IMPORT LIBRARIES

In [45]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

IMPORTING CSV FILES

In [2]:

```
df2=pd.read_csv(r'C:\Users\manpr\Downloads\archive\hotel_booking.csv',encoding='uni
```

DATA CLEANING

In [3]:

```
df2.head()
```

Out[3]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_nu
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	

5 rows × 32 columns

In [4]:

```
df2.tail()
```

Out[4]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_wee
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	

5 rows × 32 columns

In [6]:

```
df2.shape
```

Out[6]:

```
(119390, 32)
```

In [7]:

```
df2.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 [8]:

```
df2.columns
```

Out[8]:

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [9]:

```
df2.isnull().sum()
```

Out[9]:

```
hotel                                  0
is_canceled                            0
lead_time                              0
arrival_date_year                      0
arrival_date_month                     0
arrival_date_week_number               0
arrival_date_day_of_month              0
stays_in_weekend_nights                0
stays_in_week_nights                   0
adults                                 0
children                               4
babies                                 0
meal                                   0
country                              488
market_segment                         0
distribution_channel                   0
is_repeated_guest                      0
previous_cancellations                 0
previous_bookings_not_canceled         0
reserved_room_type                     0
assigned_room_type                     0
booking_changes                        0
deposit_type                           0
agent                              16340
company                           112593
days_in_waiting_list                   0
customer_type                          0
adr                                    0
required_car_parking_spaces            0
total_of_special_requests              0
reservation_status                     0
reservation_status_date                0
dtype: int64
```

In [10]:

```
df2.drop(['agent','company'],axis=1,inplace=True)
```

In [11]:

```
df2.dropna(inplace=True)
```

In [12]:

```
df2.isnull().sum()
```

Out[12]:

```
hotel                             0
is_canceled                       0
lead_time                         0
arrival_date_year                 0
arrival_date_month                0
arrival_date_week_number          0
arrival_date_day_of_month         0
stays_in_weekend_nights           0
stays_in_week_nights              0
adults                            0
children                          0
babies                            0
meal                              0
country                           0
market_segment                    0
distribution_channel              0
is_repeated_guest                 0
previous_cancellations            0
previous_bookings_not_canceled    0
reserved_room_type                0
assigned_room_type                0
booking_changes                   0
deposit_type                      0
days_in_waiting_list              0
customer_type                     0
adr                               0
required_car_parking_spaces       0
total_of_special_requests         0
reservation_status                0
reservation_status_date           0
dtype: int64
```

In [74]:

```
#change data type
df2['reservation_status_date']=pd.to_datetime(df2['reservation_status_date'])
```

In [75]:

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 118897 entries, 0 to 119389
Data columns (total 30 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  int64
 2   lead_time                       118897 non-null  int64
 3   arrival_date_year               118897 non-null  int64
 4   arrival_date_month              118897 non-null  object
 5   arrival_date_week_number        118897 non-null  int64
 6   arrival_date_day_of_month       118897 non-null  int64
 7   stays_in_weekend_nights         118897 non-null  int64
 8   stays_in_week_nights            118897 non-null  int64
 9   adults                          118897 non-null  int64
 10  children                        118897 non-null  datetime64[ns]
 11  babies                          118897 non-null  int64
 12  meal                            118897 non-null  object
 13  country                         118897 non-null  object
 14  market_segment                  118897 non-null  object
 15  distribution_channel            118897 non-null  object
 16  is_repeated_guest               118897 non-null  int64
 17  previous_cancellations          118897 non-null  int64
 18  previous_bookings_not_canceled  118897 non-null  int64
 19  reserved_room_type              118897 non-null  object
 20  assigned_room_type              118897 non-null  object
 21  booking_changes                 118897 non-null  int64
 22  deposit_type                    118897 non-null  object
 23  days_in_waiting_list            118897 non-null  int64
 24  customer_type                   118897 non-null  object
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64
 27  total_of_special_requests       118897 non-null  int64
 28  reservation_status              118897 non-null  object
 29  reservation_status_date         118897 non-null  datetime64[ns]
dtypes: datetime64[ns](2), float64(1), int64(16), object(11)
memory usage: 28.1+ MB
```

In [76]:

```
df2.describe(include='object')
```

Out[76]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel
count	118897	118897	118897	118897	118897	118897
unique	2	12	5	177	7	5
top	City Hotel	August	BB	PRT	Online TA	TA/TO
freq	79301	13852	91862	48585	56402	97729

In [77]:

```
df2.describe()
```

Out[77]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_da
count	118897.000000	118897.000000	118897.000000	118897.000000	
mean	0.371347	104.312018	2016.157657	27.166674	
std	0.483167	106.903570	0.707462	13.589966	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	161.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

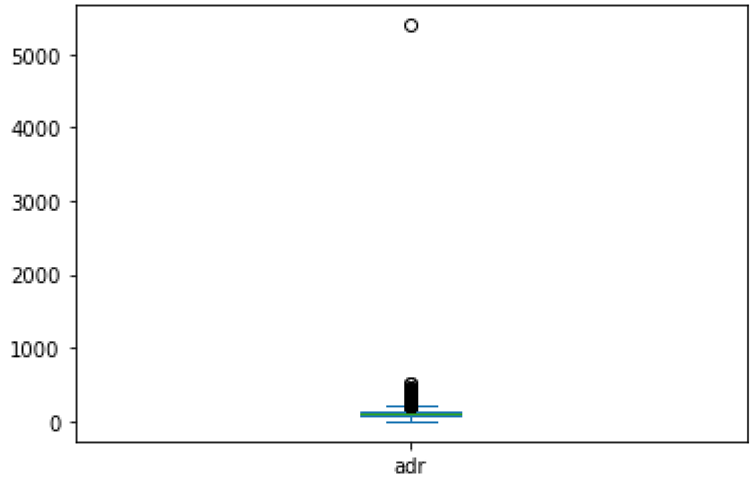
Will remove the values which are too beyond the range for all the columns.Here,the column 'adr' has maximum outlier.

In [15]:

```
#Make a box plot
df2['adr'].plot(kind='box')
```

Out[15]:

```
<AxesSubplot:>
```



In [25]:

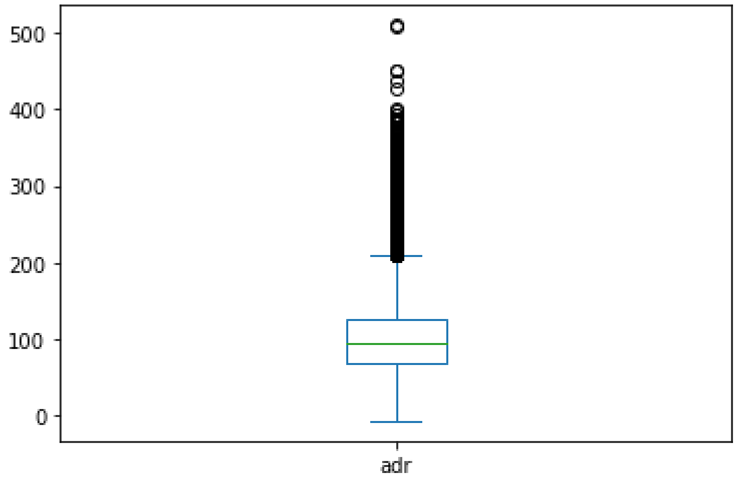
```
df2=df2[df2['adr']<5000]
```

In [27]:

```
df2['adr'].plot(kind='box')
```

Out[27]:

```
<AxesSubplot:>
```



df2.describe()

DATA ANALYSIS AND VISUALISATIONS

In [28]:

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 118897 entries, 0 to 119389
Data columns (total 30 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           118897 non-null  object 
 1   is_canceled                     118897 non-null  int64  
 2   lead_time                       118897 non-null  int64  
 3   arrival_date_year               118897 non-null  int64  
 4   arrival_date_month              118897 non-null  object 
 5   arrival_date_week_number        118897 non-null  int64  
 6   arrival_date_day_of_month       118897 non-null  int64  
 7   stays_in_weekend_nights         118897 non-null  int64  
 8   stays_in_week_nights            118897 non-null  int64  
 9   adults                          118897 non-null  int64  
 10  children                        118897 non-null  float64
 11  babies                          118897 non-null  int64  
 12  meal                            118897 non-null  object 
 13  country                         118897 non-null  object 
 14  market_segment                  118897 non-null  object 
 15  distribution_channel            118897 non-null  object 
 16  is_repeated_guest               118897 non-null  int64  
 17  previous_cancellations          118897 non-null  int64  
 18  previous_bookings_not_canceled  118897 non-null  int64  
 19  reserved_room_type              118897 non-null  object 
 20  assigned_room_type              118897 non-null  object 
 21  booking_changes                 118897 non-null  int64  
 22  deposit_type                    118897 non-null  object 
 23  days_in_waiting_list            118897 non-null  int64  
 24  customer_type                   118897 non-null  object 
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64  
 27  total_of_special_requests       118897 non-null  int64  
 28  reservation_status              118897 non-null  object 
 29  reservation_status_date         118897 non-null  object 
dtypes: float64(2), int64(16), object(12)
memory usage: 28.1+ MB
```

Canceled percentage

In [51]:

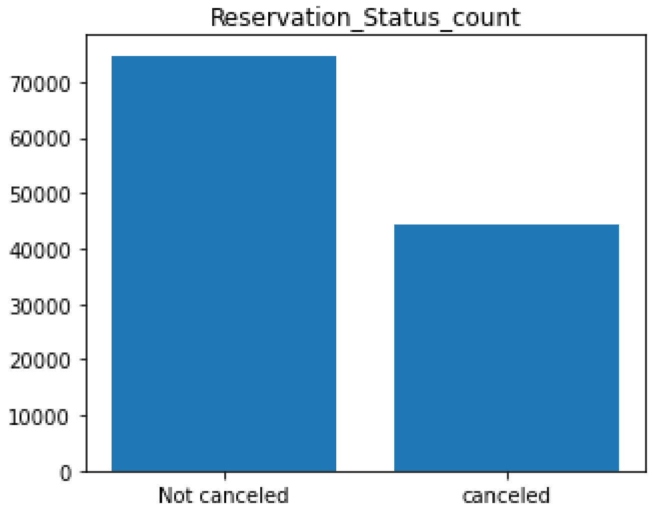
```
cancelled_perc=df2['is_canceled'].value_counts(normalize=True)
print(cancelled_perc*100)

plt.figure(figsize=(5,4))
plt.title('Reservation_Status_count')
plt.bar(['Not canceled','canceled'],df2['is_canceled'].value_counts())
```

```
0    62.865337
1    37.134663
Name: is_canceled, dtype: float64
```

Out[51]:

```
<BarContainer object of 2 artists>
```



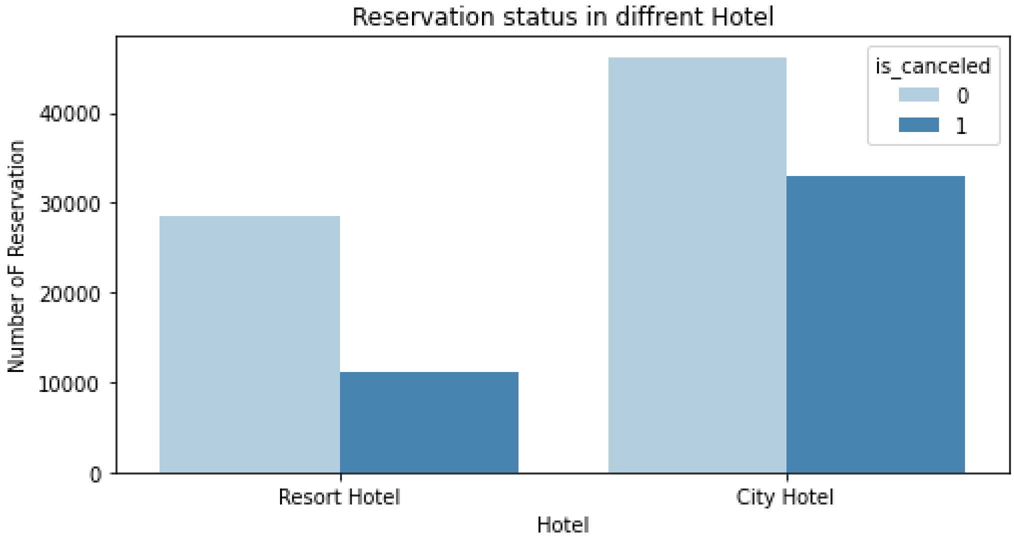
Cancelation Rate per Hotel

In [53]:

```
plt.figure(figsize=(8,4))
ax=sns.countplot(x='hotel',data=df2,palette='Blues',hue='is_canceled')
plt.title("Reservation status in diffrent Hotel")
plt.xlabel('Hotel')
plt.ylabel('Number oF Reservation')
```

Out[53]:

```
Text(0, 0.5, 'Number oF Reservation')
```



In [80]:

```
#Cancellation rate for Resort_Hotel
resort_hotel_canceled=df2[df2['hotel']=='Resort Hotel']
resort_hotel_canceled['is_canceled'].value_counts(normalize=True)
```

Out[80]:

```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

In [81]:

```
#Cancellation rate for City_Hotel
city_hotel_canceled=df2[df2['hotel']=='City Hotel']
city_hotel_canceled['is_canceled'].value_counts(normalize=True)
```

Out[81]:

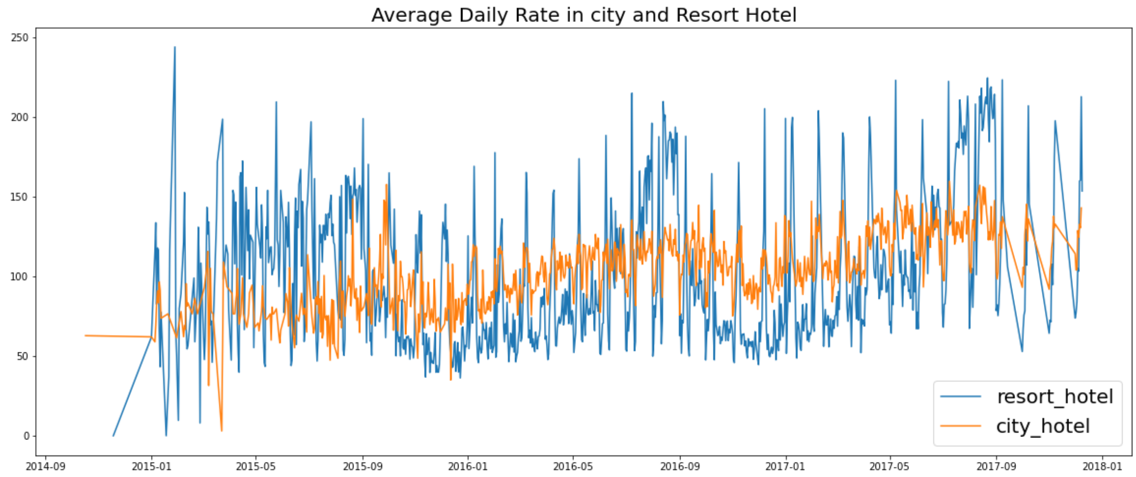
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [83]:

```
resort_hotel=resort_hotel_canceled.groupby('reservation_status_date')[['adr']].mean
city_hotel=city_hotel_canceled.groupby('reservation_status_date')[['adr']].mean()
```

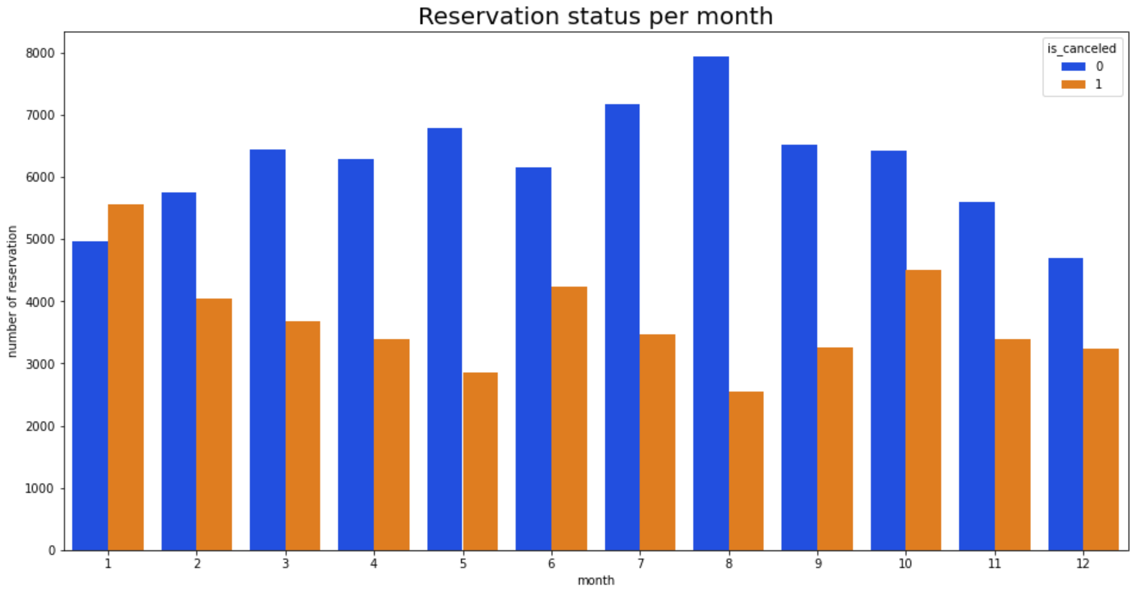
In [92]:

```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate in city and Resort Hotel',size=20)
plt.plot(resort_hotel.index,resort_hotel['adr'],label='resort_hotel')
plt.plot(city_hotel.index,city_hotel['adr'],label='city_hotel')
plt.legend(fontsize=20)
plt.show()
```



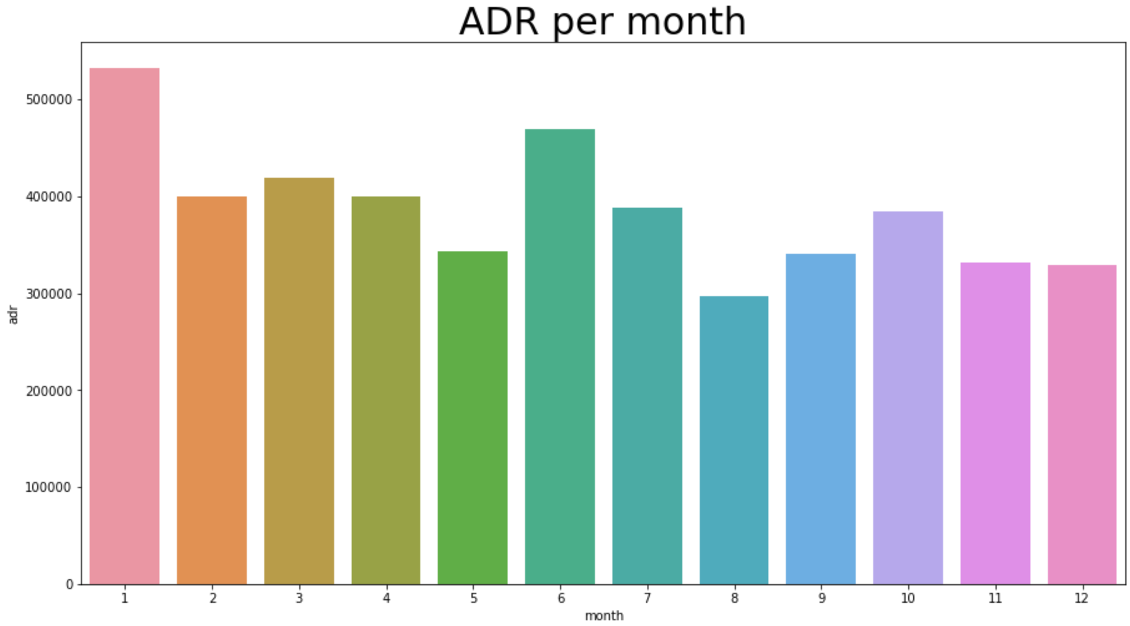
In [95]:

```
df2['month']=df2['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
bx=sns.countplot(x='month',hue='is_canceled',palette='bright',data=df2)
plt.title('Reservation status per month',size=20)
plt.xlabel('month')
plt.ylabel('number of reservation')
plt.show()
```



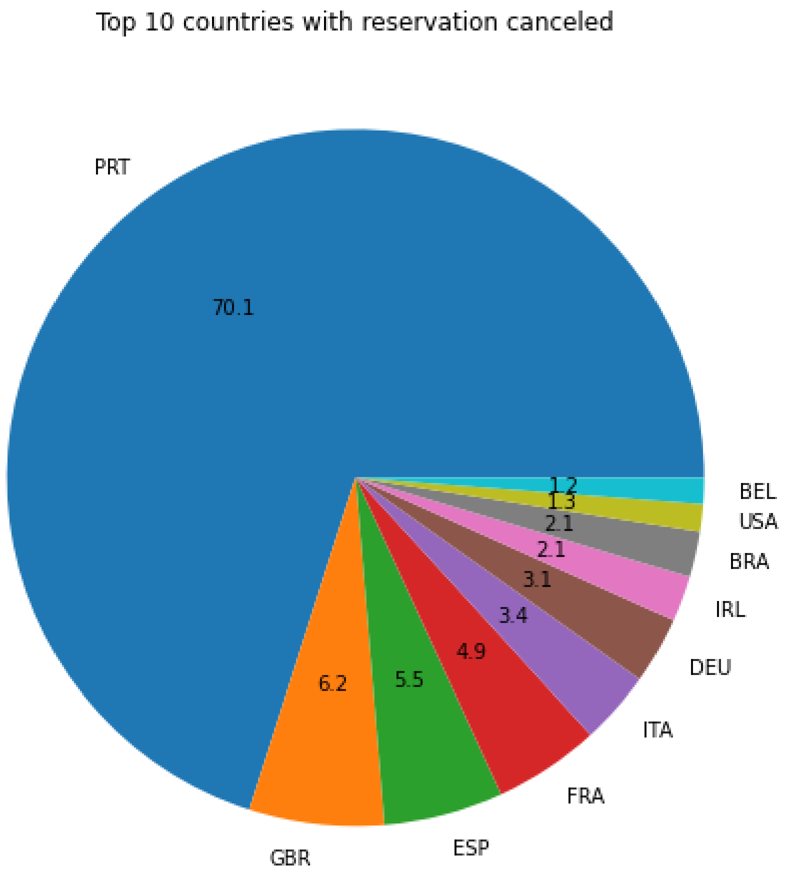
In [97]:

```
cancelled_ADR=df2[df2['is_canceled']==1].groupby('month')[['adr']].sum().reset_inde
plt.figure(figsize=(15,8))
plt.title('ADR per month',size=30)
sns.barplot(x='month',y='adr',data=cancelled_ADR)
plt.show()
```



In [124]:

```
cancelled_data=df2[df2['is_canceled']==1]
top_10_countries=cancelled_data['country'].value_counts()[ :10]
plt.figure(figsize=(8,8))
plt.title('Top 10 countries with reservation canceled')
plt.pie(top_10_countries,labels=top_10_countries.index,autopct='%.1f')
plt.show()
```



THANKS

In [ ]:

```

```