Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Importing the dataset

In [2]:
```
data=pd.read_csv('/Users/bbkpa/Downloads/hotel_bookings 2.csv')
data
```

Out[2]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month
0	Resort Hotel	0	342	2015	July
1	Resort Hotel	0	737	2015	July
2	Resort Hotel	0	7	2015	July
3	Resort Hotel	0	13	2015	July
4	Resort Hotel	0	14	2015	July
...	...	...	...	...	...
119385	City Hotel	0	23	2017	August
119386	City Hotel	0	102	2017	August
119387	City Hotel	0	34	2017	August
119388	City Hotel	0	109	2017	August
119389	City Hotel	0	205	2017	August

119390 rows × 32 columns

Exploratory Data Analysis and Data Cleaning

In [3]:
```
data.shape
```

Out[3]: (119390, 32)

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           119390 non-null  object 
 1   is_canceled                     119390 non-null  int64  
 2   lead_time                       119390 non-null  int64  
 3   arrival_date_year               119390 non-null  int64  
 4   arrival_date_month              119390 non-null  object 
 5   arrival_date_week_number        119390 non-null  int64  
 6   arrival_date_day_of_month       119390 non-null  int64  
 7   stays_in_weekend_nights         119390 non-null  int64  
 8   stays_in_week_nights            119390 non-null  int64  
 9   adults                          119390 non-null  int64  
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64  
 12  meal                            119390 non-null  object 
 13  country                         118902 non-null  object 
 14  market_segment                  119390 non-null  object 
 15  distribution_channel            119390 non-null  object 
 16  is_repeated_guest               119390 non-null  int64  
 17  previous_cancellations          119390 non-null  int64  
 18  previous_bookings_not_canceled  119390 non-null  int64  
 19  reserved_room_type              119390 non-null  object 
 20  assigned_room_type              119390 non-null  object 
 21  booking_changes                 119390 non-null  int64  
 22  deposit_type                    119390 non-null  object 
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64  
 26  customer_type                   119390 non-null  object 
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

In [5]:
```
data['reservation_status_date']=pd.to_datetime(data['reservation_status_date'],form
data
```

Out[5]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month
0	Resort Hotel	0	342	2015	July
1	Resort	0	737	2015	July
2	Resort Hotel	0	7	2015	July

3	Resort Hotel	0	13	2015	July
4	Resort Hotel	0	14	2015	July
...	...	...	...	...	...
119385	City Hotel	0	23	2017	August
119386	City Hotel	0	102	2017	August
119387	City Hotel	0	34	2017	August
119388	City Hotel	0	109	2017	August
119389	City Hotel	0	205	2017	August

119390 rows × 32 columns

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
```

```
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

Information about all categorical columns

In [7]:
```
data.describe(include='object')
```

Out[7]:

	hotel	arrival_date_month	meal	country	market_segment	distrib
count	119390	119390	119390	118902	119390	
unique	2	12	5	177	8	
top	City Hotel	August	BB	PRT	Online TA	
freq	79330	13877	92310	48590	56477	

All unique values in the categorical columns of the dataset

In [8]:
```
for col in data.describe(include='object'):
    print(col)
    print(data[col].unique())
    print('_'*50)
```

```
hotel
['Resort Hotel' 'City Hotel']
__________________________________________________
arrival_date_month
['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
__________________________________________________
meal
['BB' 'FB' 'HB' 'SC' 'Undefined']
__________________________________________________
country
['PRT' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' nan 'ROU' 'NOR' 'OMN' 'ARG' 'POL'
 'DEU' 'BEL' 'CHE' 'CN' 'GRC' 'ITA' 'NLD' 'DNK' 'RUS' 'SWE' 'AUS' 'EST'
 'CZE' 'BRA' 'FIN' 'MOZ' 'BWA' 'LUX' 'SVN' 'ALB' 'IND' 'CHN' 'MEX' 'MAR'
 'UKR' 'SMR' 'LVA' 'PRI' 'SRB' 'CHL' 'AUT' 'BLR' 'LTU' 'TUR' 'ZAF' 'AGO'
 'ISR' 'CYM' 'ZMB' 'CPV' 'ZWE' 'DZA' 'KOR' 'CRI' 'HUN' 'ARE' 'TUN' 'JAM'
 'HRV' 'HKG' 'IRN' 'GEO' 'AND' 'GIB' 'URY' 'JEY' 'CAF' 'CYP' 'COL' 'GGY'
 'KWT' 'NGA' 'MDV' 'VEN' 'SVK' 'FJI' 'KAZ' 'PAK' 'IDN' 'LBN' 'PHL' 'SEN'
 'SYC' 'AZE' 'BHR' 'NZL' 'THA' 'DOM' 'MKD' 'MYS' 'ARM' 'JPN' 'LKA' 'CUB'
 'CMR' 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SYR' 'SGP' 'BDI'
 'SAU' 'VNM' 'PLW' 'QAT' 'EGY' 'PER' 'MLT' 'MWI' 'ECU' 'MDG' 'ISL' 'UZB'
 'NPL' 'BHS' 'MAC' 'TGO' 'TWN' 'DJI' 'STP' 'KNA' 'ETH' 'IRQ' 'HND' 'RWA'
 'KHM' 'MCO' 'BGD' 'IMN' 'TJK' 'NIC' 'BEN' 'VGB' 'TZA' 'GAB' 'GHA' 'TMP'
 'GLP' 'KEN' 'LIE' 'GNB' 'MNE' 'UMI' 'MYT' 'FRO' 'MMR' 'PAN' 'BFA' 'LBY'
 'MLI' 'NAM' 'BOL' 'PRY' 'BRB' 'ABW' 'AIA' 'SLV' 'DMA' 'PYF' 'GUY' 'LCA'
```

```
 'ATA' 'GTM' 'ASM' 'MRT' 'NCL' 'KIR' 'SDN' 'ATF' 'SLE' 'LAO']
--------------------------------------------------
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']
--------------------------------------------------
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
--------------------------------------------------
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
--------------------------------------------------
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K']
--------------------------------------------------
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
--------------------------------------------------
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
--------------------------------------------------
reservation_status
['Check-Out' 'Canceled' 'No-Show']
--------------------------------------------------
```

Descriptive Statistics

In [9]:
```
data.describe()
```

Out[9]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_num
count	119390.000000	119390.000000	119390.000000	119390.00
mean	0.370416	104.011416	2016.156554	27.16
min	0.000000	0.000000	2015.000000	1.00
25%	0.000000	18.000000	2016.000000	16.00
50%	0.000000	69.000000	2016.000000	28.00
75%	1.000000	160.000000	2017.000000	38.00
max	1.000000	737.000000	2017.000000	53.00
std	0.482918	106.863097	0.707476	13.60

8 rows × 21 columns

All unique values in the categorical column 'stays_in_weekend_nights' of the dataset

In [10]:
```
data['stays_in_weekend_nights'].value_counts()
```

Out[10]:
```
stays_in_weekend_nights
0     51998
2     33308
1     30626
4      1855
3      1259
6       153
5        79
8        60
7        19
9        11
10        7
12        5
13        3
16        3
14        2
18        1
19        1
Name: count, dtype: int64
```

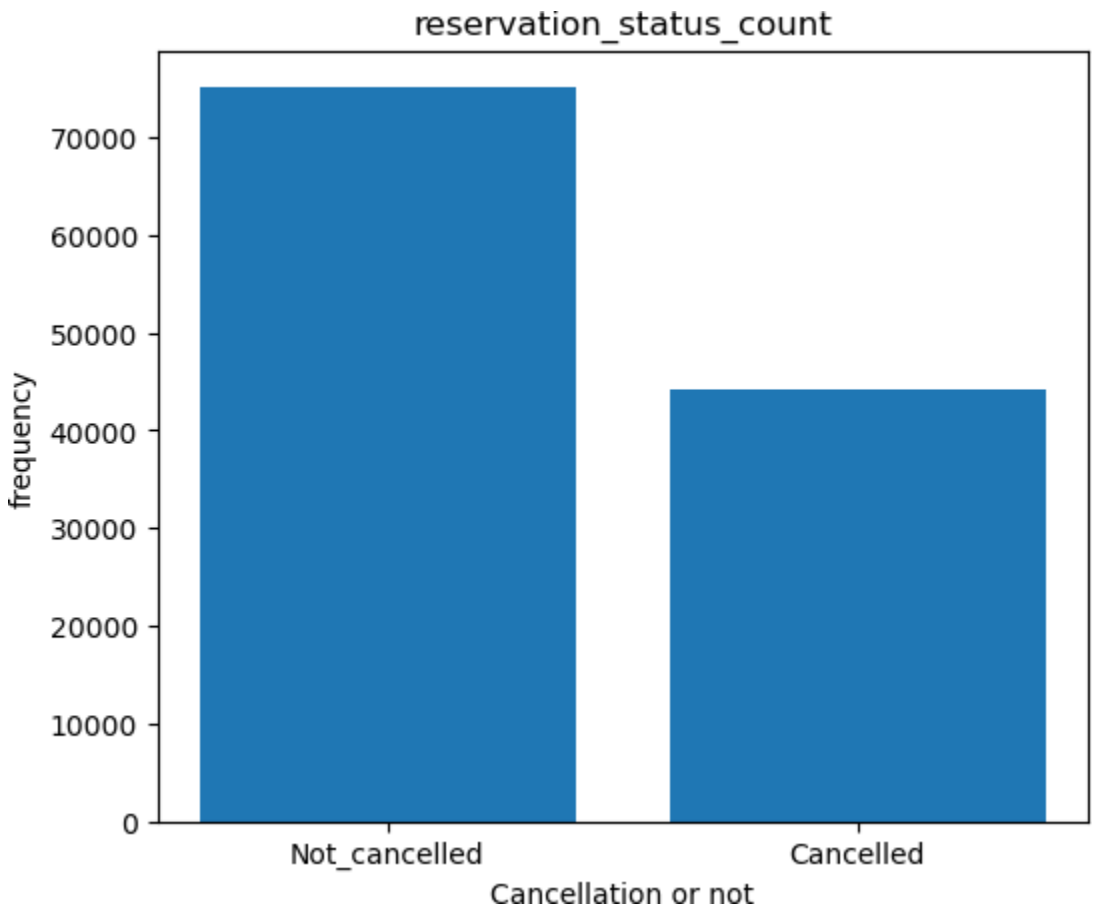
Data Analysis and Visualization

In [11]:
```
data=data[data['adr']<5000]
```

Reservation status in both the hotels

In [12]:
```
canceled_per=data['is_canceled'].value_counts(normalize=True)
print(canceled_per)
plt.figure(figsize=(6,5))
plt.bar(['Not_cancelled','Cancelled'],data['is_canceled'].value_counts())
plt.xlabel('Cancellation or not')
plt.ylabel('frequency')
plt.title('reservation_status_count')
plt.show()
```

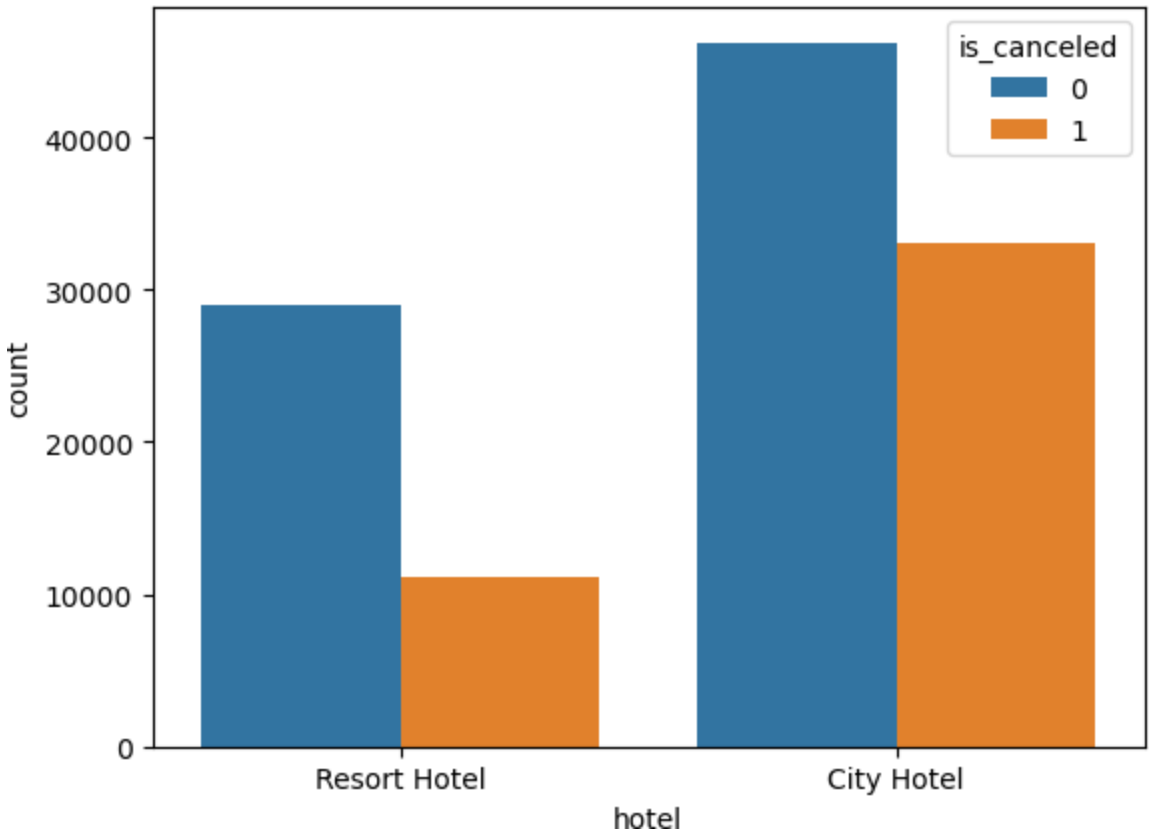
```
is_canceled
0    0.629589
1    0.370411
Name: proportion, dtype: float64
```



Reservation Status in both hotels(City And Resort) seperately

In [13]:
```
data['hotel']=data['hotel'].astype(str)
data['is_canceled']=data['is_canceled'].astype(str)
```

In [14]:
```
sns.countplot(data=data,x='hotel',hue='is_canceled')
plt.show()
```



In [15]:
```
data[data['hotel']=='Resort  Hotel'].sample(5)
```

Out[15]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	a
12802	Resort Hotel	1	133	2017	July	
25914	Resort Hotel	0	132	2016	July	
16634	Resort Hotel	0	68	2015	August	
11956	Resort Hotel	1	318	2017	June	
31298	Resort Hotel	0	2	2016	December	

5 rows × 32 columns

Reservation status in Resort hotel (0-Not Cancelled) (1-Cancelled)

In [16]:
```
resort_hotel=data[data['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)
```

Out[16]:
```
is_canceled
0    0.722366
```

```
1    0.277634
Name: proportion, dtype: float64
```

Reservation status in City hotel (0-Not Cancelled) (1-Cancelled)

In [17]:
```
city_hotel=data[data['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)
```

Out[17]:
```
is_canceled
0    0.582738
1    0.417262
Name: proportion, dtype: float64
```

Average daily rate in Resort hotel per day

In [18]:
```
resort_hotel=resort_hotel.groupby('reservation_status_date')[['adr']].mean()
resort_hotel
```

Out[18]:

reservation_status_date	adr
2014-11-18	0.000000
2015-01-01	61.966667
2015-01-05	115.363333
2015-01-06	133.677143
2015-01-07	82.485455
...	...
2017-12-05	103.287534
2017-12-06	159.808929
2017-12-07	160.306275
2017-12-08	212.767222
2017-12-09	153.570000

913 rows × 1 columns

Average daily rate in City hotel per day

In [19]:
```
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel
```

Out[19]:

reservation_status_date	adr

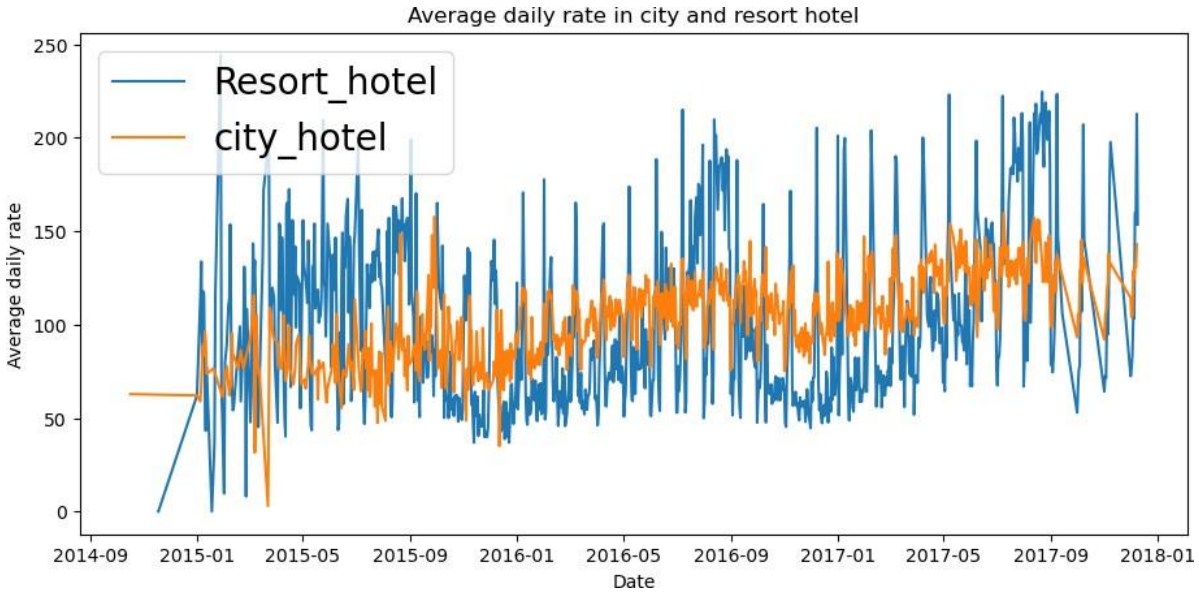
2014-10-17	62.800000
2015-01-01	62.063158
2015-01-05	58.900000
2015-01-06	69.216667
2015-01-07	82.877500
...	...
2017-12-04	128.755465
2017-12-05	124.544536
2017-12-06	132.725882
2017-12-07	130.473617
2017-12-08	142.949080

864 rows × 1 columns

Variation of Average daily rate in both City and Resort hotel throughout the years

In [22]:
```
plt.figure(figsize=(11,5))
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort_hotel')
plt.plot(city_hotel.index,city_hotel['adr'],label='city_hotel')
plt.xlabel('Date')
plt.ylabel('Average daily rate')
plt.title('Average daily rate in city and resort hotel')
plt.legend(fontsize=20)
```

Out[22]: <matplotlib.legend.Legend at 0x281c4fdcc50>



reservation_status_date are converted into months

In [23]:
```
data['month']=data['reservation_status_date'].dt.month
data
```

Out[23]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month
0	Resort Hotel	0	342	2015	July
1	Resort Hotel	0	737	2015	July
2	Resort Hotel	0	7	2015	July
3	Resort Hotel	0	13	2015	July
4	Resort Hotel	0	14	2015	July
...	...	...	...	...	...
119385	City Hotel	0	23	2017	August
119386	City Hotel	0	102	2017	August
119387	City Hotel	0	34	2017	August
119388	City Hotel	0	109	2017	August
119389	City Hotel	0	205	2017	August

119389 rows × 33 columns

In [24]:
```
data['month']=data['month'].astype(str)
data['is_canceled']=data['is_canceled'].astype(str)
data
```

Out[24]:

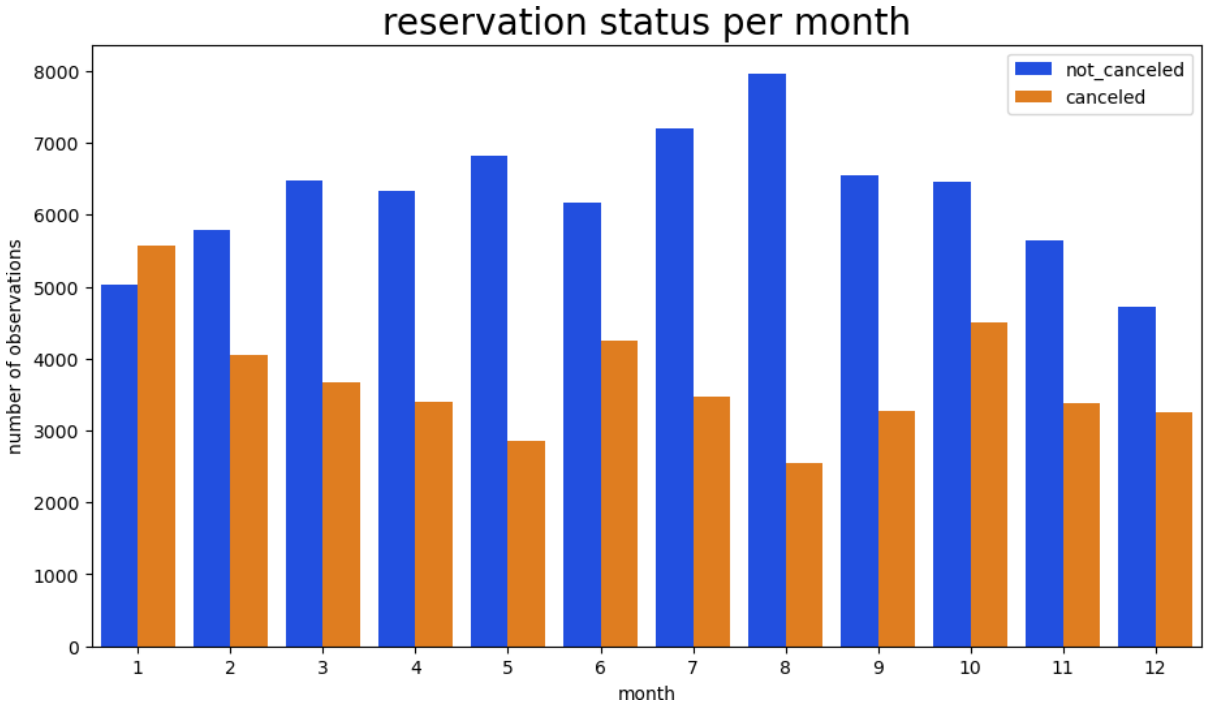
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month
0	Resort Hotel	0	342	2015	July
1	Resort Hotel	0	737	2015	July
2	Resort Hotel	0	7	2015	July
3	Resort Hotel	0	13	2015	July
4	Resort Hotel	0	14	2015	July
...	...	...	...	...	...

119385	City Hotel	0	23	2017	August
119386	City	0	102	2017	August
119387	City Hotel	0	34	2017	August
119388	City	0	109	2017	August
119389	City Hotel	0	205	2017	August

119389 rows × 33 columns

Variation of cancellation and not cancelled hotel bookings throughout the months

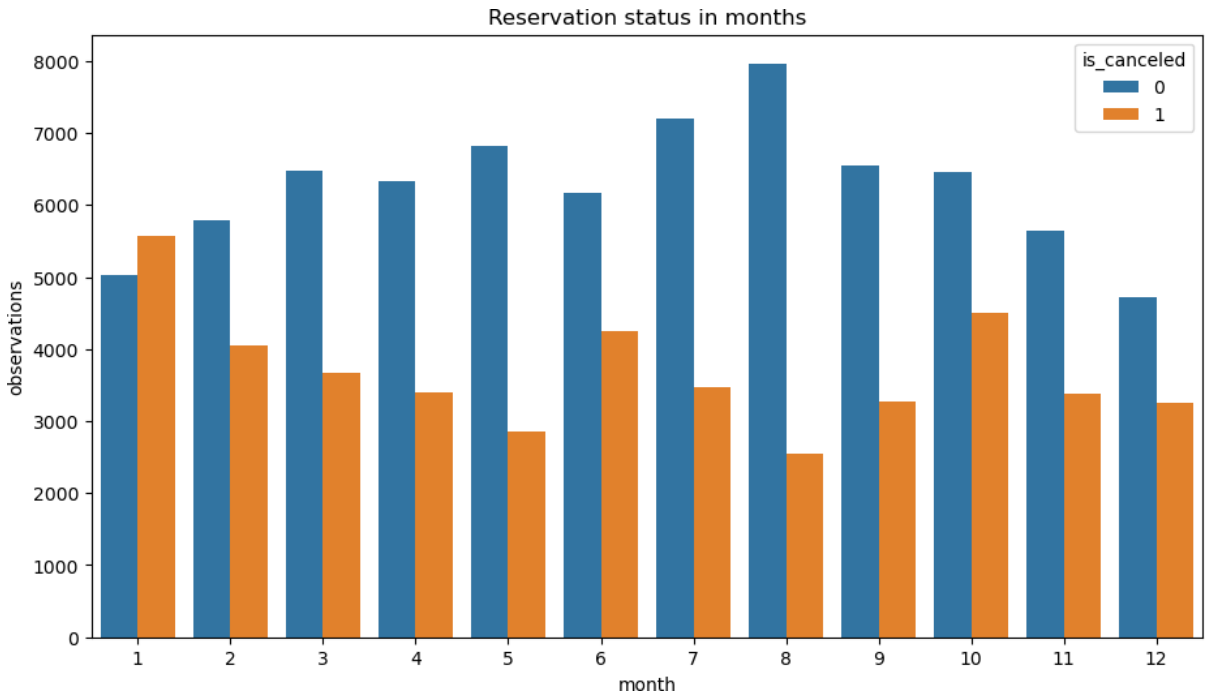
In [27]:
```
data['month']=data['reservation_status_date'].dt.month
plt.figure(figsize=(11,6))
ax1=sns.countplot(x='month',hue='is_canceled',data=data,palette='bright')
legend_labels,_ =ax1. get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('reservation status per month',size=20)
plt.xlabel('month')
plt.ylabel('number of observations')
plt.legend(['not_canceled','canceled'])
plt.show()
```



In [28]:
```
plt.figure(figsize=(11,6))
sns.countplot(data=data,x='month',hue='is_canceled')
plt.title('Reservation status in months')
```

```
plt.xlabel('month')
plt.ylabel('observations')
```

Out[28]: Text(0, 0.5, 'observations')



In [29]:
```
data[['is_canceled','adr']]
```

Out[29]:

	is_canceled	adr
0	0	0.00
1	0	0.00
2	0	75.00
3	0	75.00
4	0	98.00
...	...	...
119385	0	96.14
119386	0	225.43
119387	0	157.71
119388	0	104.40
119389	0	151.20

119389 rows × 2 columns

In [30]:
```
data['is_canceled']=list(map(int,data['is_canceled']))
data
```

Out[30]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month

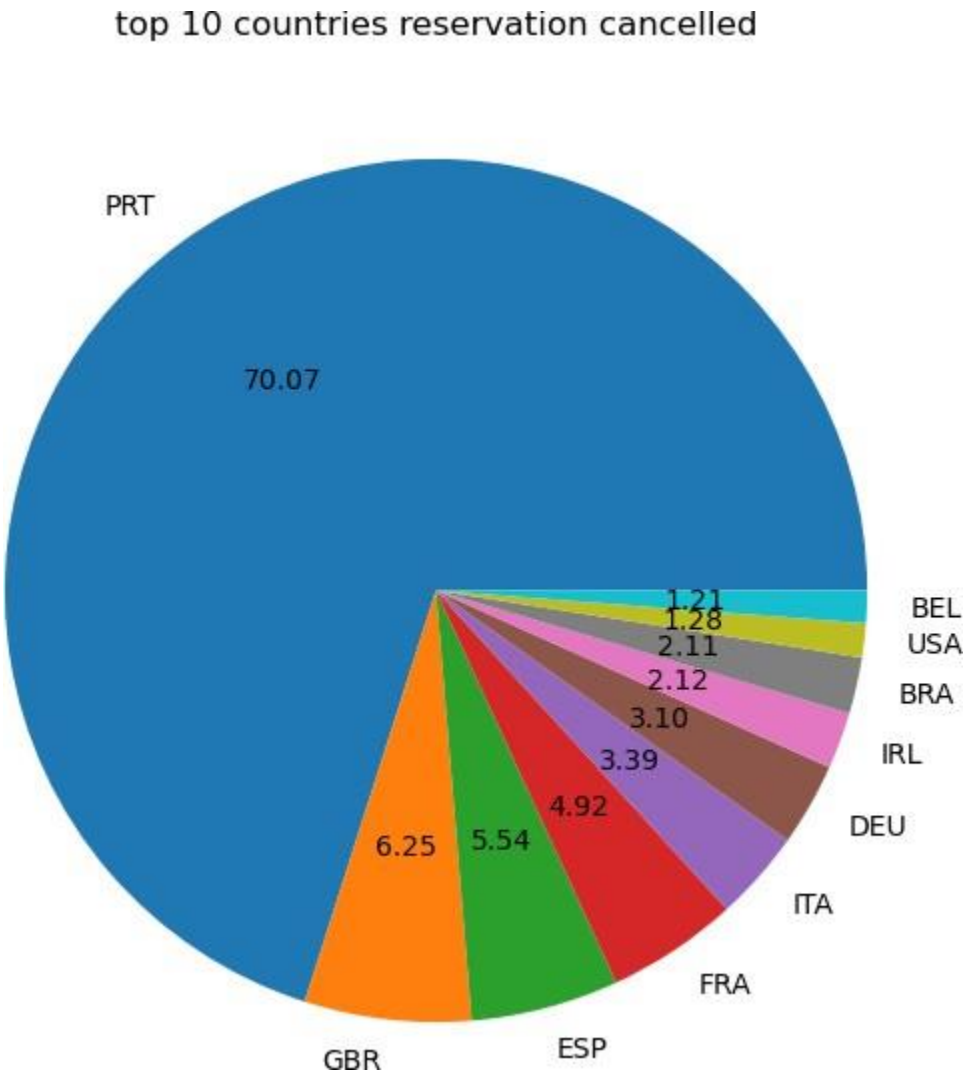
0	Resort Hotel	0	342	2015	July
1	Resort Hotel	0	737	2015	July
2	Resort Hotel	0	7	2015	July
3	Resort Hotel	0	13	2015	July
4	Resort Hotel	0	14	2015	July
...	...	...	...	...	...
119385	City Hotel	0	23	2017	August
119386	City Hotel	0	102	2017	August
119387	City Hotel	0	34	2017	August
119388	City Hotel	0	109	2017	August
119389	City Hotel	0	205	2017	August

119389 rows × 33 columns

Top 10 countries where the cancellation of hotel bookings mostly occur

In [31]:
```
new_data=data[data['is_canceled']==1]
top_10_country=new_data['country'].value_counts()[:10]
print(top_10_country)
plt.figure(figsize=(8,7))
plt.title('top 10 countries reservation cancelled')
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.show()
```

```
country
PRT     27518
GBR      2453
ESP      2177
FRA      1934
ITA      1333
DEU      1218
IRL       832
BRA       830
USA       501
BEL       474
Name: count, dtype: int64
```



Total unique categories in market_segment column of the dataset

In [32]:
```
data['market_segment'].value_counts()
```

Out[32]:
```
market_segment
Online TA         56477
Offline TA/TO     24218
Groups            19811
Direct            12606
Corporate          5295
Complementary       743
Aviation            237
Undefined             2
Name: count, dtype: int64
```

fraction of total unique categories in market_segment column of the dataset

In [33]:
```
data['market_segment'].value_counts(normalize=True)
```

Out[33]:
```
market_segment
Online TA         0.473050
Offline TA/TO     0.202850
Groups            0.165937
Direct            0.105588
Corporate         0.044351
Complementary     0.006223
Aviation          0.001985
Undefined         0.000017
Name: proportion, dtype: float64
```

In [34]:
```
new_data['market_segment'].value_counts(normalize=True)
```

Out[34]:
```
market_segment
Online TA         0.468964
Groups            0.273545
Offline TA/TO     0.187911
Direct            0.043733
Corporate         0.022432
Complementary     0.002193
Aviation          0.001176
Undefined         0.000045
Name: proportion, dtype: float64
```

Average Daily Rate of the both hotels (City and Resort) per day Where reservation_status cancelled

In [35]:
```
new_data_adr=new_data.groupby('reservation_status_date')[['adr']].mean()
new_data_adr.reset_index(inplace=True)
new_data_adr
```

Out[35]:

	reservation_status_date	adr
0	2014-10-17	62.800000
1	2014-11-18	0.000000
2	2015-01-01	62.062779
3	2015-01-05	96.542222
4	2015-01-06	103.926154
...	...	...
897	2017-12-04	148.121613
898	2017-12-05	118.205000
899	2017-12-06	178.939535
900	2017-12-07	173.704444
901	2017-12-08	198.000000

902 rows × 2 columns

In [36]:
```
new_data_adr.sort_values('reservation_status_date',inplace=True)
new_data_adr
```

Out[36]:

	reservation_status_date	adr
0	2014-10-17	62.800000
1	2014-11-18	0.000000
2	2015-01-01	62.062779
3	2015-01-05	96.542222
4	2015-01-06	103.926154
...	...	...
897	2017-12-04	148.121613
898	2017-12-05	118.205000
899	2017-12-06	178.939535
900	2017-12-07	173.704444
901	2017-12-08	198.000000

902 rows × 2 columns

Average Daily Rate of the both hotels (City and Resort) per day Where reservation_status not cancelled

In [37]:
```
not_new_data=data[data['is_canceled']==0]
not_new_data_adr=not_new_data.groupby('reservation_status_date')[['adr']].mean()
not_new_data_adr.reset_index(inplace=True)
not_new_data_adr.sort_values('reservation_status_date',inplace=True)
not_new_data_adr
```

Out[37]:

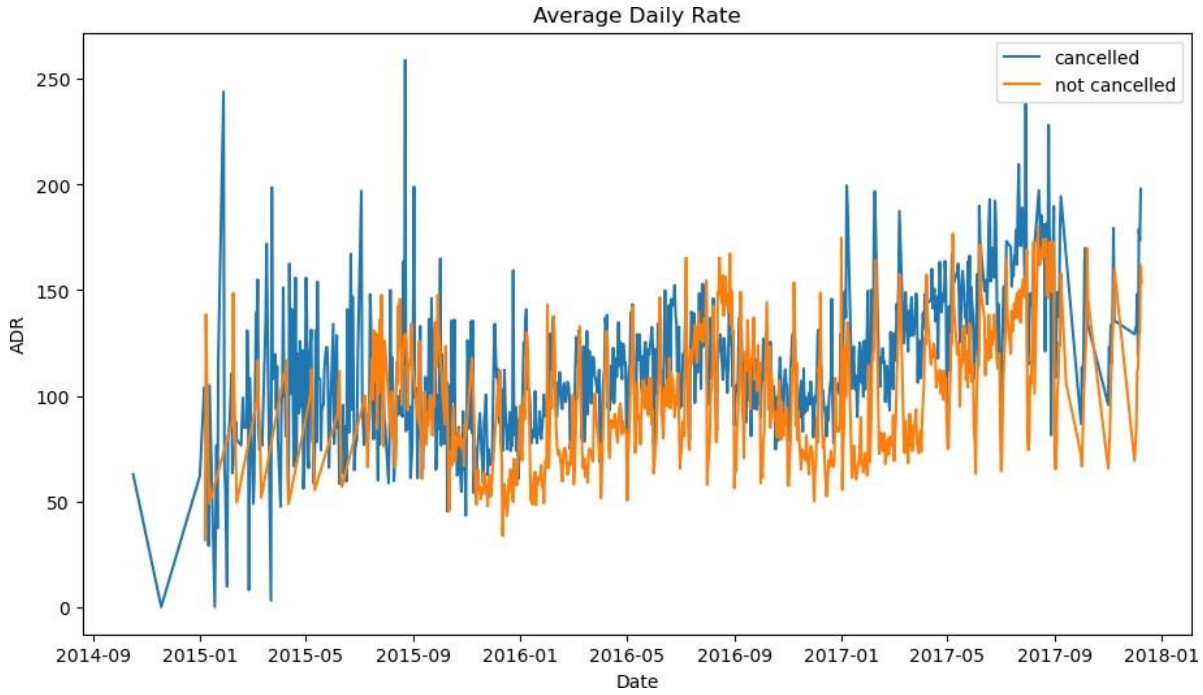
	reservation_status_date	adr
0	2015-01-07	31.650000
1	2015-01-08	138.444565
2	2015-01-09	110.008657
3	2015-01-10	86.723818
4	2015-01-11	59.431300
...	...	...
800	2017-12-05	113.367857
801	2017-12-06	127.924490
802	2017-12-07	130.153945
803	2017-12-08	161.916864

804	2017-12-09	153.570000

805 rows × 2 columns

Variation of Average daily rate in both City and Resort hotel throughout the years where cancellation and not cancelled of bookings occurs

In [38]:
```
plt.figure(figsize=(11,6))
plt.plot(new_data_adr['reservation_status_date'],new_data_adr['adr'],label='cancell
plt.plot(not_new_data_adr['reservation_status_date'],not_new_data_adr['adr'],label=
plt.legend()
plt.title('Average Daily Rate')
plt.xlabel('Date')
plt.ylabel('ADR')
plt.show()
```



In [ ]:

COMPREHENSIVE EXECUTIVE SUMMARY

The project involves an **Exploratory Data Analysis (EDA)** of hotel booking demand, focusing on cleaning, analyzing, and visualizing a dataset of 119,390 entries from various hotels. The key elements of the analysis are as follows:

1. **Data Cleaning**: The dataset was first cleaned to handle missing values and incorrect data types. Various features like hotel, is_canceled, lead_time, arrival_date_year, and others were explored and visualized to understand trends.
2. **Exploratory Analysis**:
    - Cancellations were analyzed, showing that 37% of bookings were canceled, with a deeper dive into the differences between City and Resort hotels. Resort hotels had a lower cancellation rate (27%) compared to City hotels (41%).
    - Trends such as the variation of cancellations across different months were also visualized.
    - Average Daily Rate (ADR) trends were studied separately for canceled and non-canceled bookings for both City and Resort hotels.
3. **Country-wise cancellations**: The analysis highlighted the top 10 countries with the most cancellations, with Portugal leading the list.
4. **Market Segment Analysis**: The study examined the market segments, revealing that the majority of bookings (47%) came through online travel agencies (OTA).

Conclusion:

The project successfully identified key patterns in hotel booking demand, such as cancellation rates, ADR variations, and market segmentation. Resort hotels generally performed better in terms of lower cancellations and higher ADRs. Cancellations were higher in City hotels, with Online TA being the major source of reservations and cancellations.

Suggestions:

- **Focus on Direct Bookings**: Since online platforms contributed significantly to cancellations, hotels could incentivize direct bookings to reduce dependency on intermediaries.
- **Improved Forecasting**: By identifying peak cancellation periods, hotels can adjust pricing strategies during high-risk months.
- **Targeted Marketing**: Given the cancellation rates from specific countries, hotels can optimize their marketing efforts to improve booking stability in those regions.

THANK YOU!!

AYUSH PATEL