Exploratory Data Analysis of the Hotel Booking

Sweta Seal

Data science trainees,

AlmaBetter, Bangalore

ABSTRACT:

This datasets describes data with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel.Dataset share the same structure, with 32 variables describing the 40,060 observations of resort hotel and 79,330 observations of city hotel.The distance between these two locations is 280 km and both locations border on the north atlantic. Each observation represents a hotel booking. Dataset comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017.
We will perform exploratory data analysis with python to get insight from the data.

INTRODUCTION

Hotel industry is a very volatile industry and the bookings depend on a variety of factors such as type of hotels, seasonality, days of week and many more. This makes analysing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business.

ANSWERED QUESTIONS

Hotel wise Analysis

1. Which hotel is mostly booked by customers?
2. How long do people stay in hotels?
3. How does lead time affect cancellation of hotels ?

Booking Analysis

1. Which is the most common customer _type ?
2. Which country has the highest booking ?
3. What are the effects of deposit on bookings ?
4. What are the effects of deposit on bookings ?
5. Is assigned room type the cause for high cancellation?

Market Analysis

1. Which segment of the market usually has the least number on the waiting list?
2. Which segment of the Distribution channel usually has the least number on the waiting list?
3. Which hotel is mostly preferred by Distributors ?

Time Analysis

1. How does the average daily rate (adr) change with different months ?
2. What is the busiest month of the year?
3. Which is most preferred on weekdays or weekends ?
4. How does the average daily rate (adr) change with Customer Type ?
5. What are the chances of repeated guests cancelling the booking ?
6. Which segment of the distribution channel has the highest bookings and cancellations?
7. Which market segment has the highest bookings and cancellations?
8. Which type of rooms are mostly reserved ?
9. Which type of rooms are mostly assigned ?

DATA

This data article describes two datasets with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel. Both datasets share the same structure, with 32 variables describing the 40,060 observations of resort hotel and 79,330 observations of city hotel.The distance between these two locations is 280 km and both locations border on the north atlantic. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017.

Column Information:

Variable	Type	Description
ADR	Numeric	The average daily rate (ADR) measures the average rental revenue earned for an occupied room per day
Adults	Integer	Number of adults
Agent	Categorical	ID of the travel agency that made the bookings
Arrival Date Day Of Month	Integer	Day of the month of the arrival date
Arrival Date Month	Categorical	Month of arrival date with 12 categories: “January” to “December”
Arrival Date Week Number	Integer	Week number of the arrival date
Arrival Date Year	Integer	Year of arrival date
Assigned Room Type	Categorical	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
Babies	Integer	Number of babies
Booking Changes	Integer	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
Children	Integer	Number of children

Company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
Country	Categorical	Country of origin.
Customer Type	Categorical	Type of booking, assuming one of four categories:
		Contract - when the booking has an allotment or other type of contract associated to it;
		Group – when the booking is associated to a group;
		Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;
		Transient-party – when the booking is transient, but is associated to at least other transient booking
Days In Waiting List	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
Deposit Type	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
		No Deposit – no deposit was made;
		Non Refund – a deposit was made in the value of the total stay cost;
		Refundable – a deposit was made with a value under the total cost of stay.
Distribution Channel	Categorical	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Is Canceled	Categorical	Value indicating if the booking was canceled (1) or not (0)
Is Repeated Guest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
Lead Time	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
Market Segment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages:
		Undefined/SC – no meal package;
		BB – Bed & Breakfast;
		HB – Half board (breakfast and one other meal – usually dinner);
		FB – Full board (breakfast, lunch and dinner)

Previous Bookings Not Canceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking
Previous Cancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking
Required Car Parking Spaces	Integer	Number of car parking spaces required by the customer
Reservation Status	Categorical	Reservation last status, assuming one of three categories:
		Cancelled – booking was cancelled by the customer;
		Check-Out – customer has checked in but already departed;
		No-Show – customer did not check-in and did inform the hotel of the reason why
Reservation Status Date	Date	Date at which the last status was set. This variable can be used in conjunction with the reservation status to understand when was the booking cancelled or when did the customer checked-out of the hotel
Reserved Room Type	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
Stays In Weekend Nights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
Stays In Week Nights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
Total Of Special Requests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)

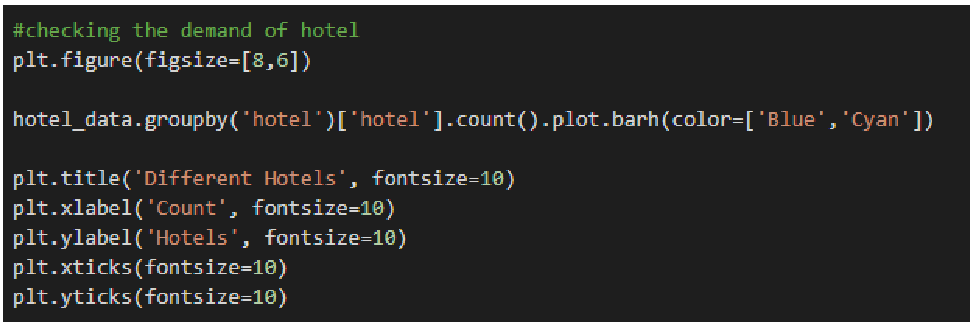
Steps involved

I. Importing libraries
II. Mounting the drive.
III. Reading the data set
IV. Displaying the data set
V. Checking the values of different columns
VI. Data cleaning and dealing with missing data
VII. EDA

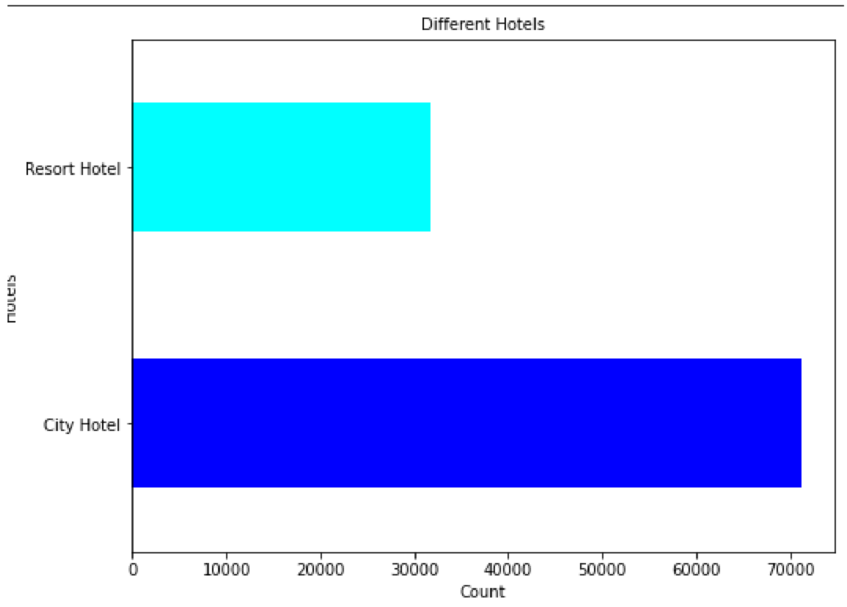
Hotel wise Analysis

1. Which hotel is mostly booked by customers?

For finding which hotel is most preferred we can group the hotel data and can plot the bar chart



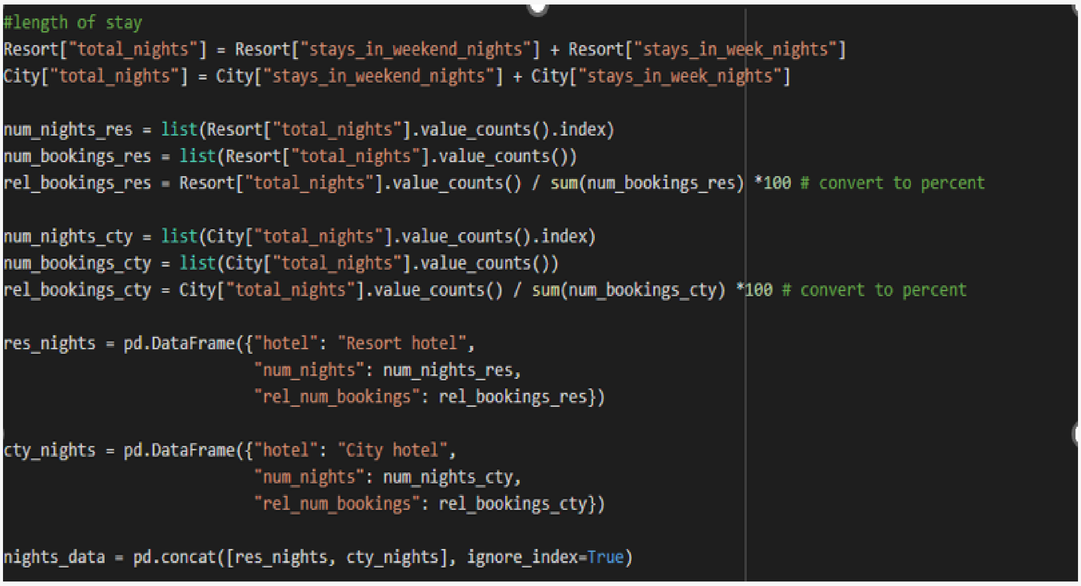
```
#checking the demand of hotel
plt.figure(figsize=[8,6])

hotel_data.groupby('hotel')['hotel'].count().plot.barh(color=['Blue','Cyan'])

plt.title('Different Hotels', fontsize=10)
plt.xlabel('Count', fontsize=10)
plt.ylabel('Hotels', fontsize=10)
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
```



Around 69.2% bookings are of City hotel and 30.8% bookings are for Resort hotel. By this we can say that a city hotel is most preferred hotel.

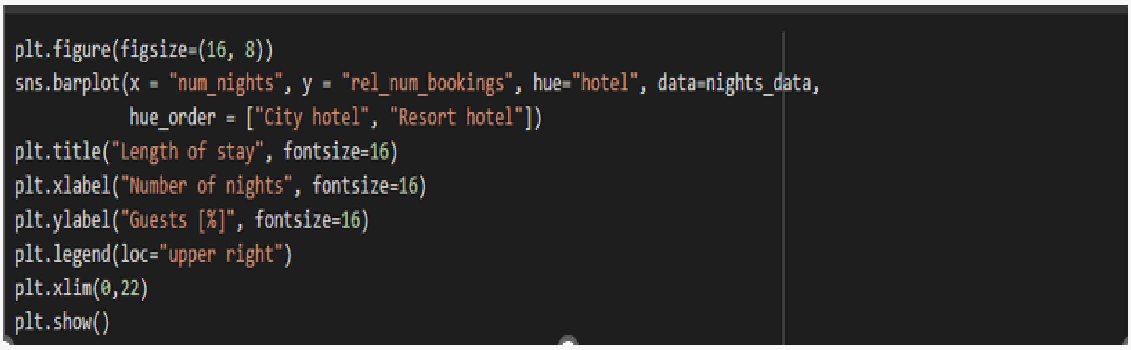
2. **How long does people stay in hotels?**

For finding the length of day first we need to create 2 data frames resort and city and counting each data frame values then converting it into percentage.

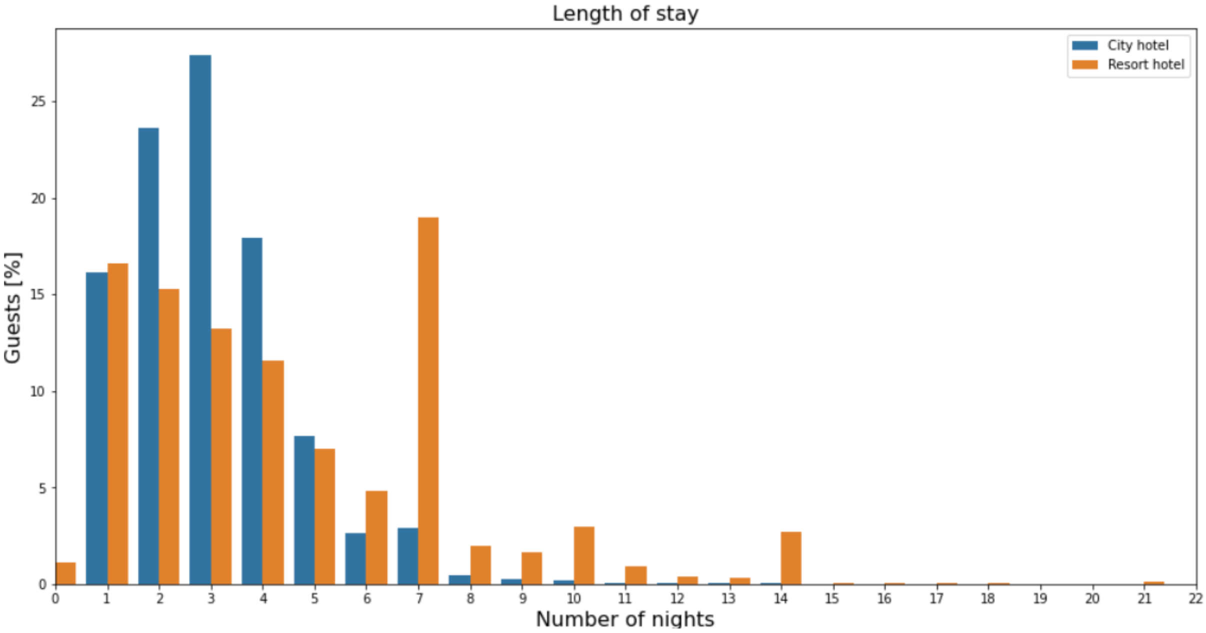


```
#length of stay
Resort["total_nights"] = Resort["stays_in_weekend_nights"] + Resort["stays_in_week_nights"]
City["total_nights"] = City["stays_in_weekend_nights"] + City["stays_in_week_nights"]

num_nights_res = list(Resort["total_nights"].value_counts().index)
num_bookings_res = list(Resort["total_nights"].value_counts())
rel_bookings_res = Resort["total_nights"].value_counts() / sum(num_bookings_res) *100 # convert to percent

num_nights_cty = list(City["total_nights"].value_counts().index)
num_bookings_cty = list(City["total_nights"].value_counts())
rel_bookings_cty = City["total_nights"].value_counts() / sum(num_bookings_cty) *100 # convert to percent

res_nights = pd.DataFrame({"hotel": "Resort hotel",
                           "num_nights": num_nights_res,
                           "rel_num_bookings": rel_bookings_res})

cty_nights = pd.DataFrame({"hotel": "City hotel",
                           "num_nights": num_nights_cty,
                           "rel_num_bookings": rel_bookings_cty})

nights_data = pd.concat([res_nights, cty_nights], ignore_index=True)
```

Then plotting the graph of length of day vs city and resort real bookings i.e excluding the bookings which are cancelled.



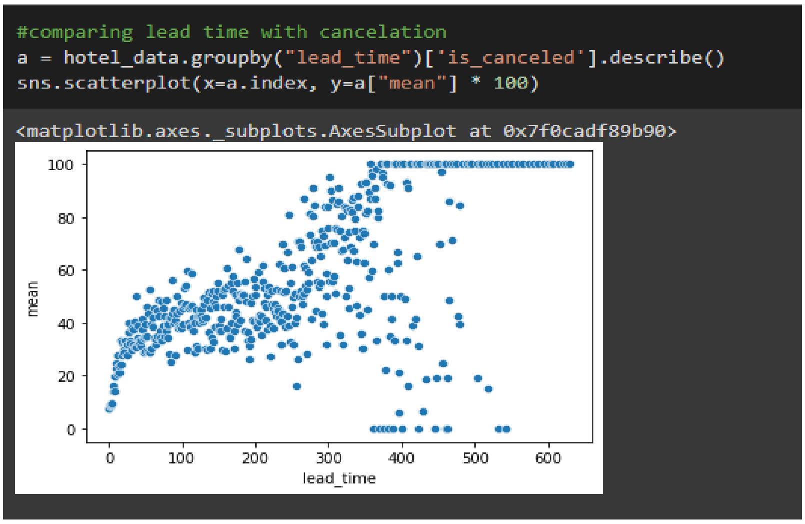
```
plt.figure(figsize=(16, 8))
sns.barplot(x = "num_nights", y = "rel_num_bookings", hue="hotel", data=nights_data,
            hue_order = ["City hotel", "Resort hotel"])
plt.title("Length of stay", fontsize=16)
plt.xlabel("Number of nights", fontsize=16)
plt.ylabel("Guests [%]", fontsize=16)
plt.legend(loc="upper right")
plt.xlim(0,22)
plt.show()
```



As we can see On average, guests of the City hotel stay 3 nights. On average, guests of the Resort hotel stay 4 nights For the city hotel there is a clear preference for 1-4 nights. For the resort hotel, 1-4 nights are also often booked, but 7 nights also stand out as being very popular.

3. How does lead time affect cancelation of hotel ?

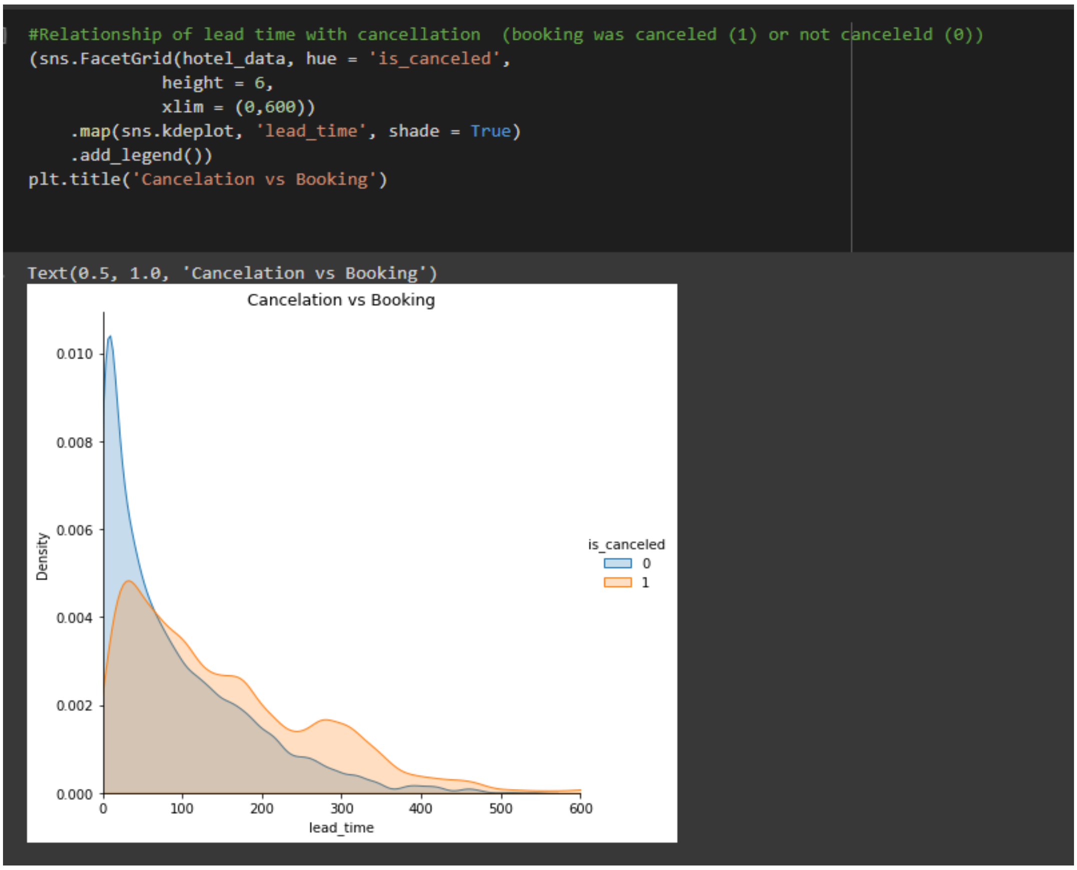
For finding the effects of lead time on cancellation we just need to group the variables i.e lead time and is cancelled and then plot the graph with seaborn library as sns.



```
#comparing lead time with cancelation
a = hotel_data.groupby("lead_time")['is_canceled'].describe()
sns.scatterplot(x=a.index, y=a["mean"] * 100)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0cadf89b90>
```

This graph is the same as above but done in seaborn kdeplot library which gives a graph between lead time and cancelled density**.**



```
#Relationship of lead time with cancellation  (booking was canceled (1) or not canceleld (0))
(sns.FacetGrid(hotel_data, hue = 'is_canceled',
             height = 6,
             xlim = (0,600))
    .map(sns.kdeplot, 'lead_time', shade = True)
    .add_legend())
plt.title('Cancelation vs Booking')
```

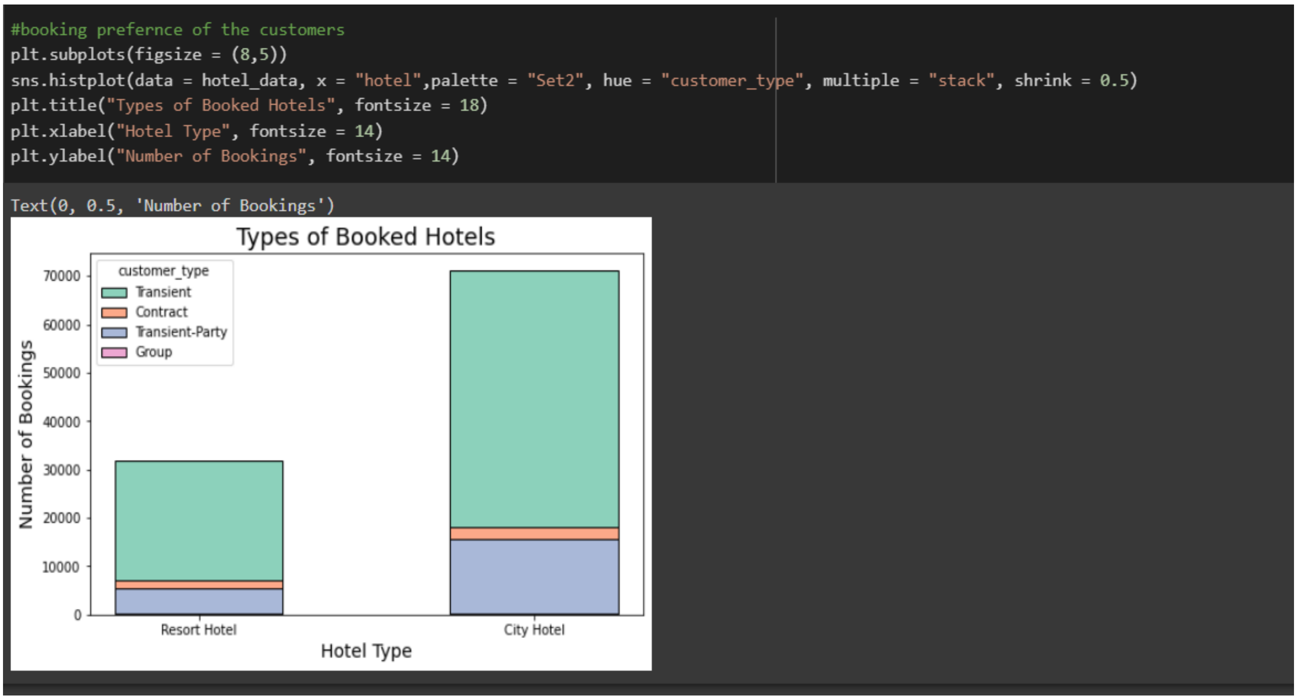
```
Text(0.5, 1.0, 'Cancelation vs Booking')
```

As we can see Lead time has a positive correlation with cancellation. i.e as lead time increases there is a high chance of cancelling the booking. Bookings made a few days before the arrival date are rarely cancelled, whereas bookings made over one year in advance are cancelled very often.

Booking Analysis

4. Which is the most common customer _type ?

For answering this question we need to just give x variable as hotel and multiple as stack and hue as customer type.



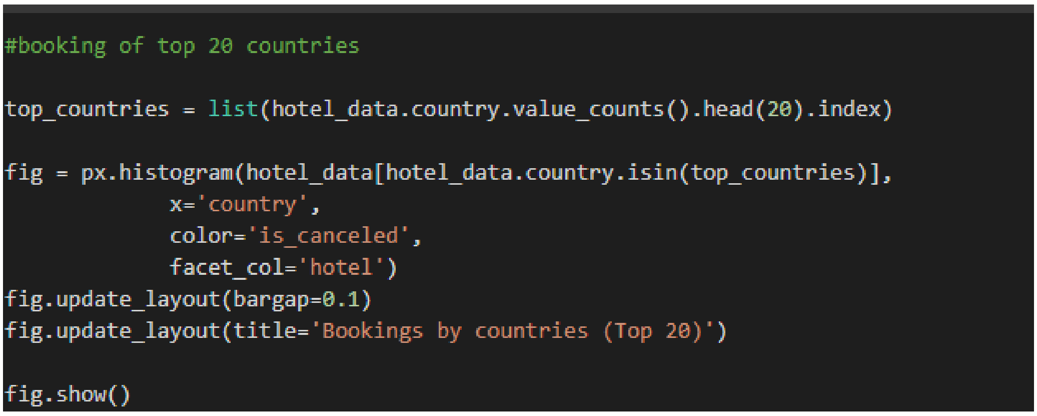
```
#booking prefernce of the customers
plt.subplots(figsize = (8,5))
sns.histplot(data = hotel_data, x = "hotel",palette = "Set2", hue = "customer_type", multiple = "stack", shrink = 0.5)
plt.title("Types of Booked Hotels", fontsize = 18)
plt.xlabel("Hotel Type", fontsize = 14)
plt.ylabel("Number of Bookings", fontsize = 14)
```

```
Text(0, 0.5, 'Number of Bookings')
```

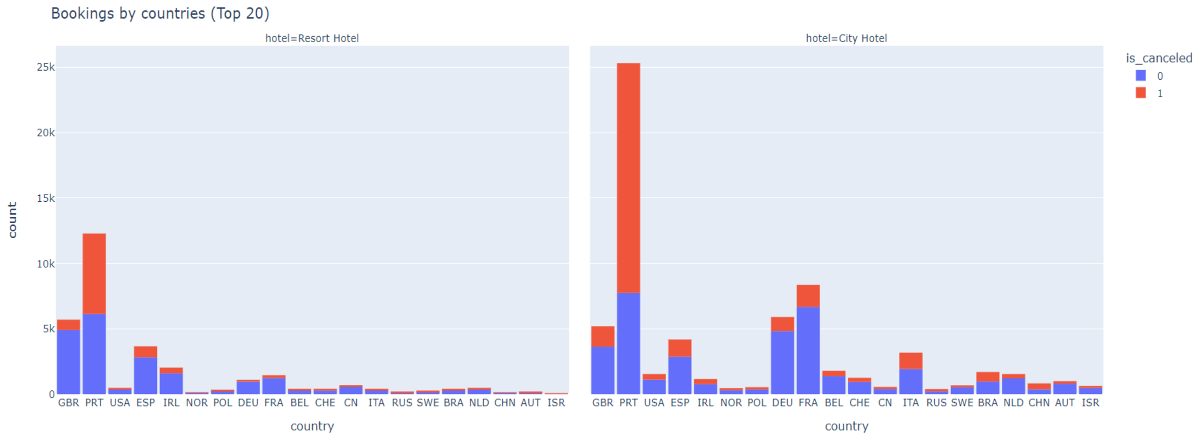
As we can see the most common type of customer are Transient type followed by the Transient-Party. And the less common type is Group type followed by Contract type.

5. Which country has the highest booking ?

For finding this first we need to create a data frame as top countries and plotly.express as library and giving x as country.



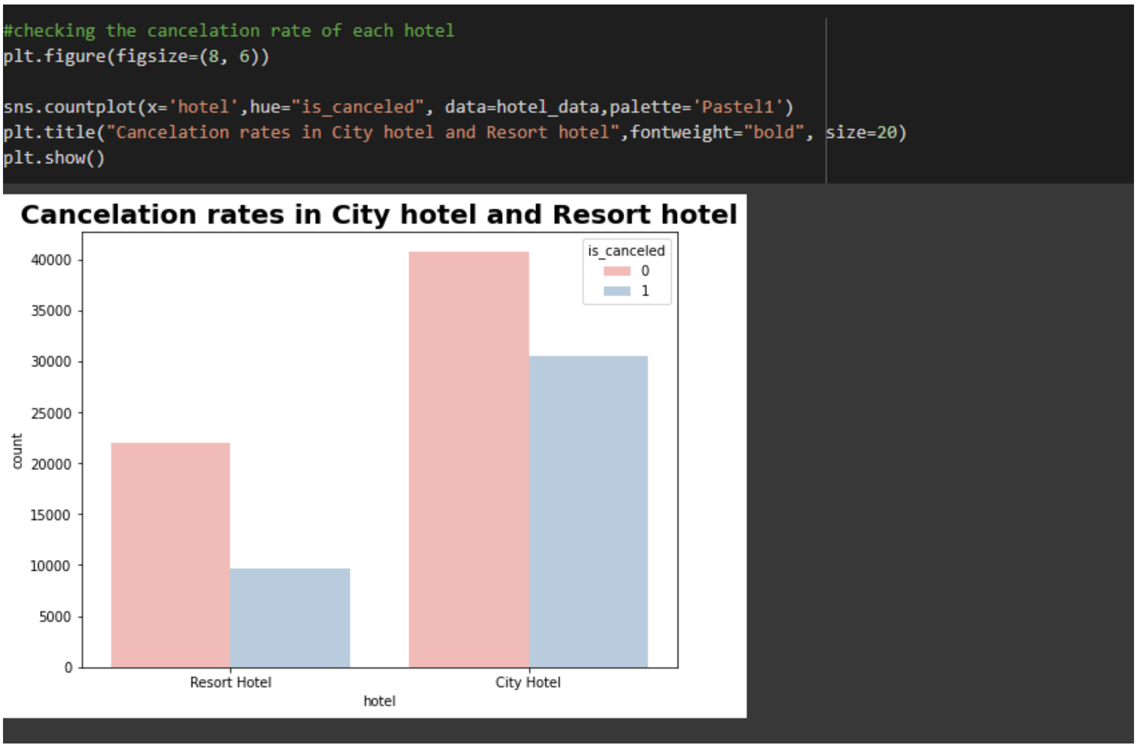
```
#booking of top 20 countries

top_countries = list(hotel_data.country.value_counts().head(20).index)

fig = px.histogram(hotel_data[hotel_data.country.isin(top_countries)],
             x='country',
             color='is_canceled',
             facet_col='hotel')
fig.update_layout(bargap=0.1)
fig.update_layout(title='Bookings by countries (Top 20)')

fig.show()
```



As we can see most guests are from Portugal followed by the United Kingdom and other countries in Europe ,as both the hotels are located in Portugal the number of people from Portugal is highest.

6. Which hotel has a high cancellation ratio ?

For finding the cancellation rate we just need to plot a graph giving x as hotel and hue as is cancelled and data as hotel data and using matplotlib.pyplot as library.

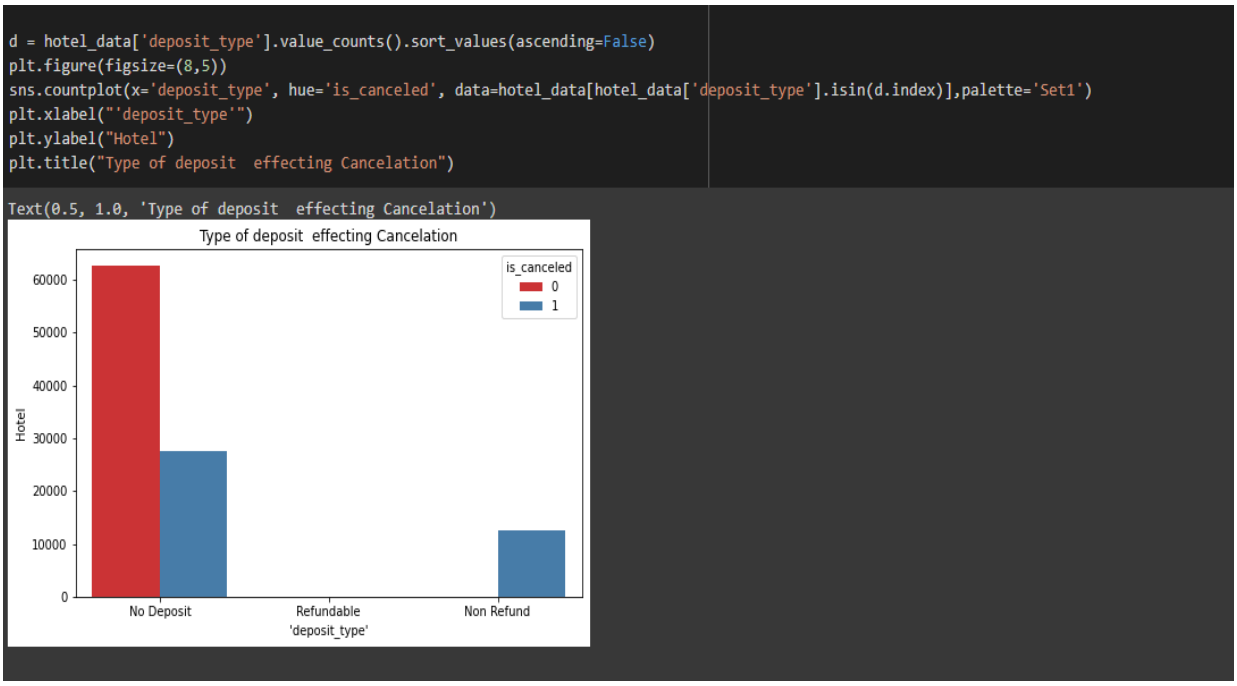


```
#checking the cancelation rate of each hotel
plt.figure(figsize=(8, 6))

sns.countplot(x='hotel',hue="is_canceled", data=hotel_data,palette='Pastel1')
plt.title("Cancelation rates in City hotel and Resort hotel",fontweight="bold", size=20)
plt.show()
```

As we can see resort hotel total bookings were 31713 and city hotel total bookings were 71181. And the cancellations for resort hotel is 9684 and city hotels are 30477. i.e 30.53% and 42.81% of people are canceling from resort and city hotels. Instead of its high cancellation ratio, city hotels have 54% more non canceled bookings.

7. what are the effects of deposit on bookings ?

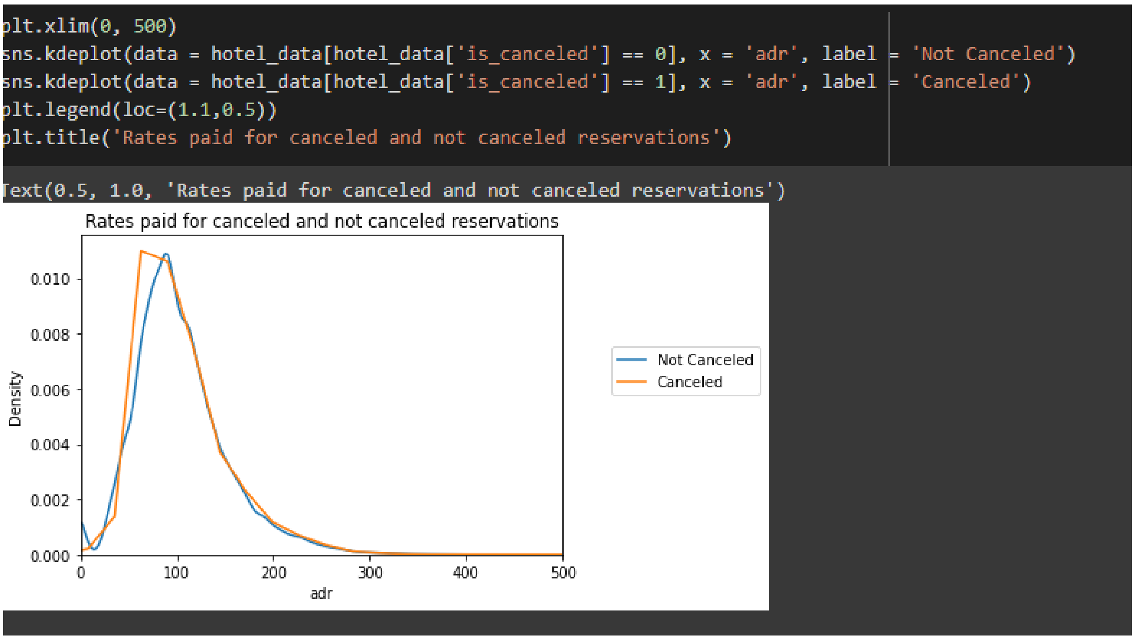
For finding this first we need to create a data frame as d which includes values of deposite type ,and then using seaborn counterplot library and x as deposit type and hue as is cancelled we can plot a graph.



```
d = hotel_data['deposit_type'].value_counts().sort_values(ascending=False)
plt.figure(figsize=(8,5))
sns.countplot(x='deposit_type', hue='is_canceled', data=hotel_data[hotel_data['deposit_type'].isin(d.index)],palette='Set1')
plt.xlabel("'deposit_type'")
plt.ylabel("Hotel")
plt.title("Type of deposit  effecting Cancelation")
```

```
Text(0.5, 1.0, 'Type of deposit  effecting Cancelation')
```

There was no deposit for City hotel where as Resorts had some deposits. No deposit may lead to cancellation of the bookings.



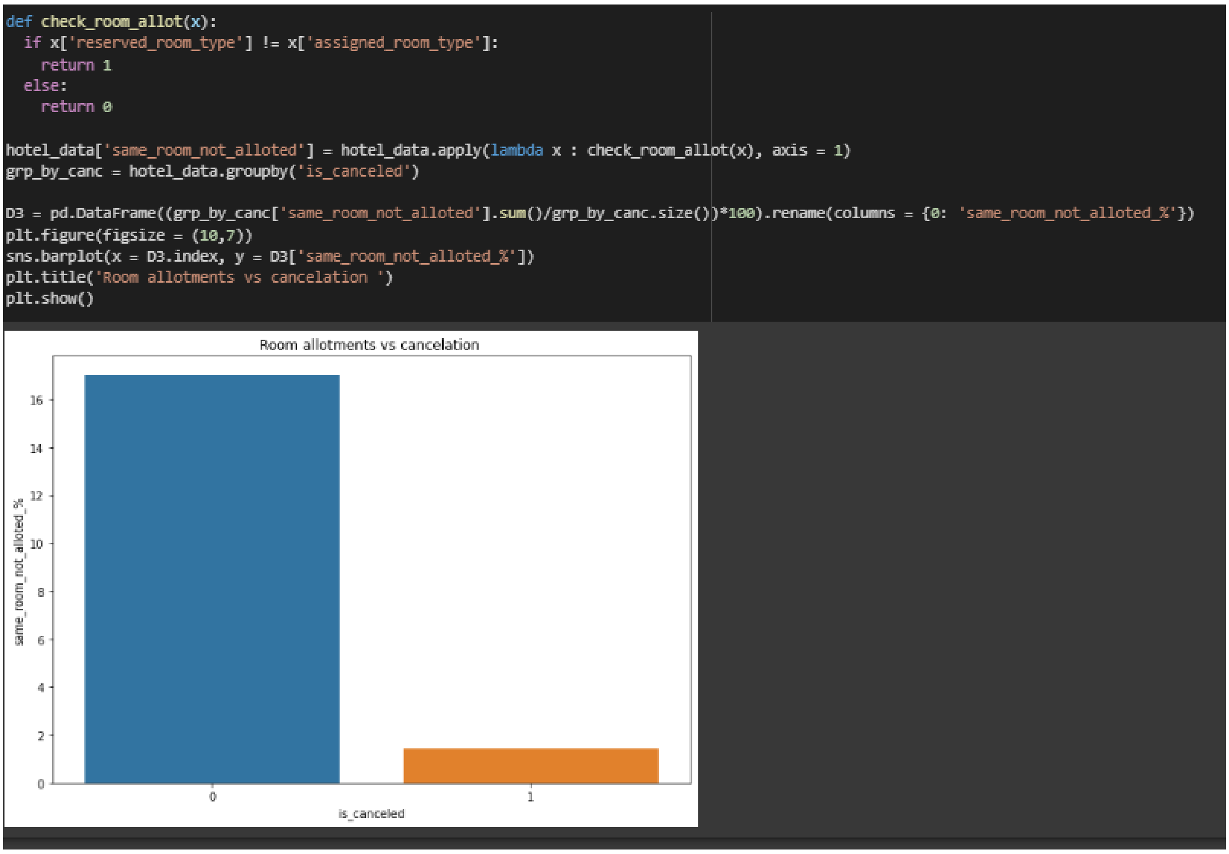
```
plt.xlim(0, 500)
sns.kdeplot(data = hotel_data[hotel_data['is_canceled'] == 0], x = 'adr', label = 'Not Canceled')
sns.kdeplot(data = hotel_data[hotel_data['is_canceled'] == 1], x = 'adr', label = 'Canceled')
plt.legend(loc=(1.1,0.5))
plt.title('Rates paid for canceled and not canceled reservations')
```

```
Text(0.5, 1.0, 'Rates paid for canceled and not canceled reservations')
```

It is interesting to note that non-refundable deposits had more cancellation than refundable deposits. Logically one would have assumed that refundable deposits have more cancellation as hotel rates are usually higher for refundable deposit type rooms and customers pay more in anticipation of cancellation.

8. **Is assigned room type the causes for high cancelation?**

For finding this we need to write an if else loop while comparing reserved rooms not equal to assigned room type . Then creating a data frame of hotel data in the same room not allotted. Then creating another data frame as D3 which includes the same room not allowed . Then by using seaborn barplot and giving x as D3.index and y as D3 same room not allotted , we can plot the graph.



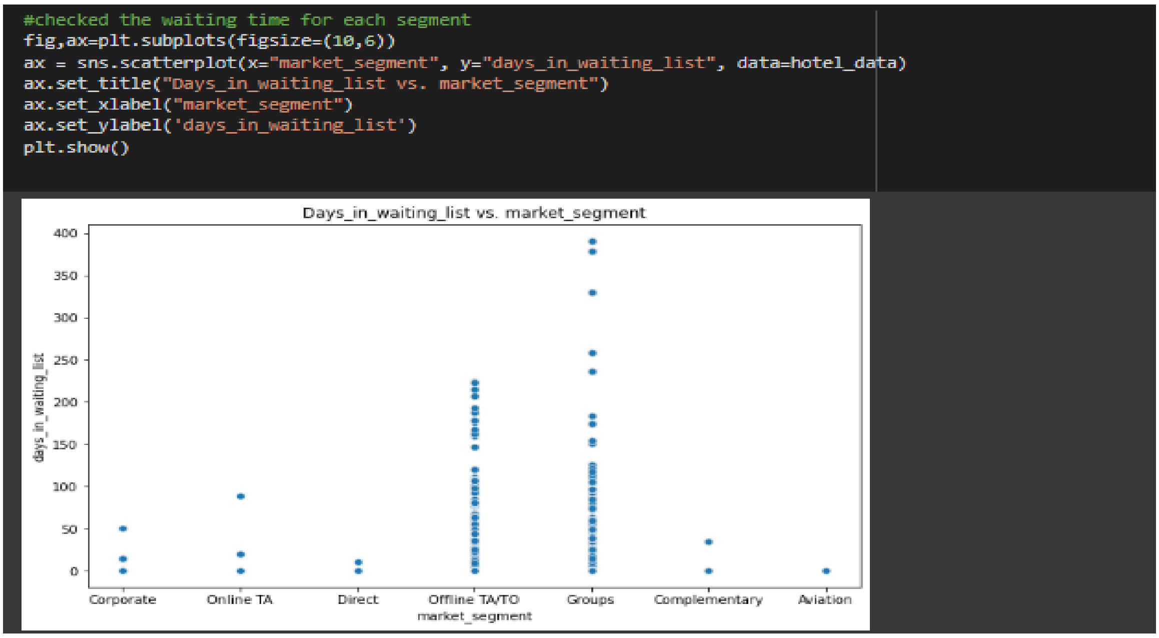
```
def check_room_allot(x):
  if x['reserved_room_type'] != x['assigned_room_type']:
    return 1
  else:
    return 0

hotel_data['same_room_not_alloted'] = hotel_data.apply(lambda x : check_room_allot(x), axis = 1)
grp_by_canc = hotel_data.groupby('is_canceled')

D3 = pd.DataFrame((grp_by_canc['same_room_not_alloted'].sum()/grp_by_canc.size())*100).rename(columns = {0: 'same_room_not_alloted_%'})
plt.figure(figsize = (10,7))
sns.barplot(x = D3.index, y = D3['same_room_not_alloted_%'])
plt.title('Room allotments vs cancelation ')
plt.show()
```

We see that not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting a different room as demanded. Less than 2% of people want to cancel their bookings due to alloted room.

Market Analysis

9. **Which segment of the market usually has the least number on the waiting list?**

For finding the market segment with least waiting list we just need to make a data frame ax as seaborn scatterplot with x as market segment and y as days in waiting list. Then plot the graph**.**

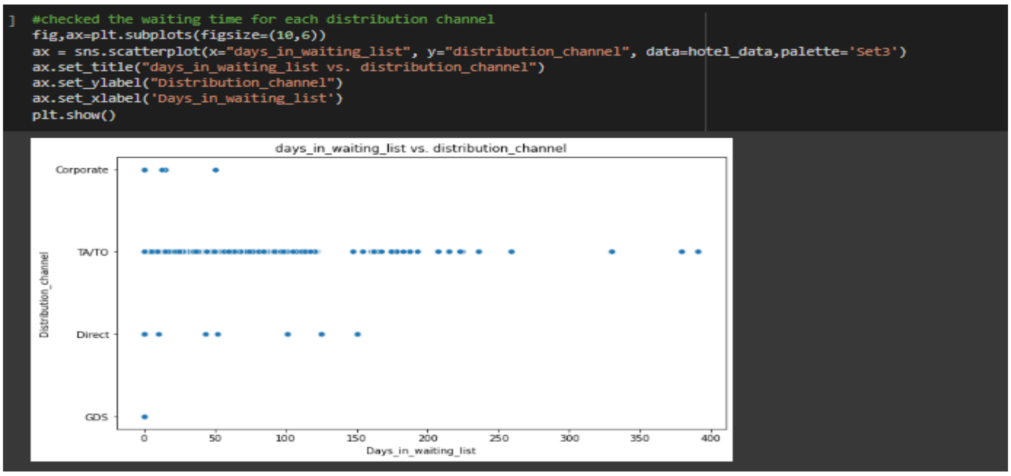


```
#checked the waiting time for each segment
fig,ax=plt.subplots(figsize=(10,6))
ax = sns.scatterplot(x="market_segment", y="days_in_waiting_list", data=hotel_data)
ax.set_title("Days_in_waiting_list vs. market_segment")
ax.set_xlabel("market_segment")
ax.set_ylabel('days_in_waiting_list')
plt.show()
```

As we can see TA/TO is mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred. While booking via TA/TO one may have to wait a little longer to confirm booking of rooms. And aviation has the least waiting list days, as there will be people who are pilots and cabin crew followed by remaining segments.

10. Which segment of the Distribution channel usually has the least number on the waiting list?

For finding the Distribution channel with least waiting list we just need to make a data frame ax as seaborn scatterplot with x as Distribution channel and y as days in waiting list. Then plot the graph.

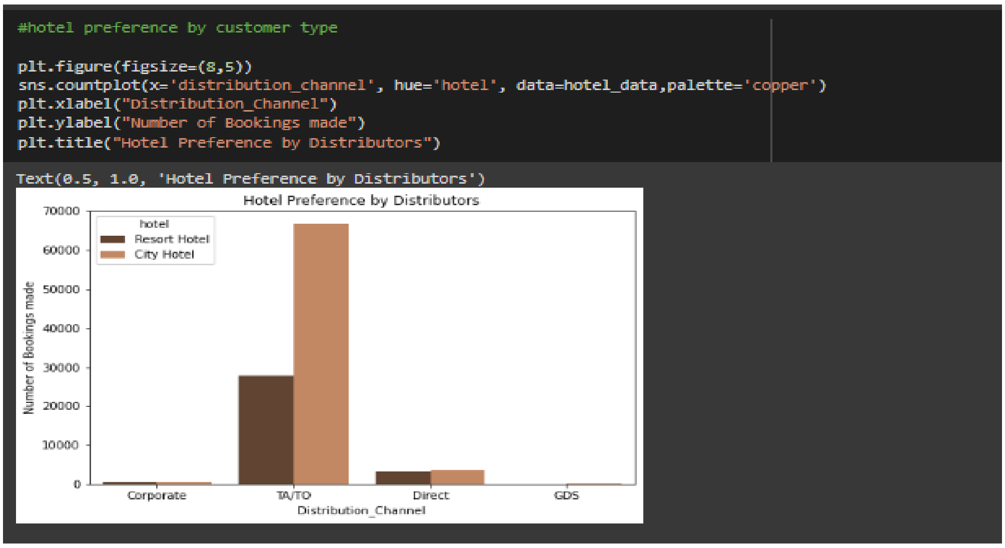


```
]  #checked the waiting time for each distribution channel
   fig,ax=plt.subplots(figsize=(10,6))
   ax = sns.scatterplot(x="days_in_waiting_list", y="distribution_channel", data=hotel_data,palette='Set3')
   ax.set_title("days_in_waiting_list vs. distribution_channel")
   ax.set_ylabel("Distribution_channel")
   ax.set_xlabel('Days_in_waiting_list')
   plt.show()
```

As we can see that global distribution systems (GDS) has the least waiting list followed by corporate and direct channel. And the highest waiting list goes for TA/TO which can extend upto 400 days.

11. Which hotel is mostly preferred by Distributors ?

For finding the hotel most preferred hotel by distributors we just need to use seaborn counterplot library with x as distribution channel and hue as hotel.



```
#hotel preference by customer type

plt.figure(figsize=(8,5))
sns.countplot(x='distribution_channel', hue='hotel', data=hotel_data,palette='copper')
plt.xlabel("Distribution_Channel")
plt.ylabel("Number of Bookings made")
plt.title("Hotel Preference by Distributors")
```

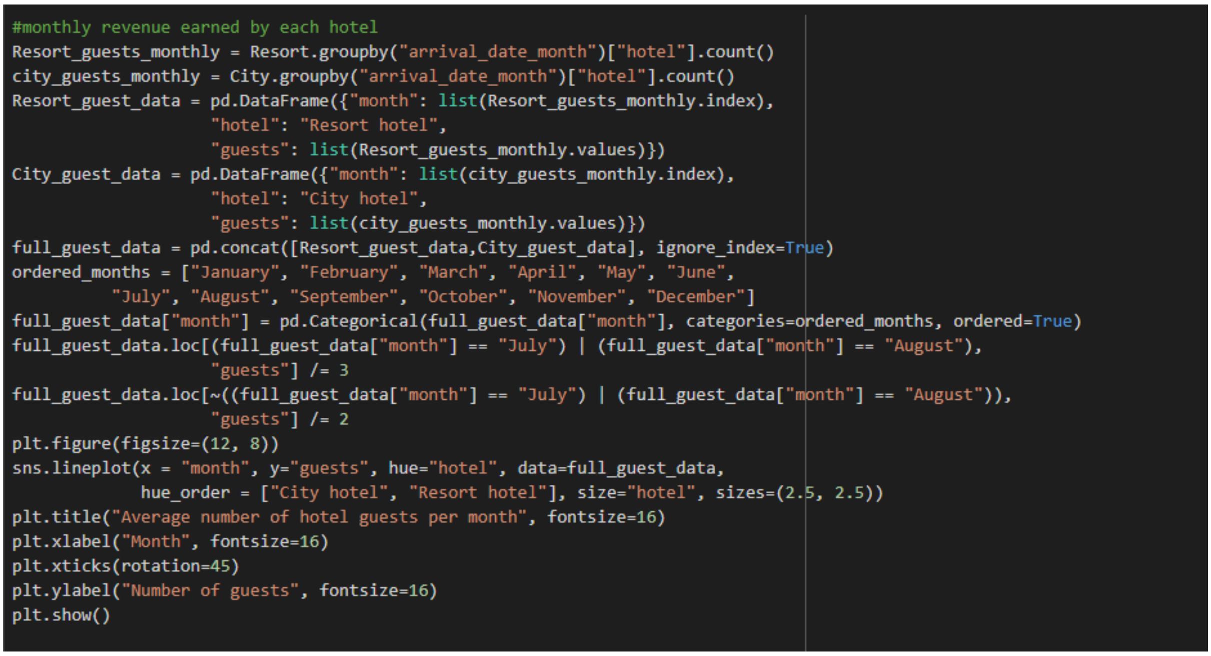
```
Text(0.5, 1.0, 'Hotel Preference by Distributors')
```

As we can see City hotel has more revenue generating deals by TA/TO channel followed by direct channel. as for resort hotels more revnue generating deals are made from TA/TO channels followed by direct channels. The preferred hotel has less influence when direct booking than by TA/TO.

Time Analysis

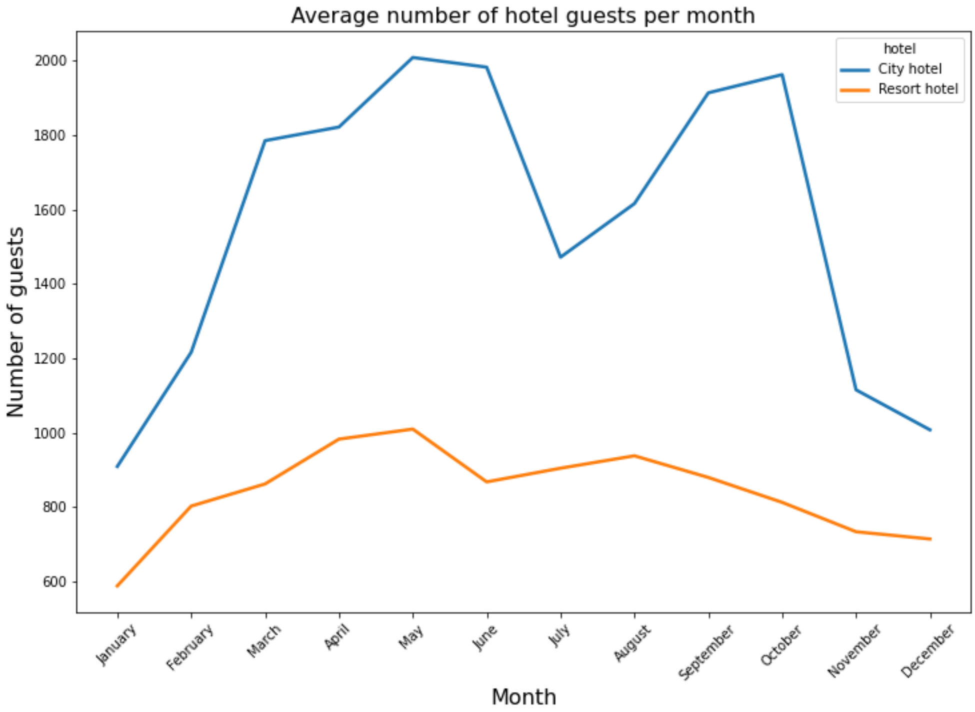
12. How average daily rate (adr) changes with different months ?

For finding adr for different months we just need to use seaborn line plot as library and x as arrival date month and



```
#monthly revenue earned by each hotel
Resort_guests_monthly = Resort.groupby("arrival_date_month")["hotel"].count()
city_guests_monthly = City.groupby("arrival_date_month")["hotel"].count()
Resort_guest_data = pd.DataFrame({"month": list(Resort_guests_monthly.index),
                    "hotel": "Resort hotel",
                    "guests": list(Resort_guests_monthly.values)})
City_guest_data = pd.DataFrame({"month": list(city_guests_monthly.index),
                    "hotel": "City hotel",
                    "guests": list(city_guests_monthly.values)})
full_guest_data = pd.concat([Resort_guest_data,City_guest_data], ignore_index=True)
ordered_months = ["January", "February", "March", "April", "May", "June",
          "July", "August", "September", "October", "November", "December"]
full_guest_data["month"] = pd.Categorical(full_guest_data["month"], categories=ordered_months, ordered=True)
full_guest_data.loc[(full_guest_data["month"] == "July") | (full_guest_data["month"] == "August"),
                    "guests"] /= 3
full_guest_data.loc[~((full_guest_data["month"] == "July") | (full_guest_data["month"] == "August")),
                    "guests"] /= 2
plt.figure(figsize=(12, 8))
sns.lineplot(x = "month", y="guests", hue="hotel", data=full_guest_data,
             hue_order = ["City hotel", "Resort hotel"], size="hotel", sizes=(2.5, 2.5))
plt.title("Average number of hotel guests per month", fontsize=16)
plt.xlabel("Month", fontsize=16)
plt.xticks(rotation=45)
plt.ylabel("Number of guests", fontsize=16)
plt.show()
```

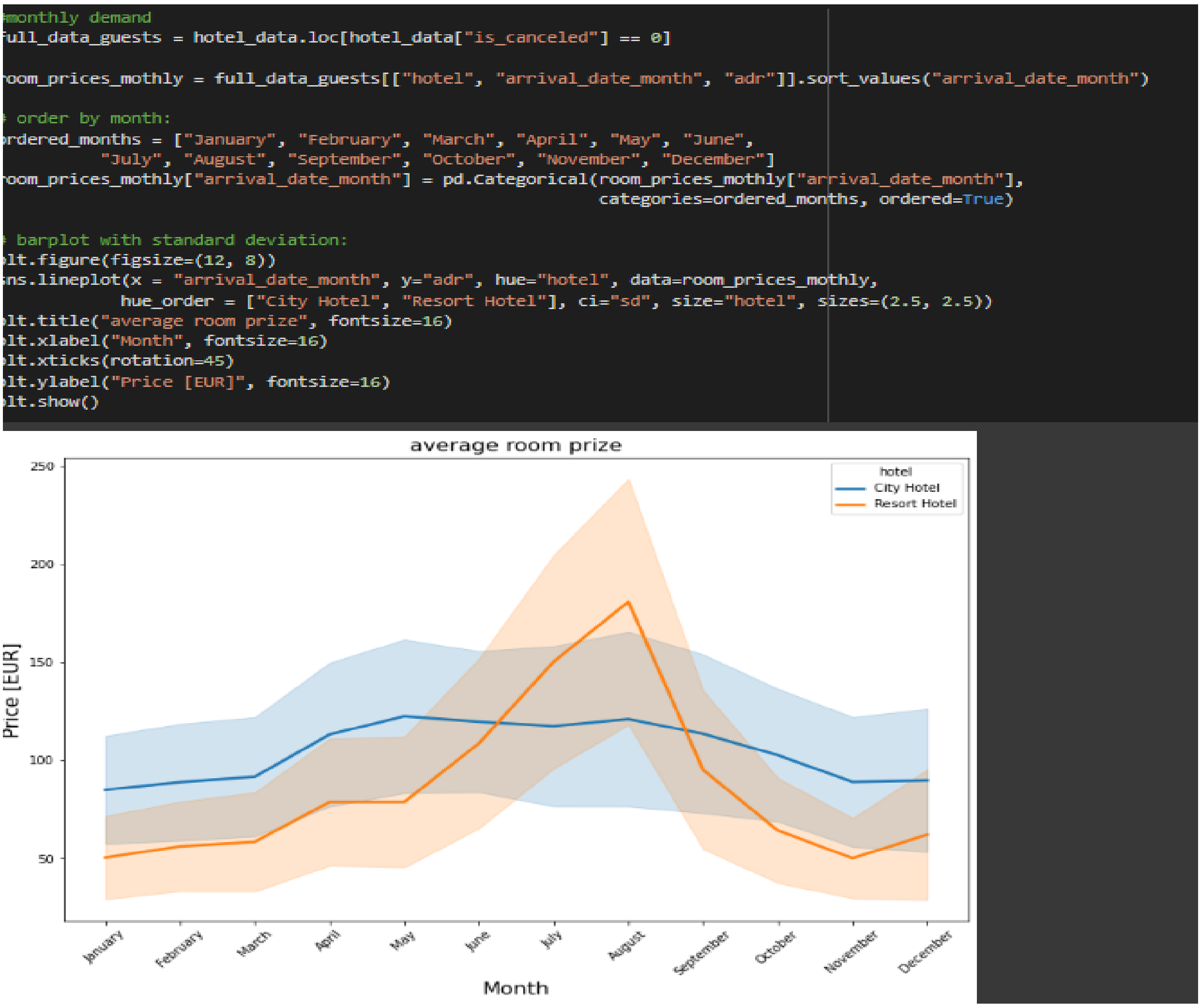
y as adr ,hue as hotel and data as hotel data.



As we can see this plot clearly shows that prices in the Resort Hotel are much higher during the summer and prices of city hotel is more during March, April & May. Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.

13. What is the busiest month of the year?

For finding the busiest month of the year we need to use sns line plot library and x as x arrival date month and y as total guests, hue as hotel and dada as hotel data.

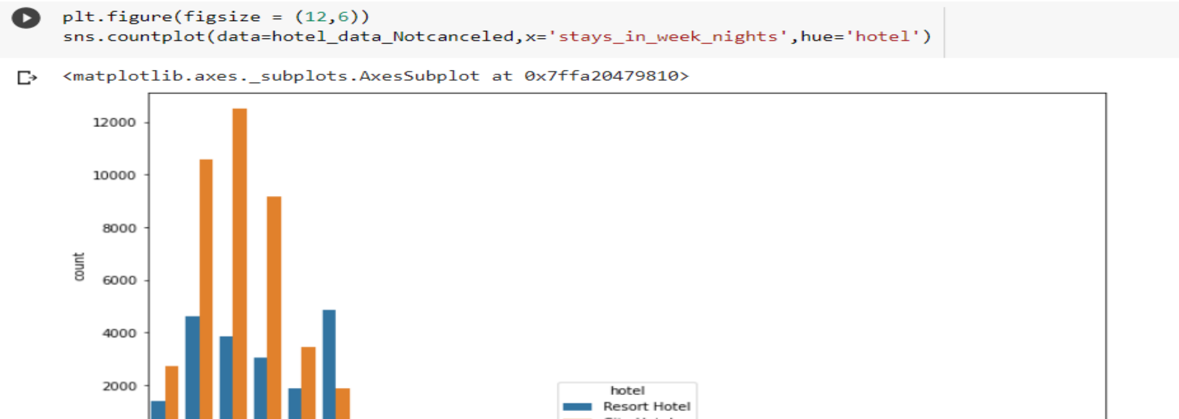


```
#monthly demand
full_data_guests = hotel_data.loc[hotel_data["is_canceled"] == 0]

room_prices_mothly = full_data_guests[["hotel", "arrival_date_month", "adr"]].sort_values("arrival_date_month")

# order by month:
ordered_months = ["January", "February", "March", "April", "May", "June",
          "July", "August", "September", "October", "November", "December"]
room_prices_mothly["arrival_date_month"] = pd.Categorical(room_prices_mothly["arrival_date_month"],
                                                          categories=ordered_months, ordered=True)

# barplot with standard deviation:
plt.figure(figsize=(12, 8))
sns.lineplot(x = "arrival_date_month", y="adr", hue="hotel", data=room_prices_mothly,
            hue_order = ["City Hotel", "Resort Hotel"], ci="sd", size="hotel", sizes=(2.5, 2.5))
plt.title("average room prize", fontsize=16)
plt.xlabel("Month", fontsize=16)
plt.xticks(rotation=45)
plt.ylabel("Price [EUR]", fontsize=16)
plt.show()
```

As we can see the city hotel and resort hotel has its peak guest values in the month of august. The resort hotel has more guests during july and august, when the prices are also highest. Guest numbers for the Resort hotel go down slightly from September. Both hotels have the fewest guests during the winter i.e November, even when price are lowest.

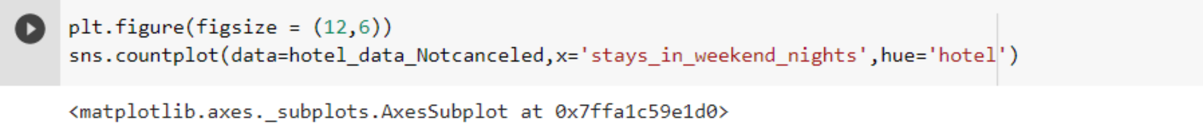
14. Which is most preferred on weekdays or weekends ?

For most preferred weekends and weekdays we created 2 figs , first create a data frame as data frame not cancelled is true and then plot a fig using seaborn counterplot library and data as hotel data not cancelled and x as stays in weekend nights and hue as hotel. And for creating weekend fig change x as stay in week nights.



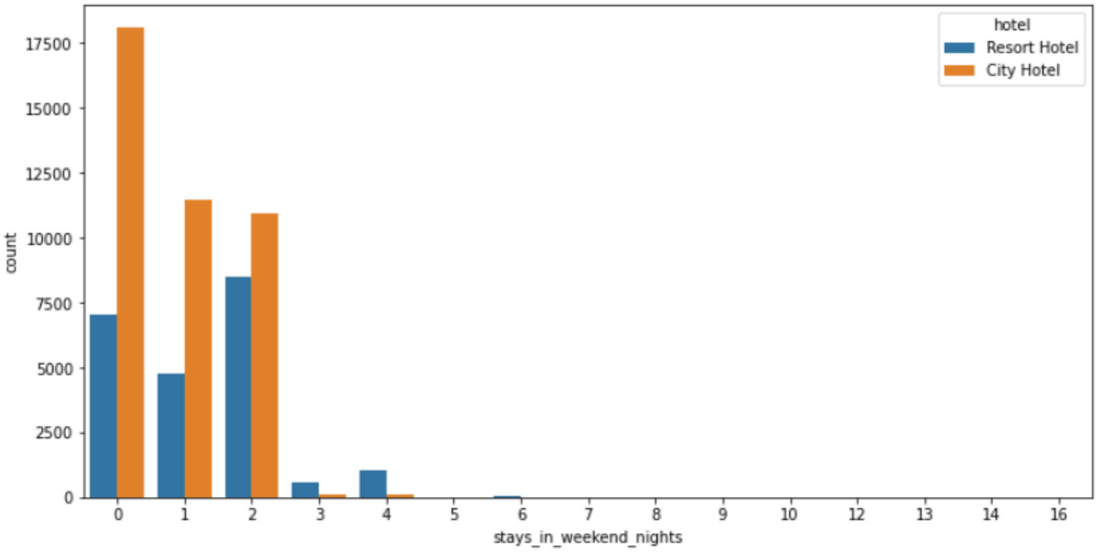
```
plt.figure(figsize = (12,6))
sns.countplot(data=hotel_data_Notcanceled,x='stays_in_week_nights',hue='hotel')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ffa20479810>
```



```
plt.figure(figsize = (12,6))
sns.countplot(data=hotel_data_Notcanceled,x='stays_in_weekend_nights',hue='hotel')
```

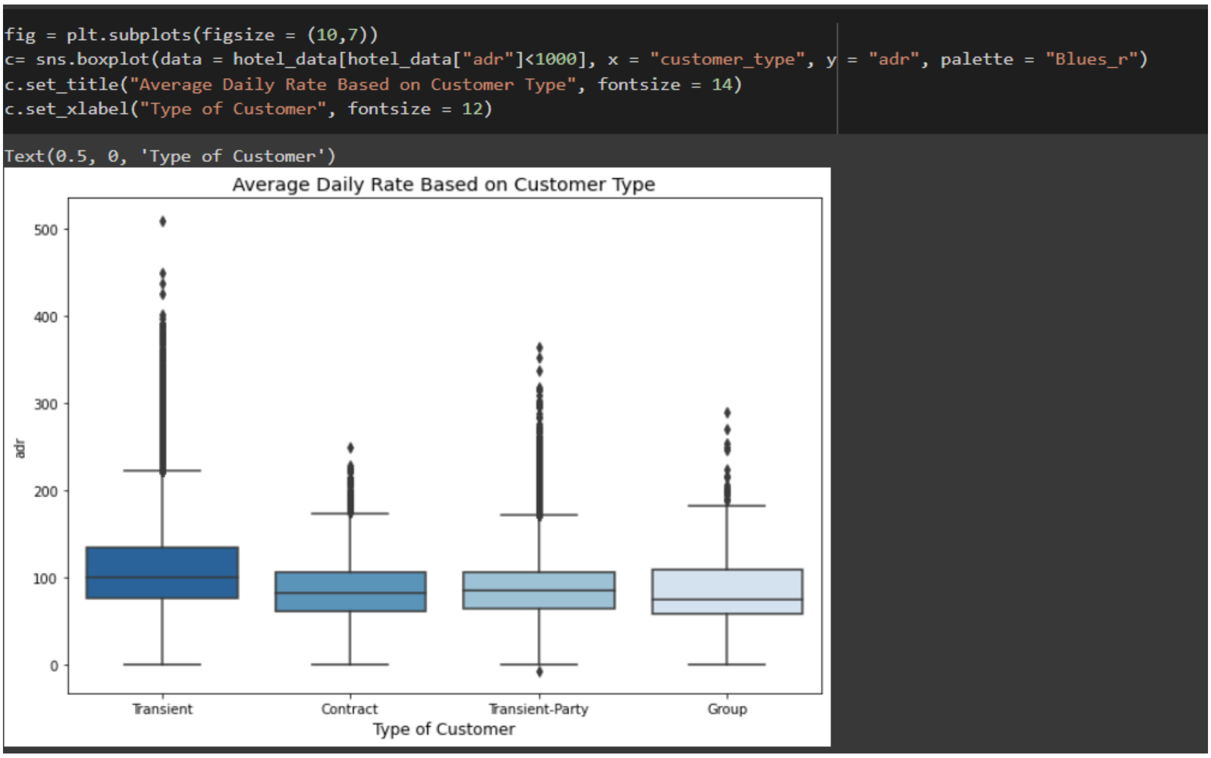
```
<matplotlib.axes._subplots.AxesSubplot at 0x7ffa1c59e1d0>
```



As we can see from the both the graphs majority of the people from city hotel tends to stay 0 weekend nights followed by 1 and 2 weekend nights ,very few people tends to stay longer than that, and many people tend to stay 1 to 3 days on an avg. In case of resort hotel people tends to stay at least 2 weekends days and most people tend to stay 4 days.

15. How average daily rate (adr) changes with Customer Type ?

For finding changes of adr with customer type first create a data frame c as seaborn boxplot data frame and data as hotel data with adr < 1000 and x as customer type and y as adr.



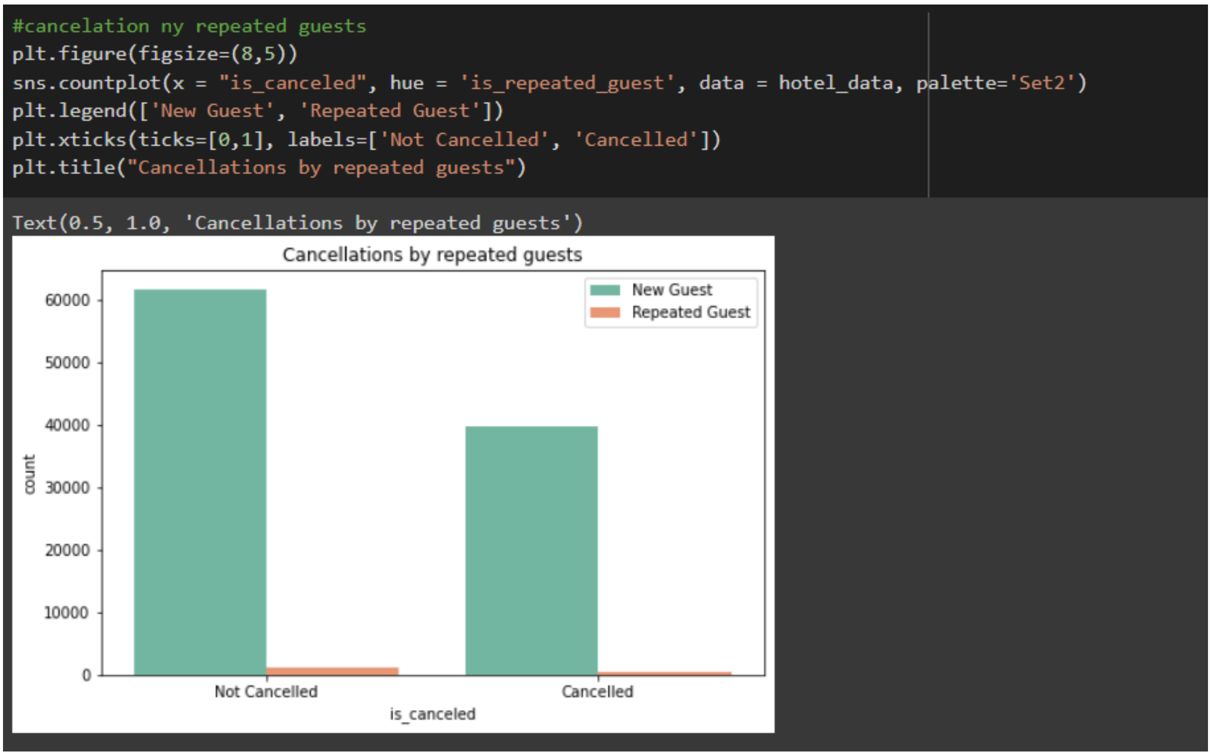
```
fig = plt.subplots(figsize = (10,7))
c= sns.boxplot(data = hotel_data[hotel_data["adr"]<1000], x = "customer_type", y = "adr", palette = "Blues_r")
c.set_title("Average Daily Rate Based on Customer Type", fontsize = 14)
c.set_xlabel("Type of Customer", fontsize = 12)
```

```
Text(0.5, 0, 'Type of Customer')
```

As we can see, the contract customer type has the least average daily rate(adr)followed by group and transient-Party and the highest average daily rate(adr) is for transient customer type.

16. what is the chances of repeated guest cancelling the booking ?

for finding the chances of repeated guest cancelling the booking we first need to use seaborn counterplot library and give x as is cancelled and hue as is repeated guest and data as hotel data .



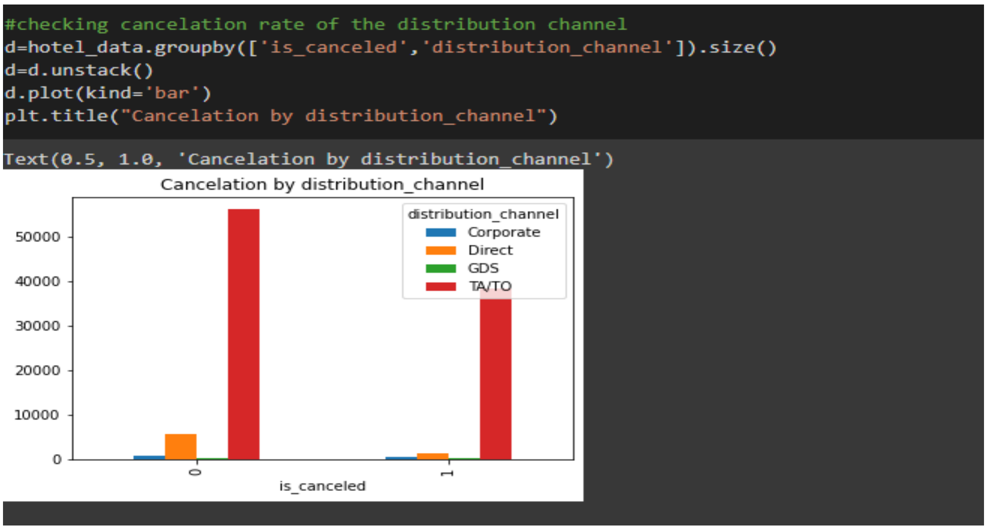
```
#cancelation ny repeated guests
plt.figure(figsize=(8,5))
sns.countplot(x = "is_canceled", hue = 'is_repeated_guest', data = hotel_data, palette='Set2')
plt.legend(['New Guest', 'Repeated Guest'])
plt.xticks(ticks=[0,1], labels=['Not Cancelled', 'Cancelled'])
plt.title("Cancellations by repeated guests")
```

```
Text(0.5, 1.0, 'Cancellations by repeated guests')
```

As we can see, most of repeated guests do not cancel their reservations. Of course there are some exceptions. Also most of the customers are not repeated guests.

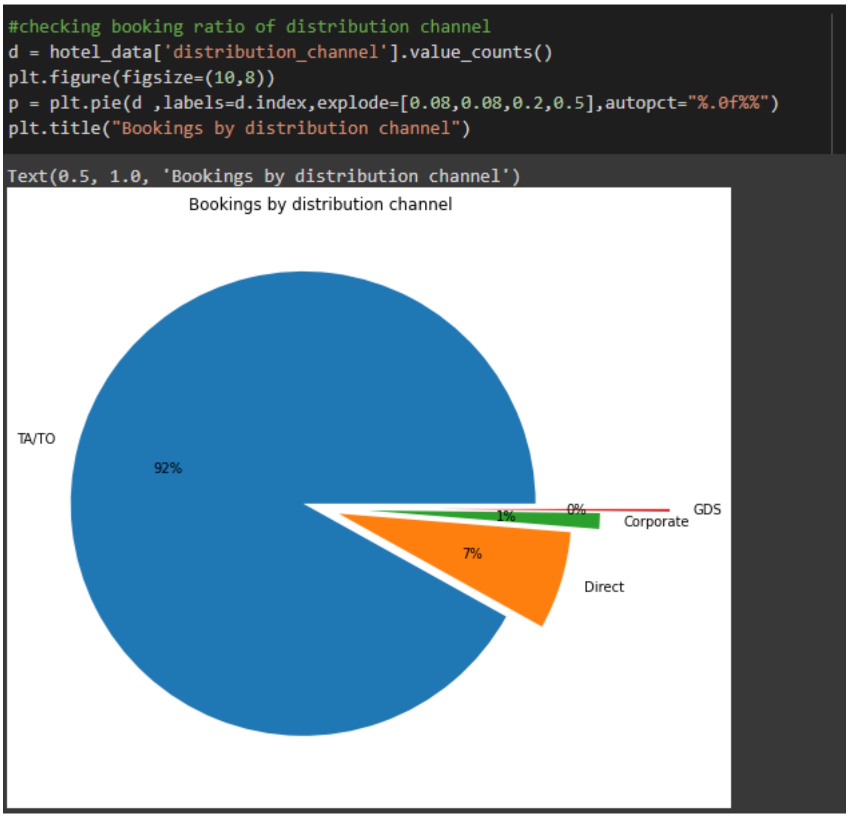
17. which segment of distribution channel have highest bookings and cancelations?

For finding which segment of booking has highest bookings and cancellations we just need to create a data frame d as hotel data distribution channel and use matplotlib pie library with d and labels as d.index. For plotting bar graph use new data frame d as hotel data with grouping of is cancelled and distribution channel. Use d as d. unstack for unstacking them



```
#checking cancelation rate of the distribution channel
d=hotel_data.groupby(['is_canceled','distribution_channel']).size()
d=d.unstack()
d.plot(kind='bar')
plt.title("Cancelation by distribution_channel")
```

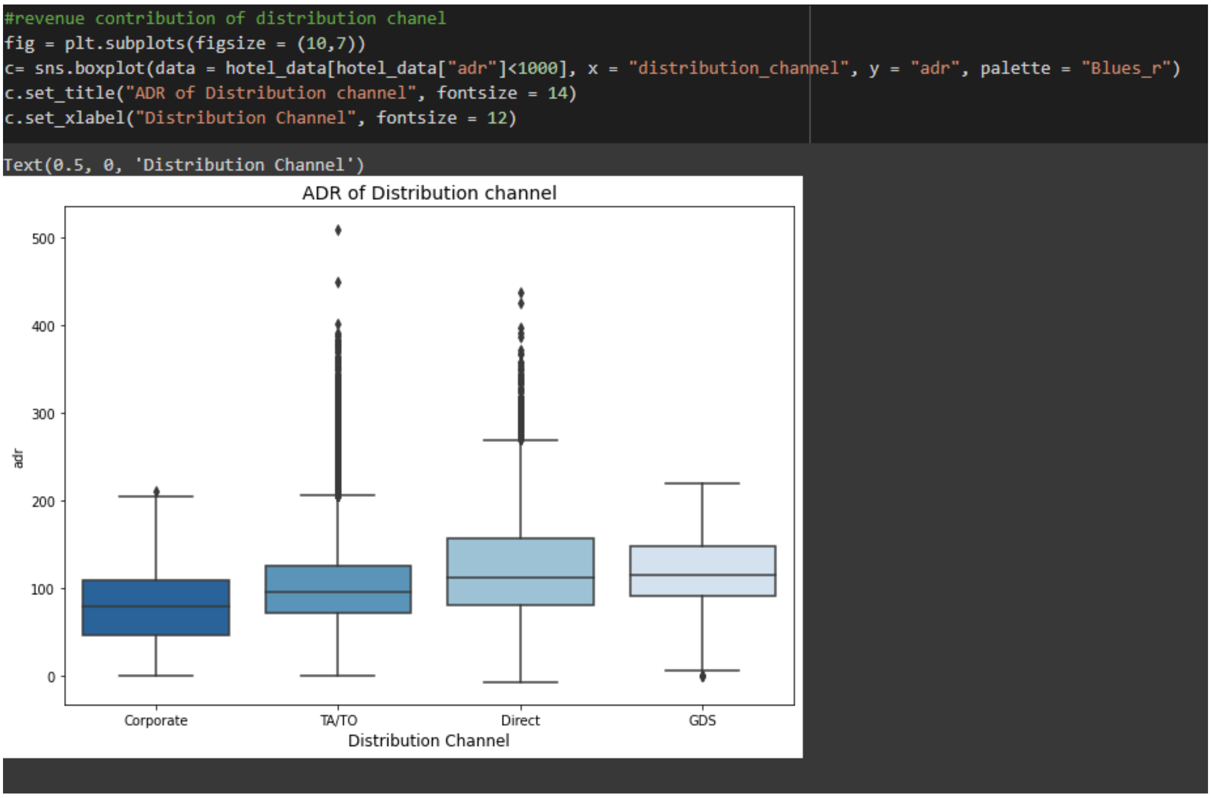
```
Text(0.5, 1.0, 'Cancelation by distribution_channel')
```



```
#checking booking ratio of distribution channel
d = hotel_data['distribution_channel'].value_counts()
plt.figure(figsize=(10,8))
p = plt.pie(d ,labels=d.index,explode=[0.08,0.08,0.2,0.5],autopct="%.0f%%")
plt.title("Bookings by distribution channel")
```

```
Text(0.5, 1.0, 'Bookings by distribution channel')
```

As we can see TA/TO has highest booking of 91% and cancellation of 40% followed by direct distribution channel.



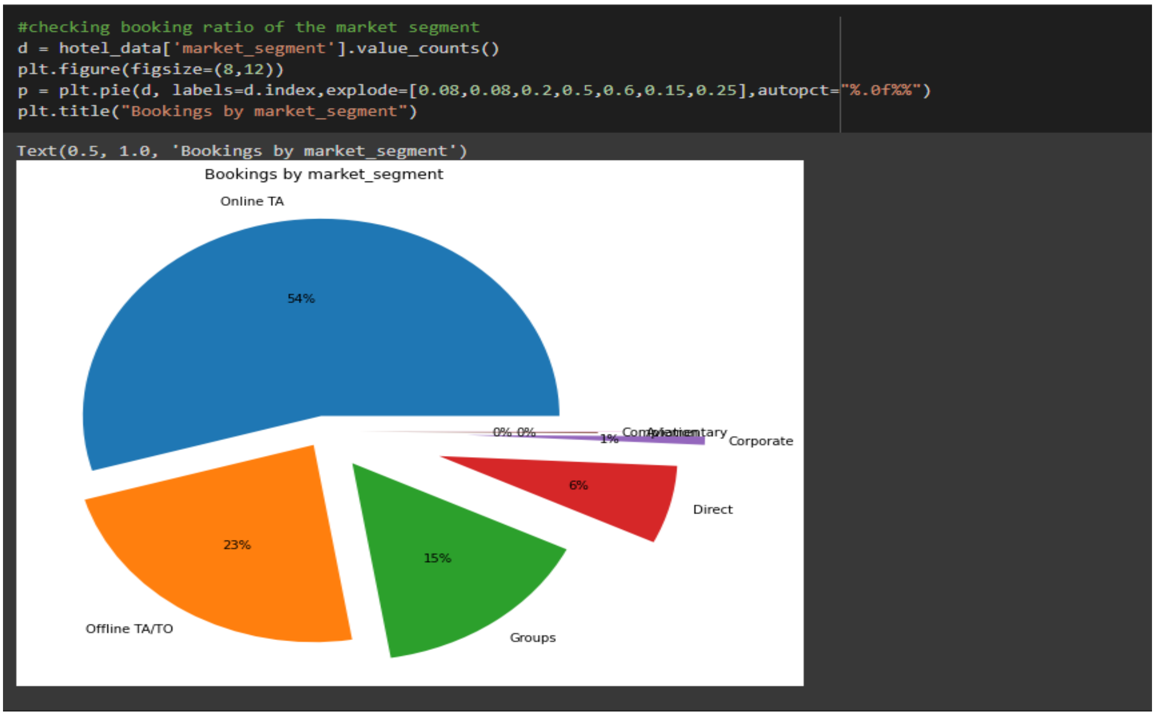
```
#revenue contribution of distribution chanel
fig = plt.subplots(figsize = (10,7))
c= sns.boxplot(data = hotel_data[hotel_data["adr"]<1000], x = "distribution_channel", y = "adr", palette = "Blues_r")
c.set_title("ADR of Distribution channel", fontsize = 14)
c.set_xlabel("Distribution Channel", fontsize = 12)
```

```
Text(0.5, 0, 'Distribution Channel')
```

We can see here the maximum revenue is contributed by the direct walkin customers onl

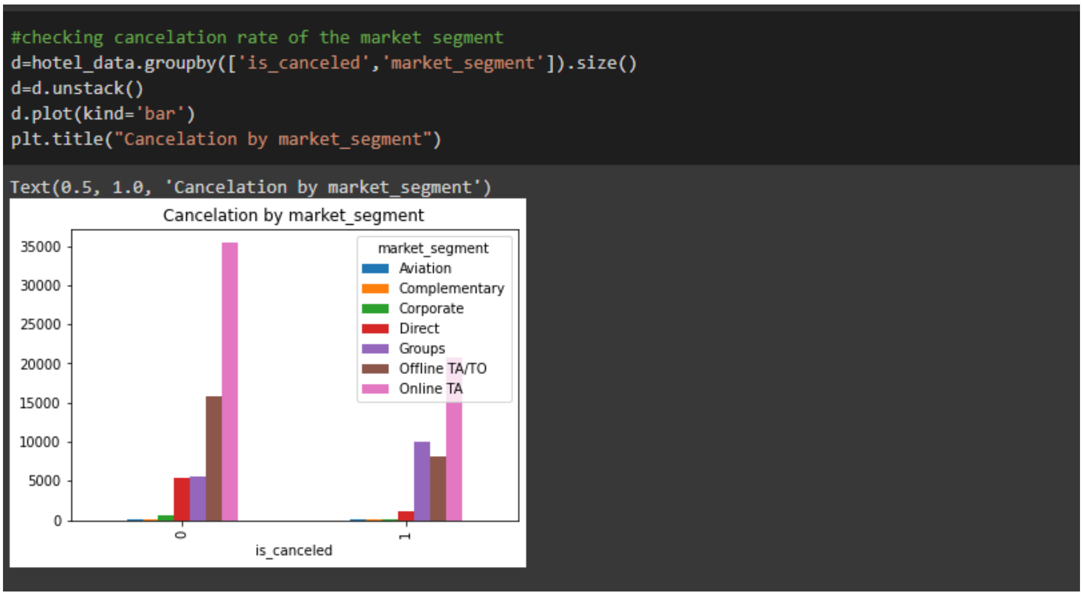
18. which market segment have highest bookings and cancelations?

For finding which segment of booking has highest bookings and cancellations we just need to create a data frame d as hotel data market segment and use matplotlib pie library with d and labels as d.index. For plotting bar graph use new data frame d as hotel data with grouping of is cancelled and market segment . Use d as d. unstack for unstacking them



```
#checking booking ratio of the market segment
d = hotel_data['market_segment'].value_counts()
plt.figure(figsize=(8,12))
p = plt.pie(d, labels=d.index,explode=[0.08,0.08,0.2,0.5,0.6,0.15,0.25],autopct="%.0f%%")
plt.title("Bookings by market_segment")
```

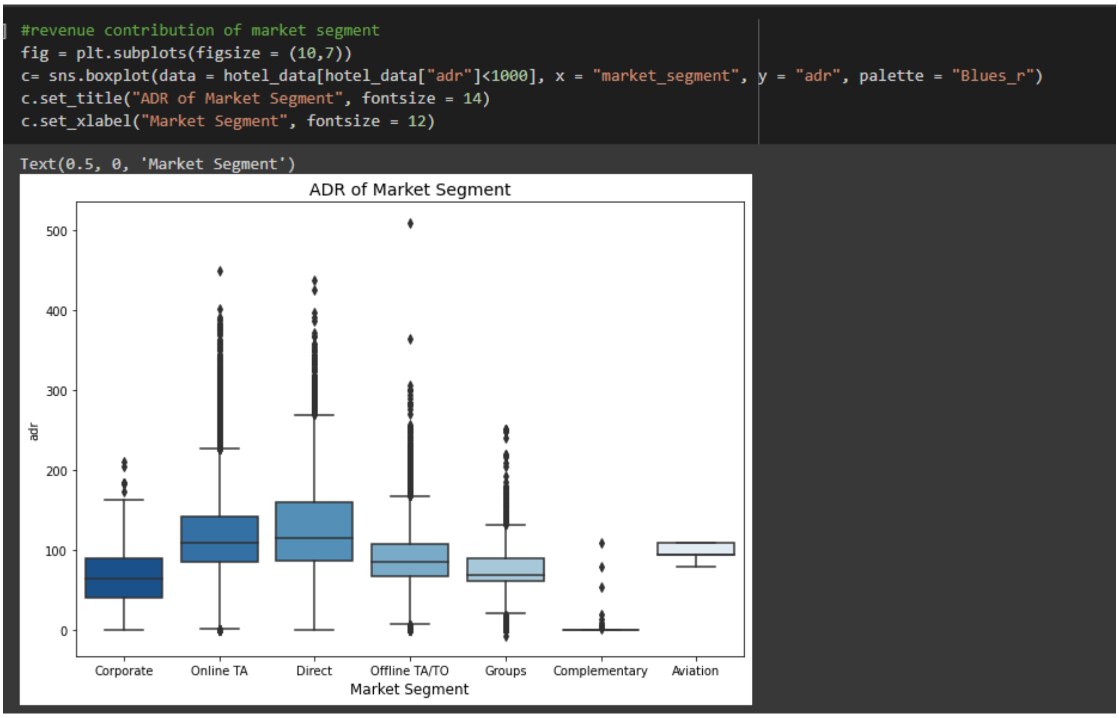
```
Text(0.5, 1.0, 'Bookings by market_segment')
```



```
#checking cancelation rate of the market segment
d=hotel_data.groupby(['is_canceled','market_segment']).size()
d=d.unstack()
d.plot(kind='bar')
plt.title("Cancelation by market_segment")
```

```
Text(0.5, 1.0, 'Cancelation by market_segment')
```

Offline TA/TO (Travel Agents/Tour Operators) and Online TA has booking rates of 23% and 54%, cancellation rate of 33.8% and 36.9%. It is surprising that the cancellation rate in these segments is high despite the application of a deposit. The fact that cancellations are made collectively like group reservations has a high cancellation rate. It is interesting to note that the Group segment has a booking rate of 15.2% and a cancellation rate of 35%.



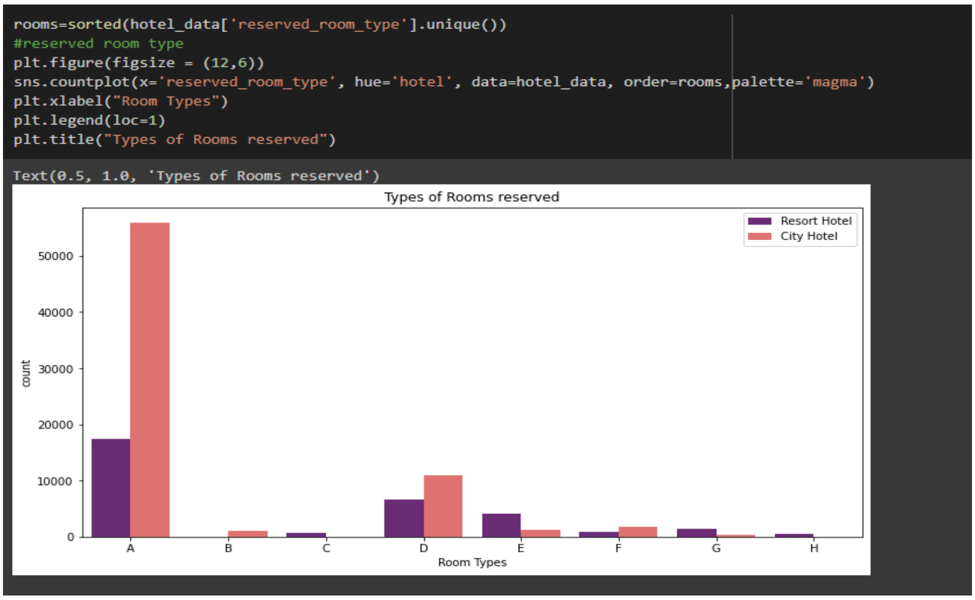
```
#revenue contribution of market segment
fig = plt.subplots(figsize = (10,7))
c= sns.boxplot(data = hotel_data[hotel_data["adr"]<1000], x = "market_segment", y = "adr", palette = "Blues_r")
c.set_title("ADR of Market Segment", fontsize = 14)
c.set_xlabel("Market Segment", fontsize = 12)
```

```
Text(0.5, 0, 'Market Segment')
```

We can see here the maximum revenue is contributed by the direct walkin customers only.

19. Which type of rooms are mostly reserved ?

For finding the room which is mostly reserved we just need to create a data frame rooms which is sorted form of hotel data reserved room type . Then by using seaborn counterplot library and x as reserved room type , hue as hotel and data as hotel data plot the graph.



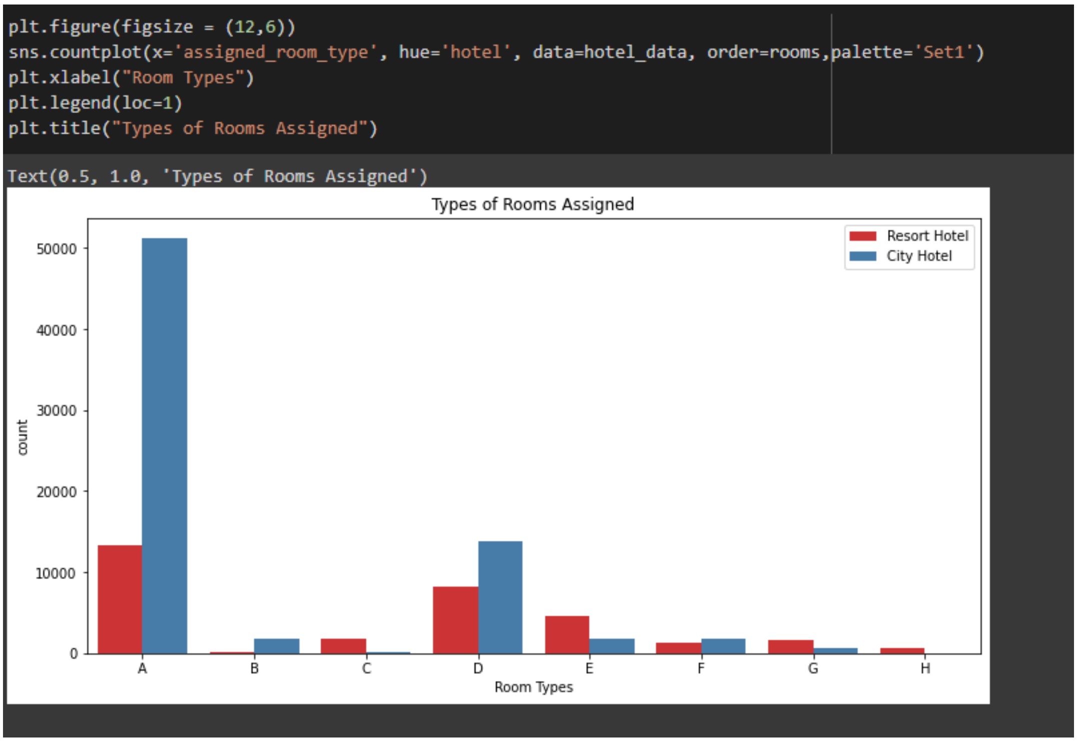
```
rooms=sorted(hotel_data['reserved_room_type'].unique())
#reserved room type
plt.figure(figsize = (12,6))
sns.countplot(x='reserved_room_type', hue='hotel', data=hotel_data, order=rooms,palette='magma')
plt.xlabel("Room Types")
plt.legend(loc=1)
plt.title("Types of Rooms reserved")
```

```
Text(0.5, 1.0, 'Types of Rooms reserved')
```

As we can see that mostly reserved rooms are A and D types .

20. **Which type of rooms are mostly assigned ?**

For finding the room which is mostly reserved we just need to create a data frame rooms which is sorted form of hotel data assigned room type . Then by using seaborn counterplot library and x as assigned room type , hue as hotel and data as hotel data plot the graph.



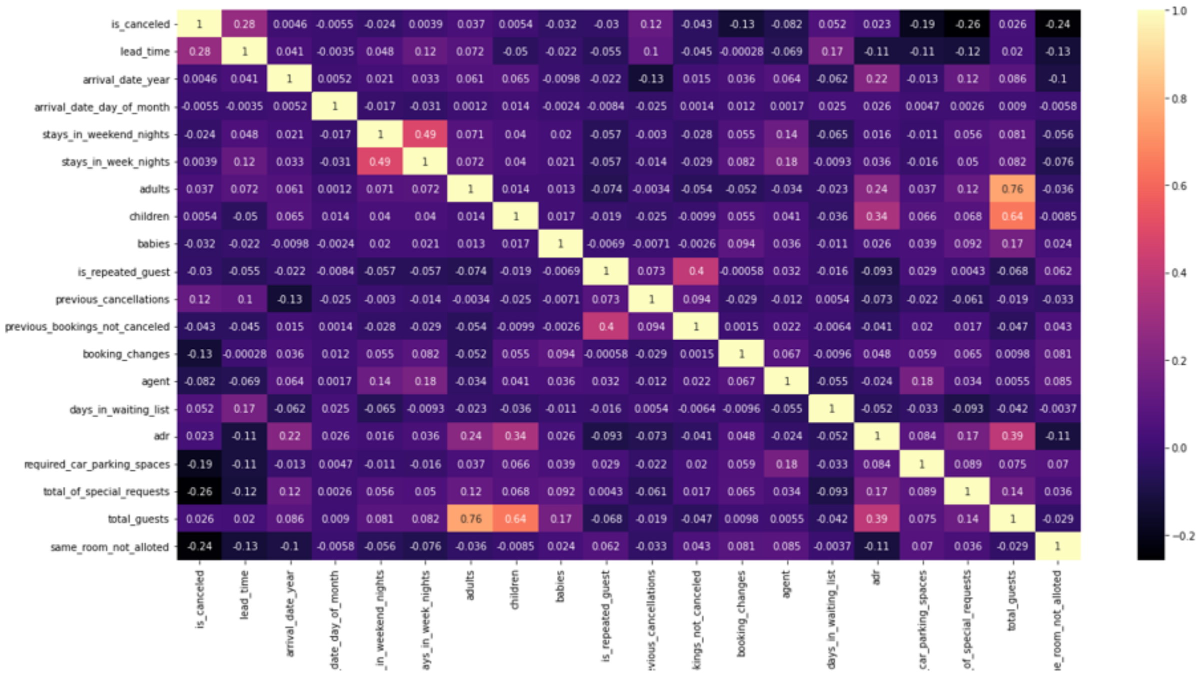
```
plt.figure(figsize = (12,6))
sns.countplot(x='assigned_room_type', hue='hotel', data=hotel_data, order=rooms,palette='Set1')
plt.xlabel("Room Types")
plt.legend(loc=1)
plt.title("Types of Rooms Assigned")
```

```
Text(0.5, 1.0, 'Types of Rooms Assigned')
```

As we can see the most assigned rooms are A and D followed by E.

Finding different correlations:

Now lets plot heat map with creating a data frame c as hotel data corr and using seaborn heatmap library.



We can see that lead time and previous cancellations have a higher correlation with is cancelled than most of the other columns.

Conclusion:

1. The City hotel is the most preferred hotel.
2. On an average, guests of City hotel stay 3 nights and guests of the Resort hotel stay 4 nights.
3. As lead time increases there is a high chance of canceling the booking.
4. The most common type of customer are Transient type followed by Transient-Party.
5. Most guests are from Portugal, as the hotels resides in that place.
6. The highest cancellation ratio of bookings is done in City hotel.But city hotel have 54% more non canceled bookings compared to resort hotel.
7. There was no deposit for City hotel where as Resorts had some deposits.It is interesting to note that non-refundable deposits had more cancellation than refundable deposits.
8. Not getting same room as demanded is not the case of cancellation of rooms.
9. When taking market segment into consideration Aviation has the least waiting list days and TA/TO has the highest waiting list days.
10. When taking distribution channel into consideration global distribution systems (GDS) has the least waiting list followed by corporate and direct channel.

11. Both city hotel and resort hotel has more revenue generating deals by TA/TO channel followed by direct channel.
12. The prices in the Resort Hotel are much higher during the summer and prices of city hotel is more during March, April & May.
13. The city hotel and resort hotel has its peak guest in the month of august.
14. The majority of the people from city hotel tends to stay 0 weekend nights and the majority of the people from resort hotel tends to stay 2 weekends days.
15. The contract customer type has the least average daily rate(adr)followed by group and transient-Party and the highest average daily rate(adr) is for transient customer type.
16. Most of the repeated guests do not cancel their reservations and most of the customers are not repeated guests.
17. With respect to market segment Online TA has the highest booking of 54% and cancellation of 36.9% followed by Offline TA/TO.
18. With respect to distribution channel TA/TO has highest booking of 91% and cancellation of 40% followed by direct distribution channel.
19. The mostly reserved rooms are A and D types and most assigned rooms are A and D followed by E.
20. The lead_time and previous_cancellations have a higher correlation with is_cancelled than most of the other columns.