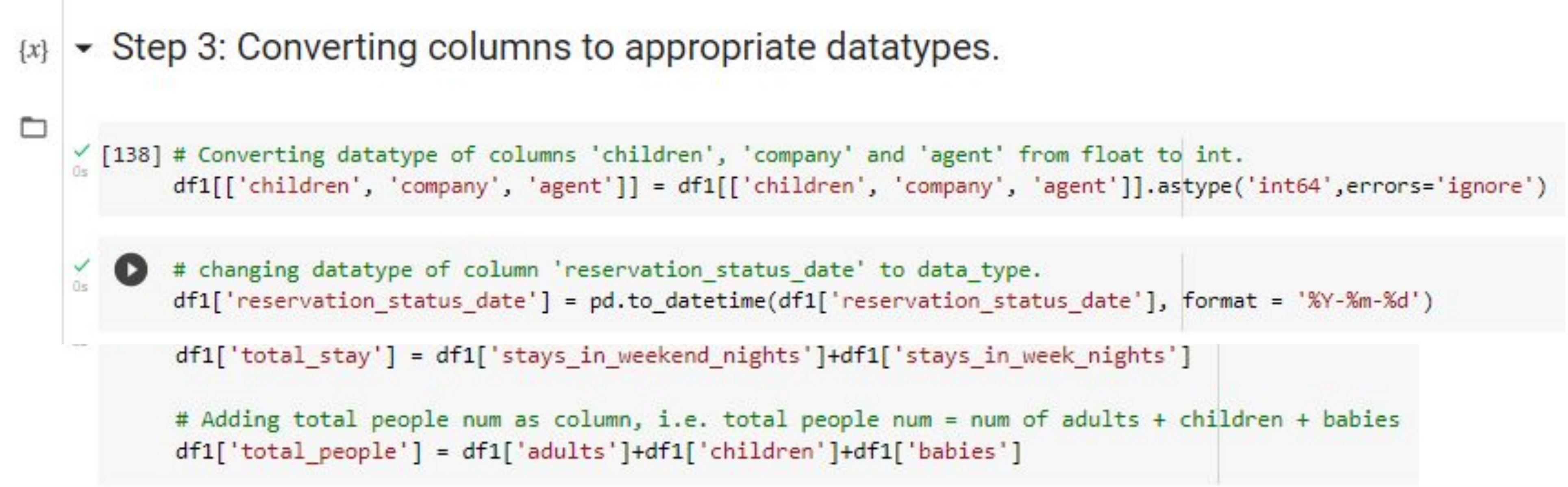
```
# Columns having missing values.
df1.isnull().sum().sort_values(ascending = False)[:6]
```

```
company               82137
agent                 12193
country                 452
children                  4
reserved_room_type        0
assigned_room_type        0
dtype: int64
```

```
[134] df1[['company','agent']] = df1[['company','agent']].fillna(0)
      df1['children'].fillna(df1['children'].mean(), inplace = True)
      df1['country'].fillna('others', inplace = True)
```

Since, company and agent columns have comany number and agent numbers as data. There may be some cases when customer didnt booked hotel via any agent or via any company. So in that case values can be null under these columns. \ We will replace null values by 0 in these columns

- ➔ Here we are looking for missing values . We found out ,Out of 32 columns 4 columns had missing values.
- ➔ Those columns are Company,Agent,Country & Children .
- ➔ We replaced the missing values in company & agent column by 0, whereas children by mean value and country by others.
- ➔ For country we will fill Missing values with 'Others'. (assuming while collecting data country was not found so user selected the 'Others' option.)
- ➔ AS the count of missing values in Children Column is only 4, so we can replace with 0 considering no childrens.



Step 3: Converting columns to appropriate datatypes.

```
[138] # Converting datatype of columns 'children', 'company' and 'agent' from float to int.
      df1[['children', 'company', 'agent']] = df1[['children', 'company', 'agent']].astype('int64',errors='ignore')
```

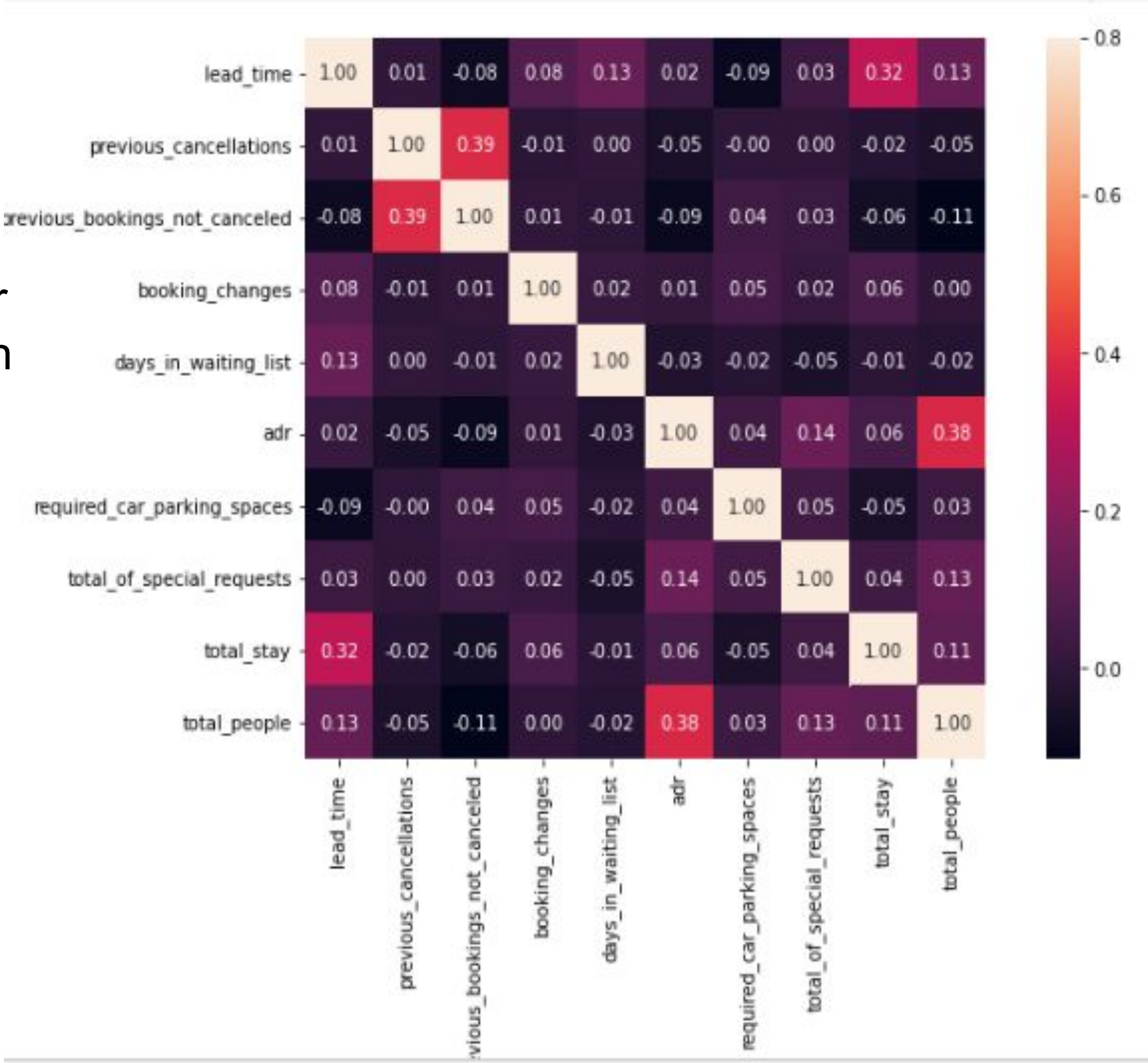
```
# changing datatype of column 'reservation_status_date' to data_type.
df1['reservation_status_date'] = pd.to_datetime(df1['reservation_status_date'], format = '%Y-%m-%d')
```

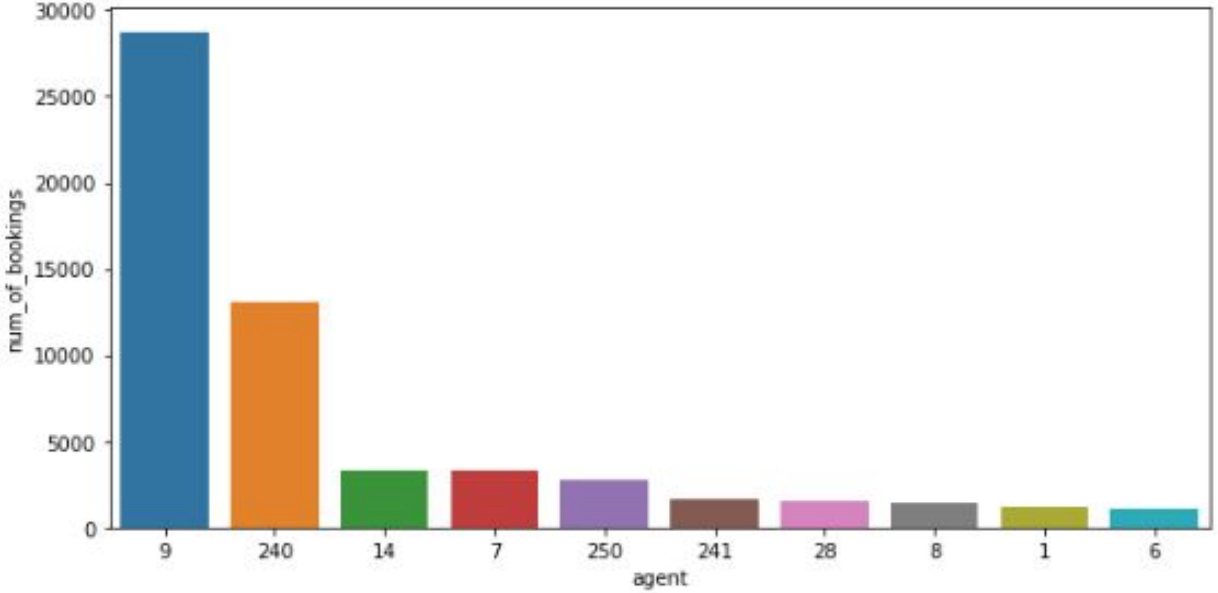
```
df1['total_stay'] = df1['stays_in_weekend_nights']+df1['stays_in_week_nights']

# Adding total people num as column, i.e. total people num = num of adults + children + babies
df1['total_people'] = df1['adults']+df1['children']+df1['babies']
```

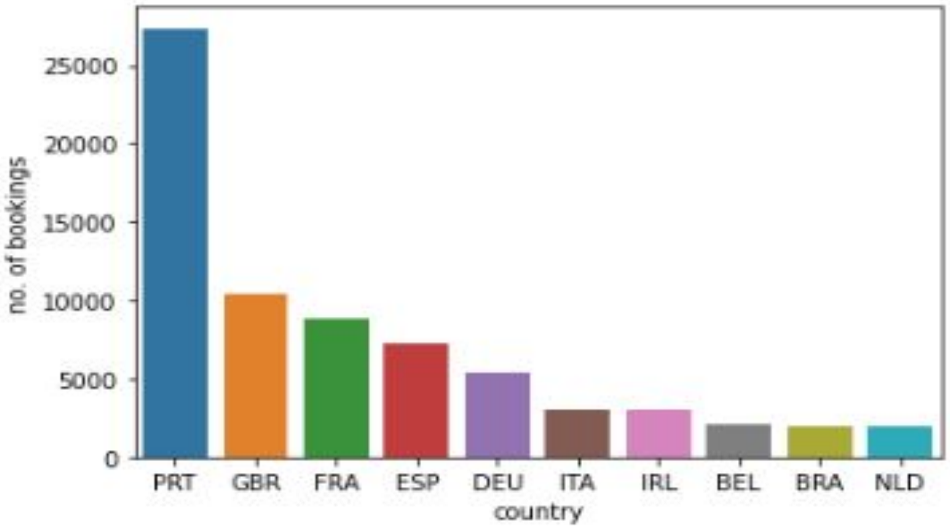
- ➔ Here we are converting the datatypes of feature children,company,agent from float64 to int64.
- ➔ Also,the feature reservation_status_date is converted from datatype object to datetime.
- ➔ After that we are also doing some feature engineering by creating new columns from existing columns.
- ➔ We did this by adding stays_in_weekend_nights and stays_in_week_nights into new column named total_stay.
- ➔ Also created new column called total_people by adding columns adults, children and babies.

EDA

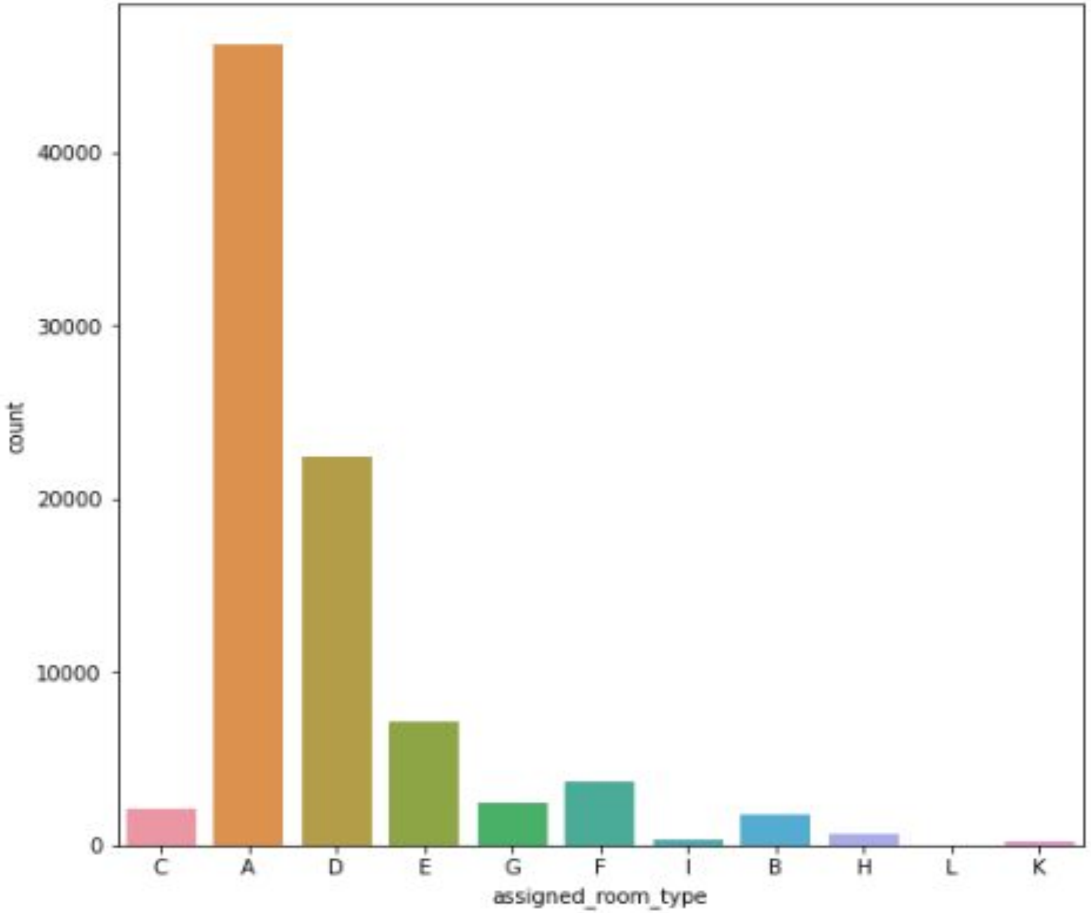
- ➔ This is the heatmap of the features.
- ➔ As we can see total_people and adr are positively correlated means as the total number of people increases the adr also increases.High adr means high revenue for the hotel.
- ➔ The column total_stay and lead_time is also positively correlated which simply means people who come for longer stays book their tickets way before early.
- ➔ The column previous _bookings_not_cancelled and previous_cancellations are also positively correlated..

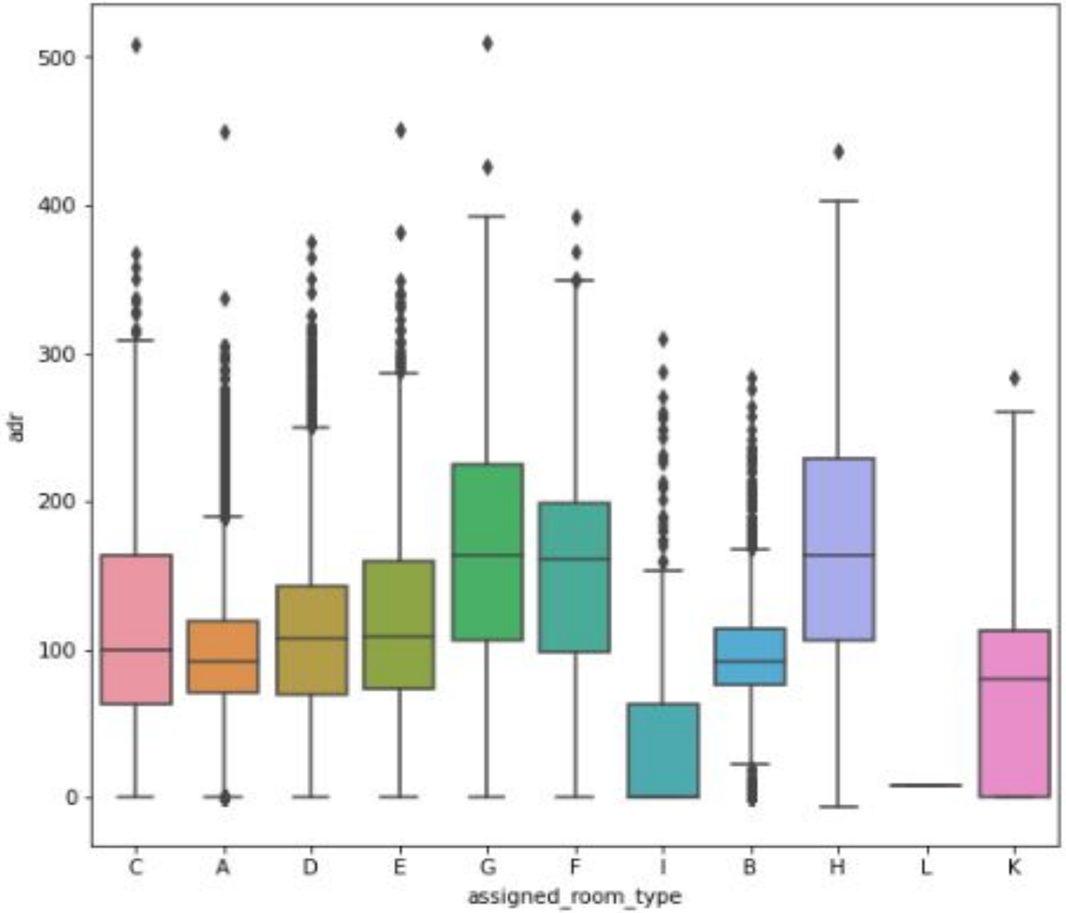




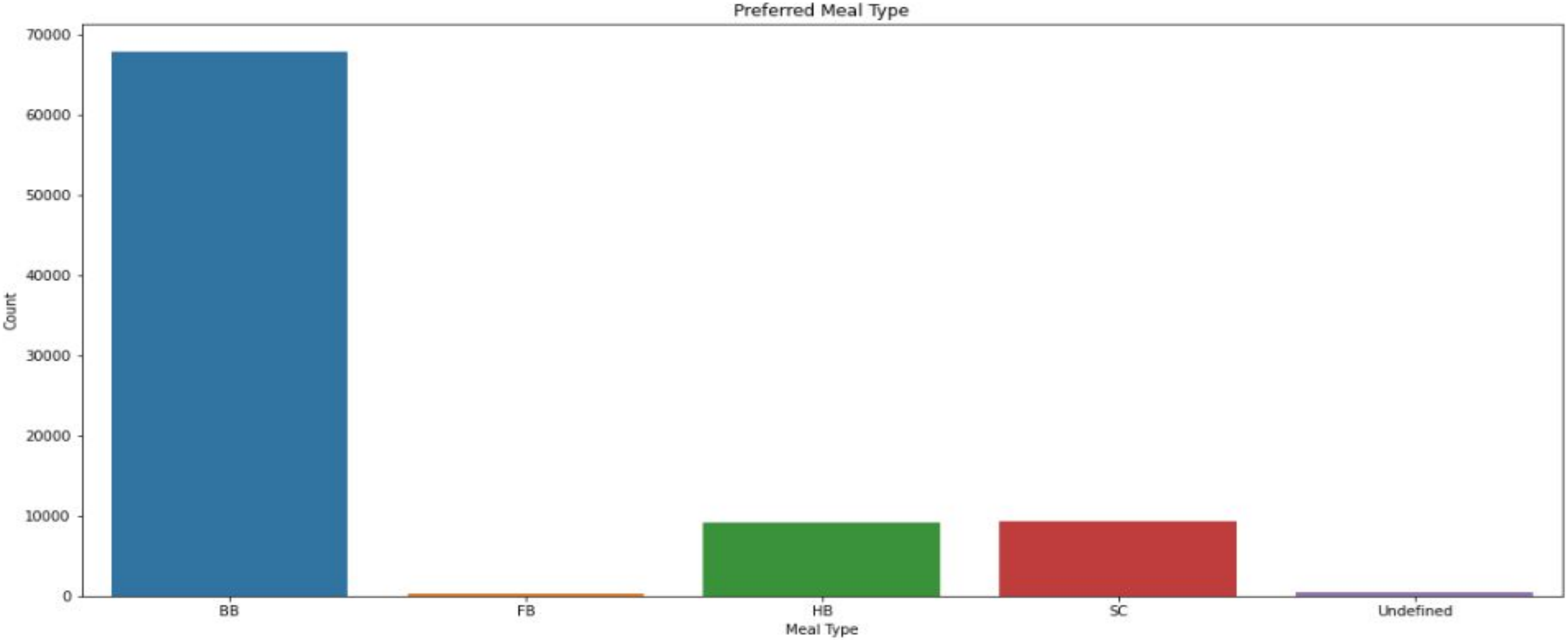


- ➔ The agent No 9,240,14,7,250 are the top 5 agents whom the customers had booked their tickets from.Maybe the Company can hire them for full time or give them some Incentives If they achieve certain target.
- ➔ The Top 10 Countries where the customers come from are Portugal, Great Britain, France, Spain, Germany, Italy ,Ireland,Belgium ,Brazil & Netherlands. Maybe the Company can do something so that people from these countries get visa-free access(if they at all are required to apply).



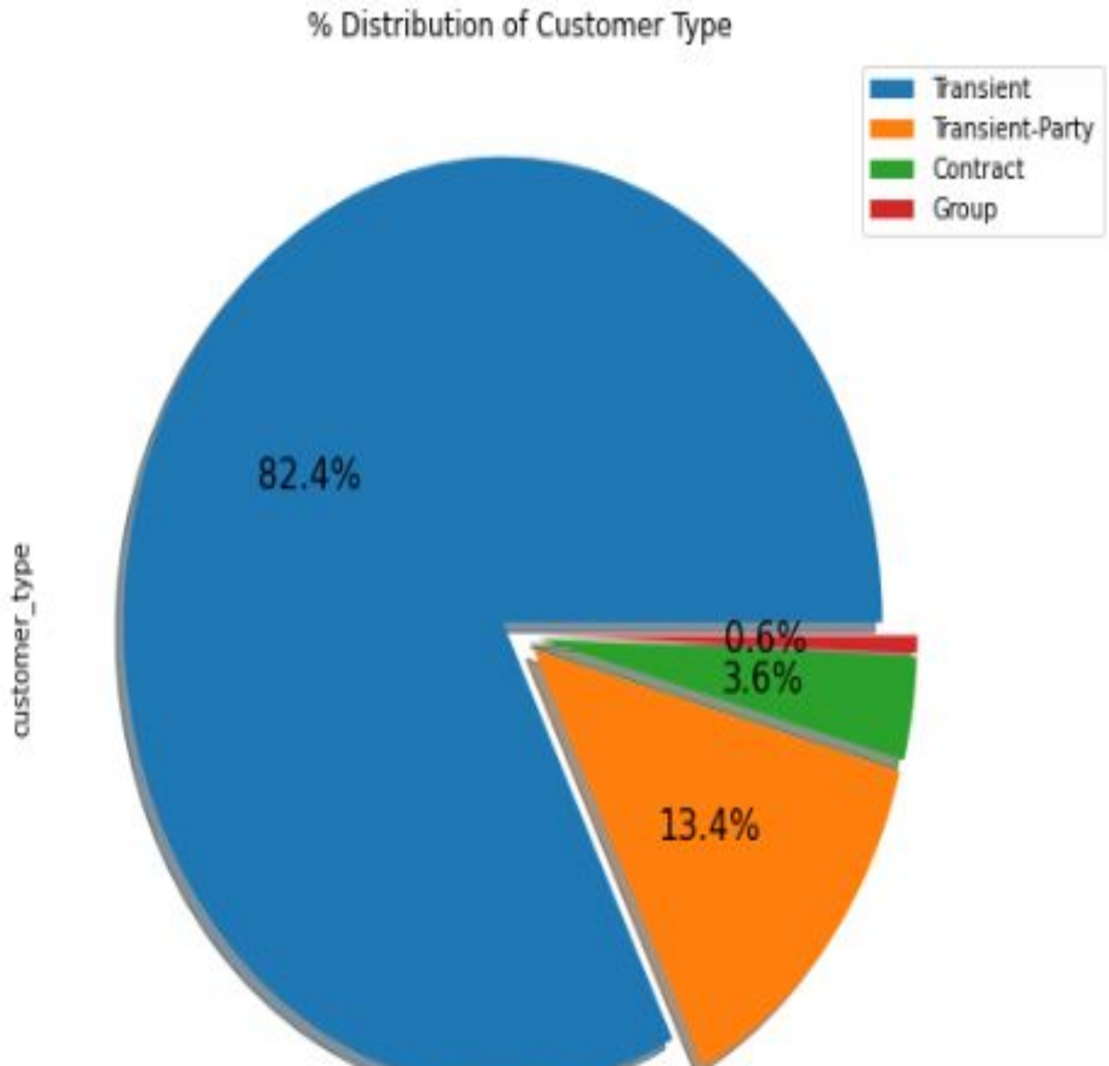


- ➔ The Room type A is the most preferred Room by the customers followed by D,E,F & G.
- ➔ Although the count of room type H is less, the highest average daily rate (adr) hotels receives is from room type H followed by G,F & C.

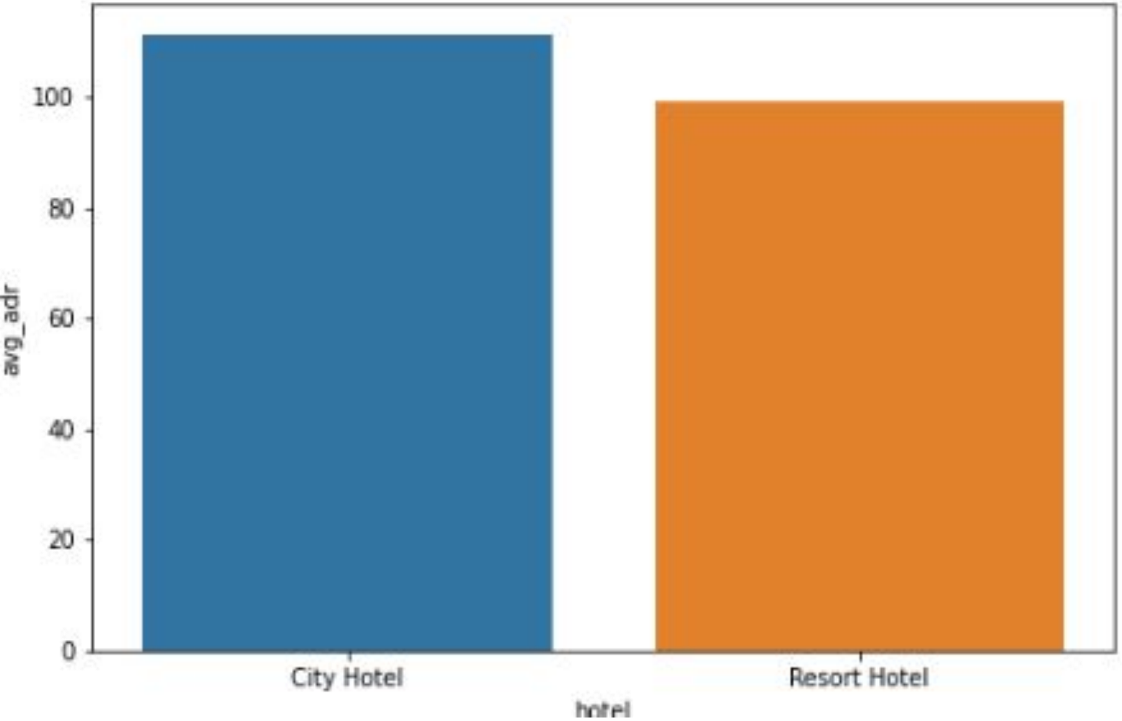


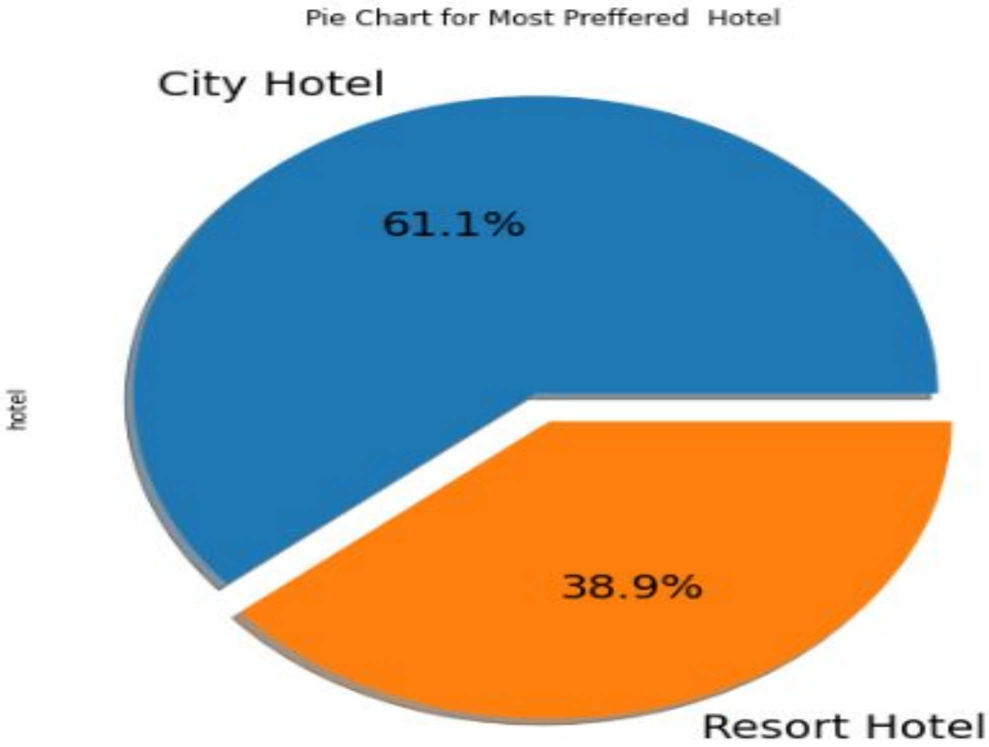
- ➔ The Most preferred meal type by the customers is BB (Bed & Breakfast) which is about 68000 i.e. roughly around (77 %) of the Customers preferred.
- ➔ The rest 27 % consist of Half Board (HB) ,Full Board (FB), Undefined (Undefined/SC).

- ➔ About 82.4 % of the customer comes from Transient type which means the booking is not part of group or the contract,and is not associated with other transient booking.
- ➔ It is followed by the Transient-Party Customer Type about (13.4%) which means when the booking is Transient, but is associated with at least other transient bookings.
- ➔ The remaining Bookings comes from the Contract about (3.6%) which means when the Booking has an allotment or other type of Contract associated with it.
- ➔ Only 0.6% of the Customers comes from the the group Customer type which means the booking is associated to a group.

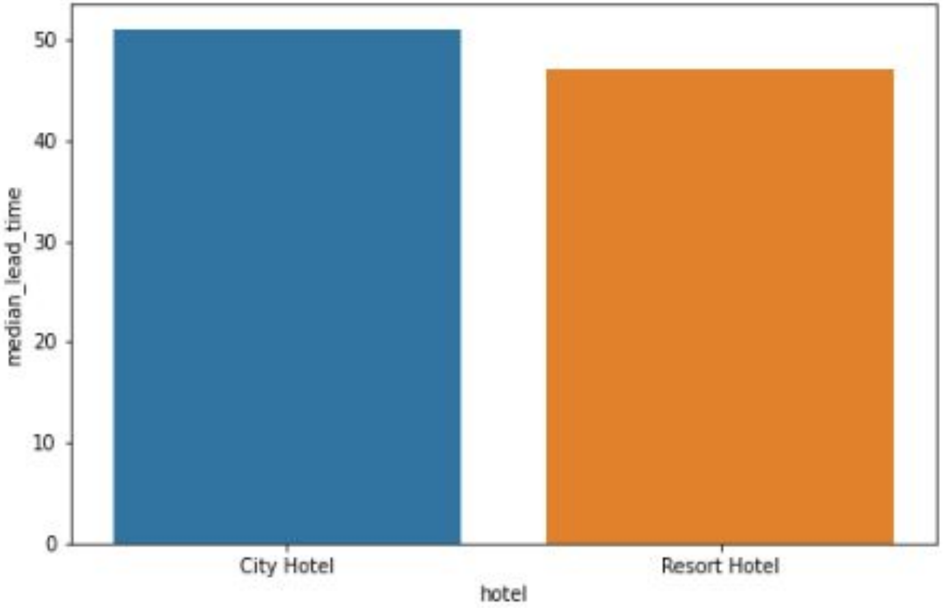


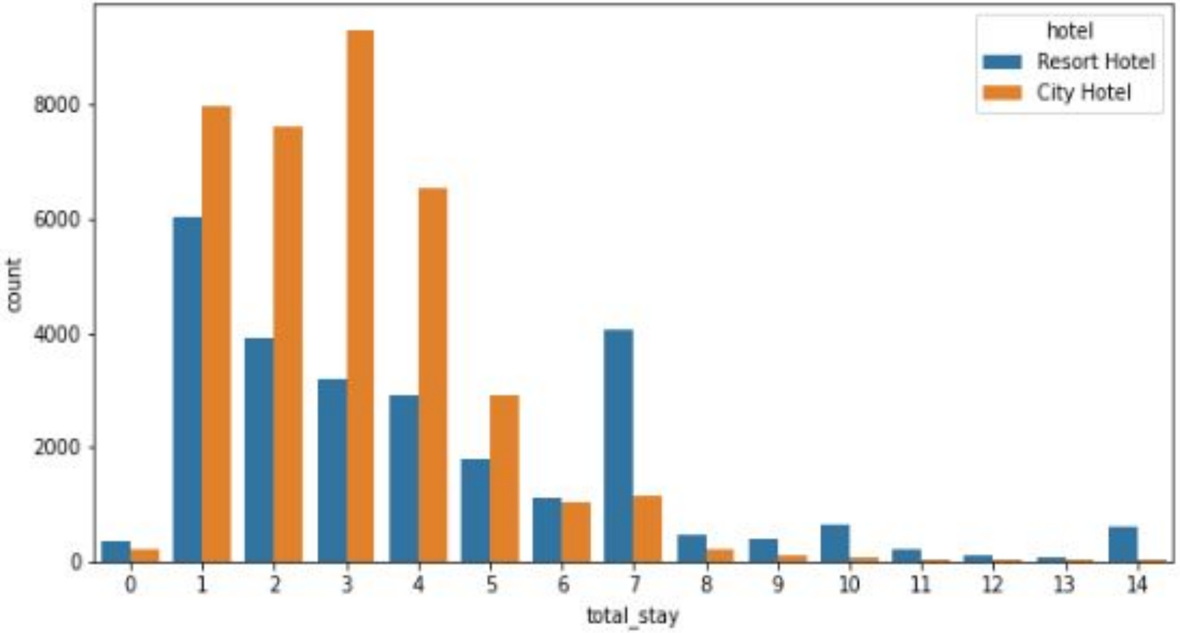
Hotel Wise Analysis





- ➔ As we can see from the pie chart about 60 % of the customers book for City Hotel and the Rest 40% for Resort Hotel.
- ➔ Although the no. of bookings for the city hotel is higher there is not much significant difference between the average adr of the two hotel type.

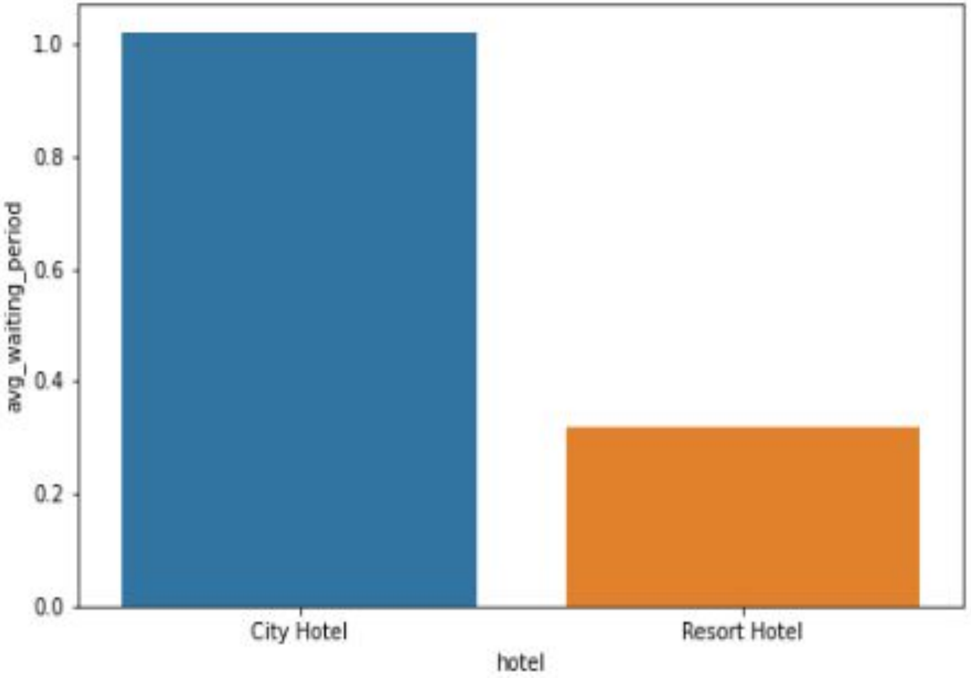


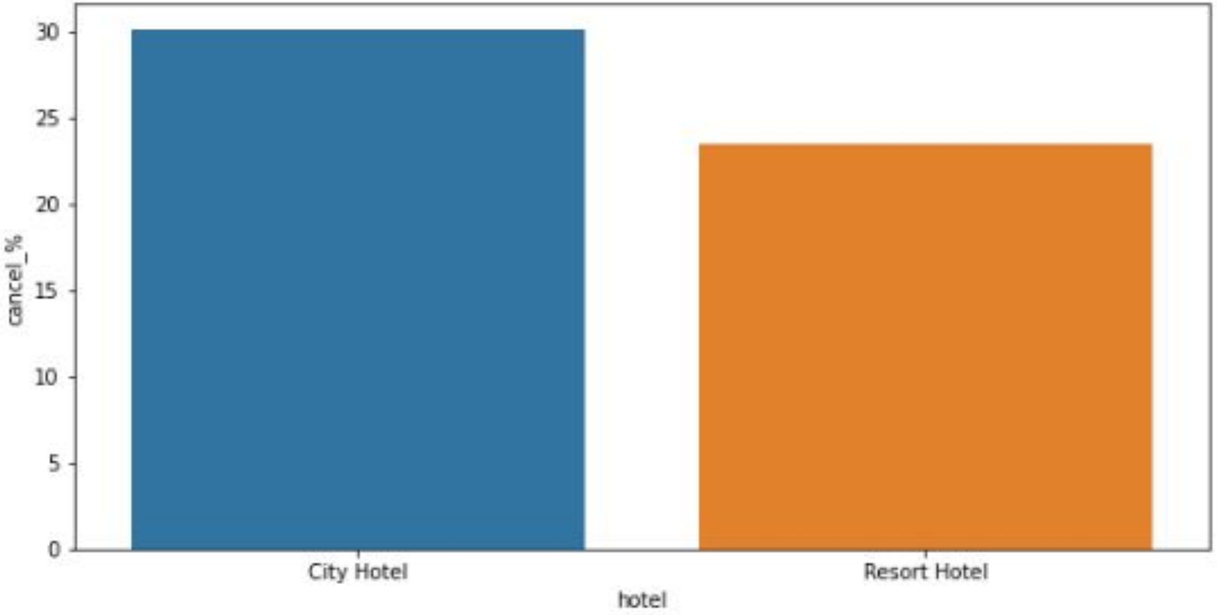


➔ As we can see from the above Bar chart there is not much difference between the lead time of City and Resort Hotel.

➔ We can notice from the above Bar chart , For the City Hotel people book for 1-6 days.

➔ Whereas In the case of Resort Hotel , People mostly book for 7 or more than 7 days, which is quite obvious because people usually come for vacation for longer stays.

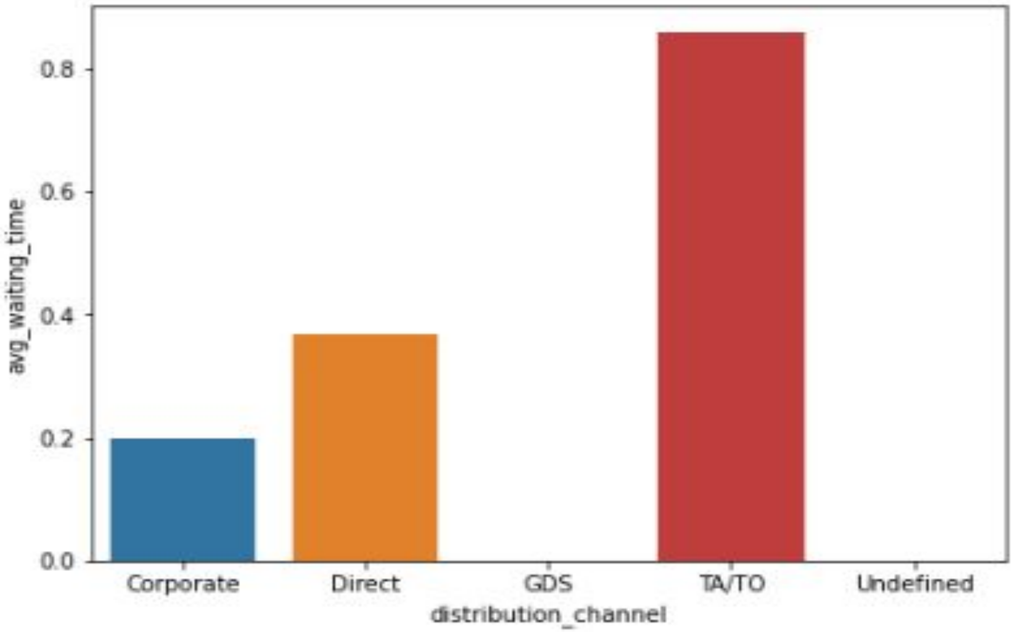


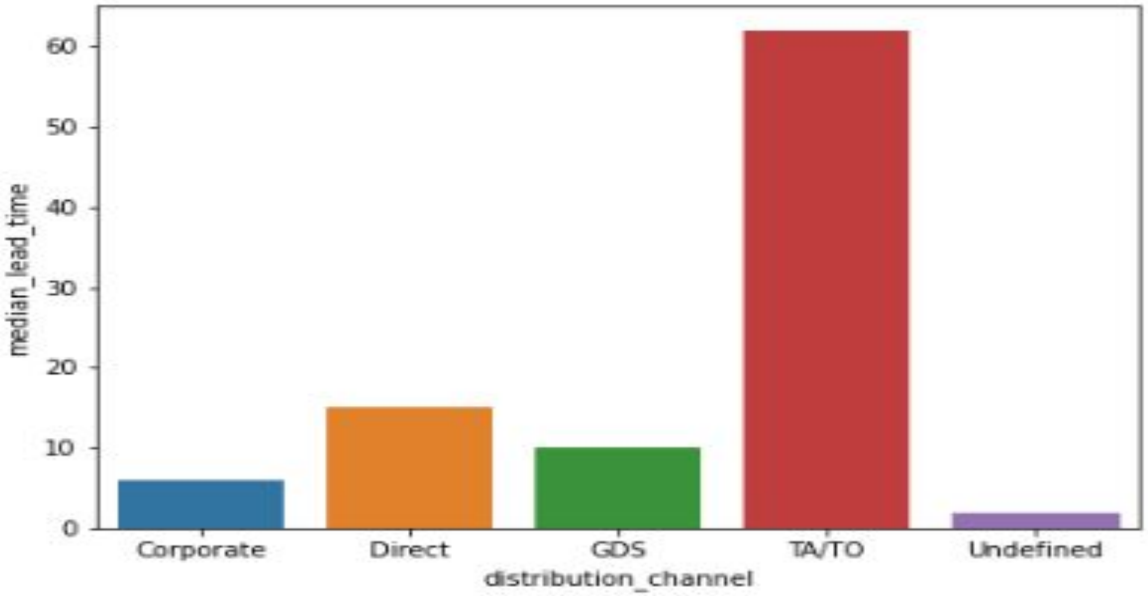


➔ The Avg waiting period in City Hotel is higher compared to Resort Hotel which can be understood as usually people for the city hotel did not book way prior as it is not in the case of the Resort Hotel.

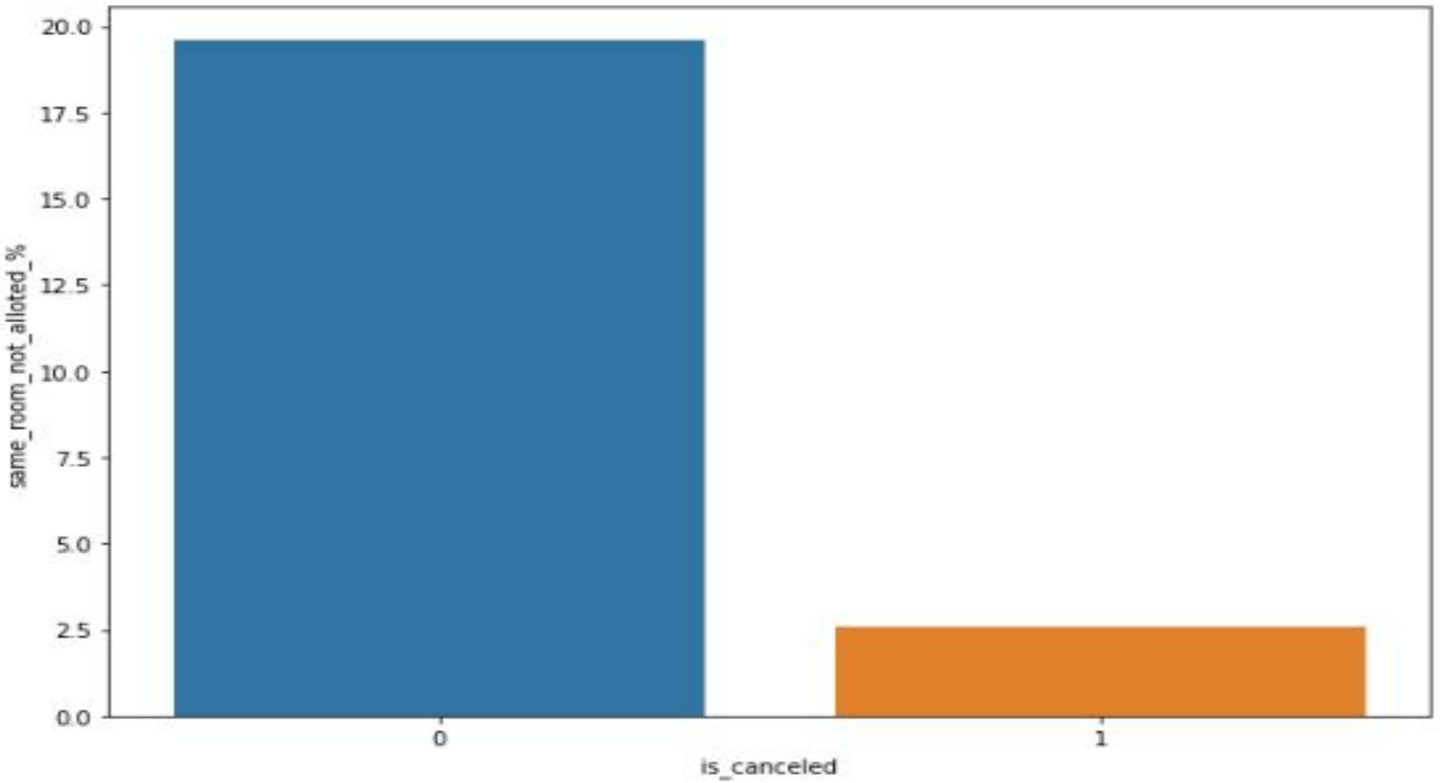
➔ As the no of the bookings is higher for City Hotel, The cancellation Percentage is also relatively higher(30%) compared to Resort Hotel (22%)

EDA Distribution Channel wise Analysis

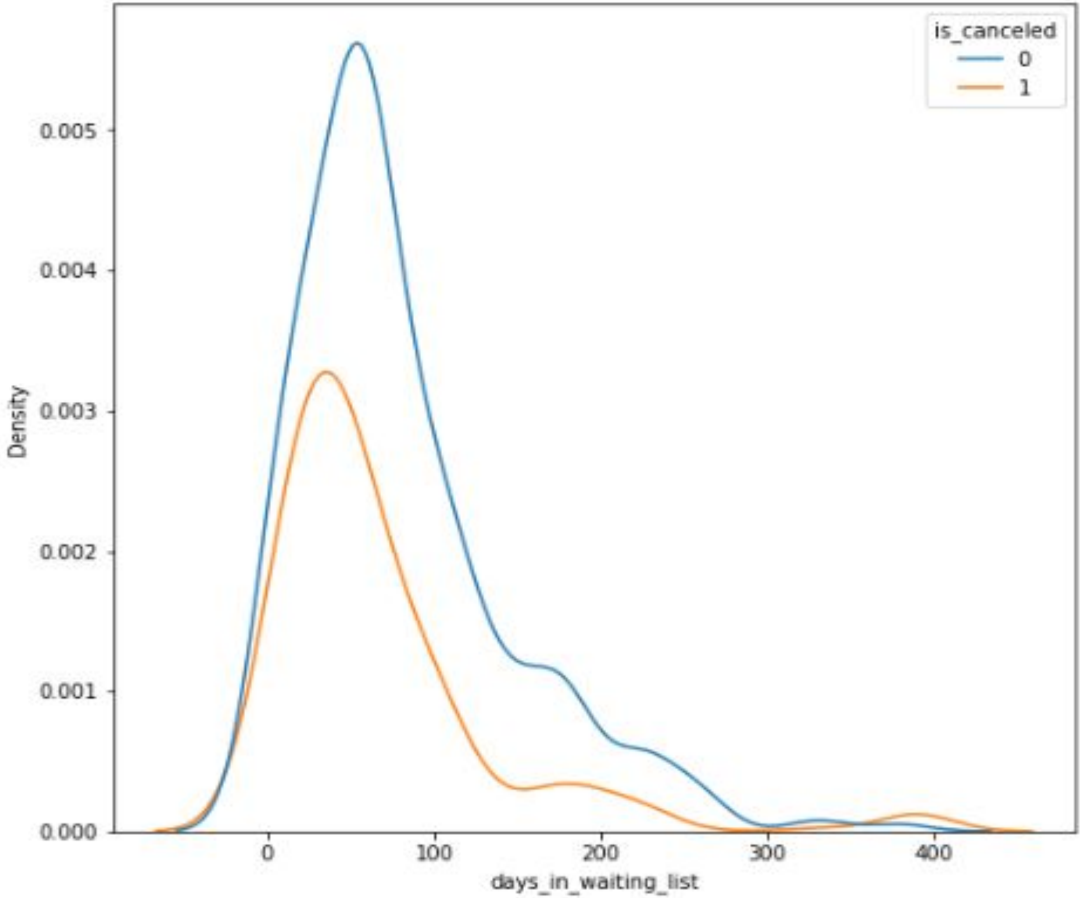


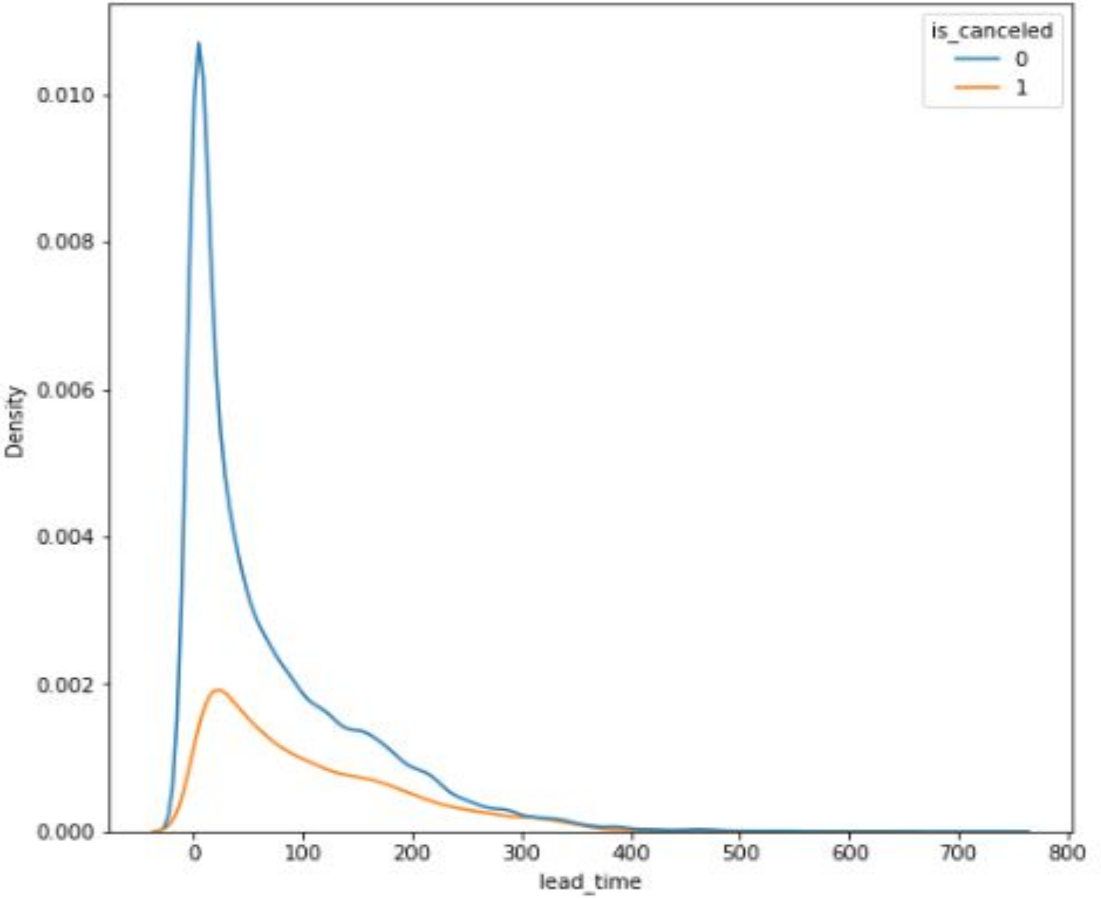


- ➔ As we can see the average waiting time for customers who come from TA/TO is highest followed by the Direct and Corporate Customers.
- ➔ Same is the case with the median lead time of TA/TO Customers followed by the rest of the Distribution Channel



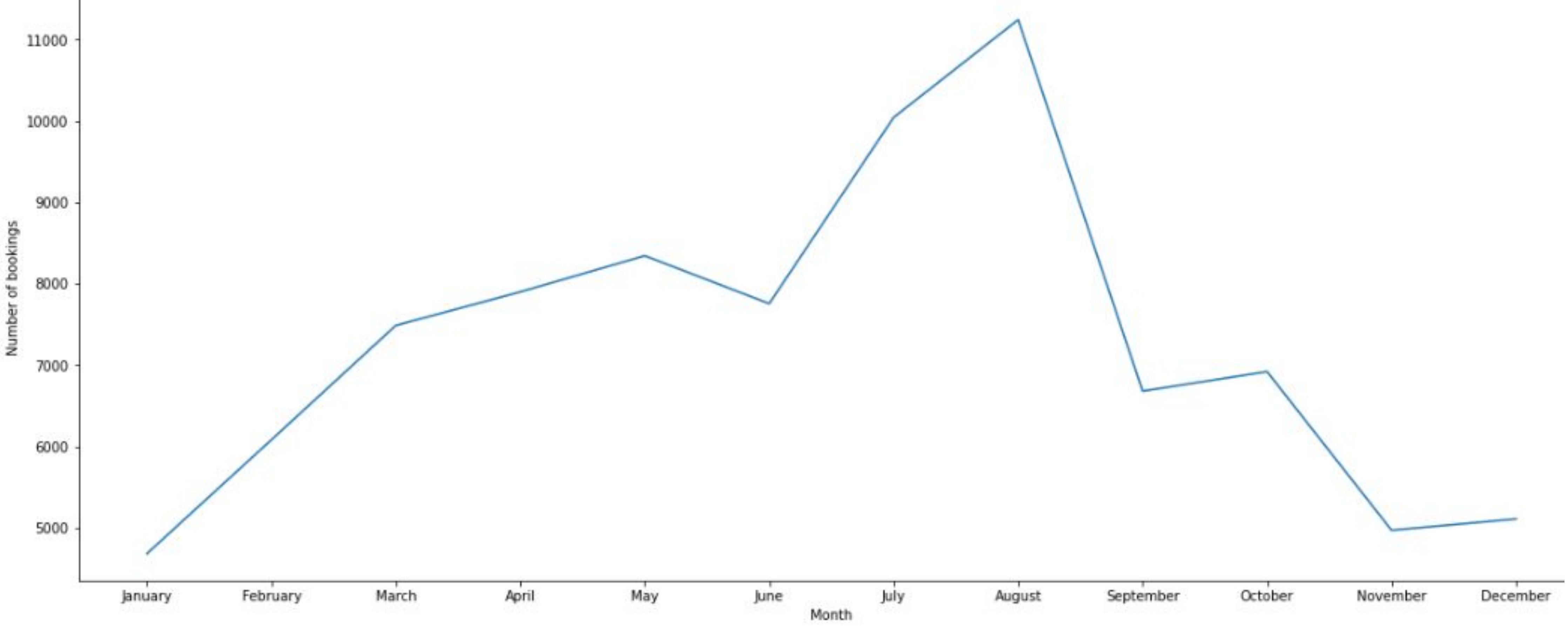
- ➔ As we can see from the above Bar Chart , that even though the same room is not alloted most of the people did not cancel there bookings.
- ➔ The Hotel Company can use this insight to fulfil the need of the customer where the customer books the room and is currently not available.





➔ We see that most of the bookings that are cancelled have waiting period of less 150 days but also most of bookings that are not cancelled also have waiting period less than 150 days. Hence this shows that waiting period has no effect on cancellation.

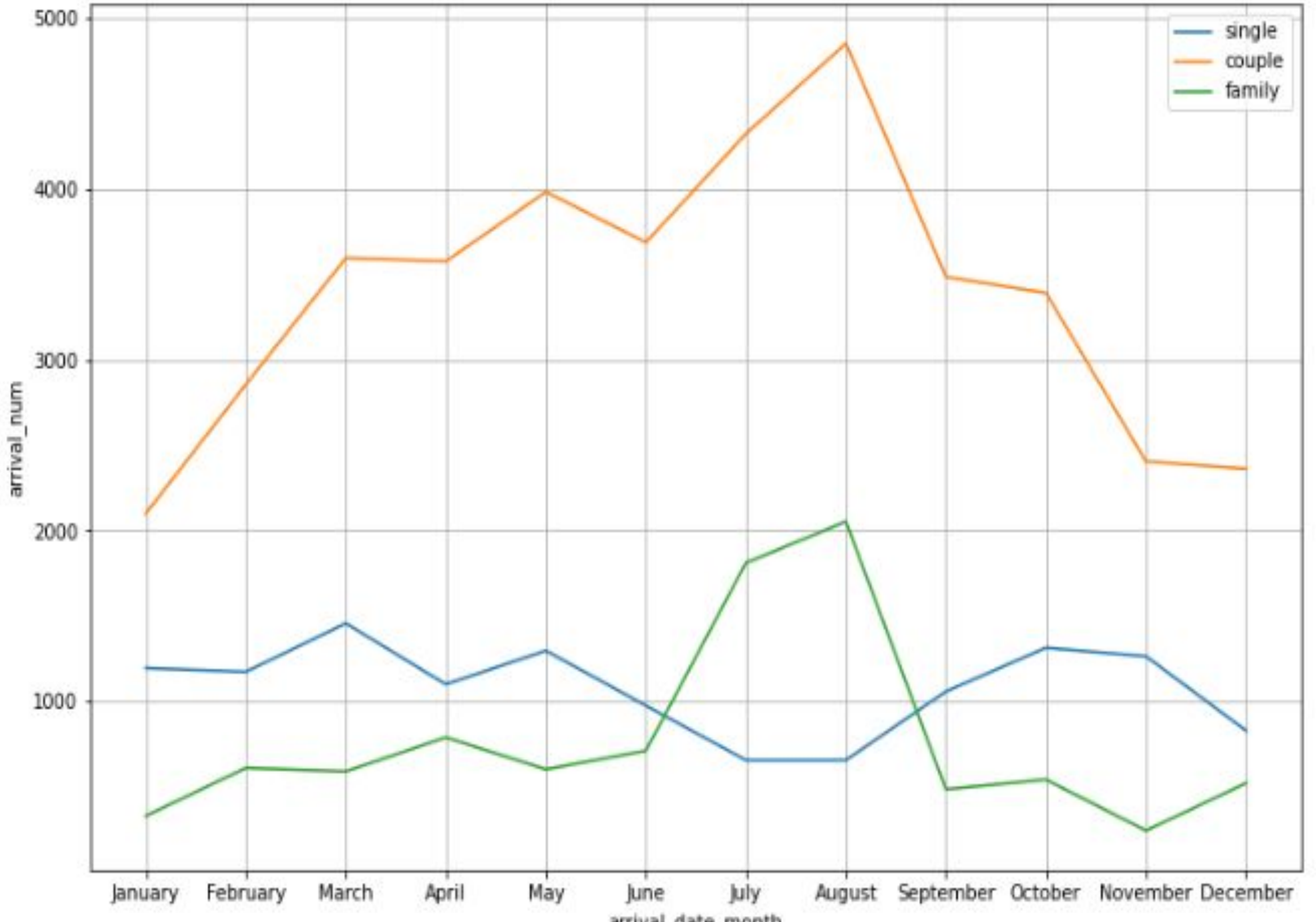
➔ Also, lead time has no effect on cancellation of bookings, as both curves of cancelation and not cancellation are similar for lead time too.



- ➔ As we can see in the line chart, from June to September most of the bookings happened. It’s Summer time. After September bookings Starts declining.
- ➔ There is not much bookings at the end of the year specifically November, December & January

➔ Mostly bookings are done by couples (although we are not sure that they are couple as data doesn't tell about that)

➔ It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.



Conclusions:-

- City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- 27.5 % bookings were got cancelled out of all the bookings.
- Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- Most of the customers (91.6%) do not require car parking spaces.
- 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Maximum number of guests were from Portugal, i.e. more than 25000 guests.
- Most of the bookings for City hotels and Resort hotel were happened in 2016.
- 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 almost 30 %.
- Average lead time for resort hotel is high.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Resort hotels have the most repeated guests.
- Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
- Almost 19 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.

THANK YOU