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 [3]:

```
df=pd.read_csv(r"C:\Users\Jai Shree Shyam\Desktop\Python Project\hotel_booking.csv'
```

Explore and Cleaning the data

In [4]:

```
df.head()
```

Out[4]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
0	Resort Hotel	0	342	2015	July	27
1	Resort Hotel	0	737	2015	July	27
2	Resort Hotel	0	7	2015	July	27
3	Resort Hotel	0	13	2015	July	27
4	Resort Hotel	0	14	2015	July	27

5 rows × 36 columns

In [5]:

```
df.tail()
```

Out[5]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num
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 × 36 columns

In [12]:

```
df.shape
```

Out[12]:

```
(119390, 36)
```

In [13]:

```
df.columns
```

Out[13]:

```
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', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

In [14]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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 
 32  name                            119390 non-null  object 
 33  email                           119390 non-null  object 
 34  phone-number                    119390 non-null  object 
 35  credit_card                     119390 non-null  object 
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

In [15]:

```
## Here, reservation_status_date is object type, have to convert into date format f

df["reservation_status_date"]= pd.to_datetime(df["reservation_status_date"])
```

In [19]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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]
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
dtypes: datetime64[ns](1), float64(4), int64(16), object(15)
memory usage: 32.8+ MB
```

In [21]:

```
df.describe().T
```

Out[21]:

	count	mean	std	min	25%	50%	7
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	16
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	201
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	3
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	2
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	22
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	27
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	12
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	

In [23]:

```
# Checking distribution of categorical columns
df.describe(include=object).T
```

Out[23]:

	count	unique	top	freq
hotel	119390	2	City Hotel	79330
arrival_date_month	119390	12	August	13877
meal	119390	5	BB	92310
country	118902	177	PRT	48590
market_segment	119390	8	Online TA	56477
distribution_channel	119390	5	TA/TO	97870
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
customer_type	119390	4	Transient	89613
reservation_status	119390	3	Check-Out	75166
name	119390	81503	Michael Johnson	48
email	119390	115889	Michael.C@gmail.com	6
phone-number	119390	119390	669-792-1661	1
credit_card	119390	9000	************4923	28

In [24]:

```
# Fetching all categorical columns with all unique values.

for col in df.describe(include="object").columns:
    print(col)
    print(df[col].unique())
    print('--'*60)
```

```
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']
----------------------------------------------------------------------------------
--------------------------------------
name
['Ernest Barnes' 'Andrea Baker' 'Rebecca Parker' ... 'Wesley Aguilar'
 'Caroline Conley MD' 'Ariana Michael']
----------------------------------------------------------------------------------
```

```
--------------------------------------
email
['Ernest.Barnes31@outlook.com' 'Andrea_Baker94@aol.com'
 'Rebecca_Parker@comcast.net' ... 'Mary_Morales@hotmail.com'
 'MD_Caroline@comcast.net' 'Ariana_M@xfinity.com']
----------------------------------------------------------------------------------
--------------------------------------
phone-number
['669-792-1661' '858-637-6955' '652-885-2745' ... '395-518-4100'
 '531-528-1017' '422-804-6403']
----------------------------------------------------------------------------------
--------------------------------------
credit_card
['************4322' '************9157' '************3734' ...
 '************9170' '************6349' '************7959']
----------------------------------------------------------------------------------
--------------------------------------
```

In [25]:

```python
df.isna().sum()
```

Out[25]:

```
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
name                                    0
email                                   0
phone-number                            0
credit_card                             0
dtype: int64
```

In [26]:

```python
# Here we drop agent column which is not required for analysis
# And company column has almost null equvivalent to no_of_row so it also not requir
```

In [27]:

```python
df.drop(["company","agent"],axis=1,inplace=True)
df.dropna(inplace=True)
```

In [28]:
```
# Further more customer_name, email and phone-number, credit_card columns are not r

df.drop(["name","email","phone-number","credit_card"],axis=1,inplace=True)
```

In [29]:
```
df.head()
```

Out[29]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
0	Resort Hotel	0	342	2015	July	27
1	Resort Hotel	0	737	2015	July	27
2	Resort Hotel	0	7	2015	July	27
3	Resort Hotel	0	13	2015	July	27
4	Resort Hotel	0	14	2015	July	27

5 rows × 30 columns

In [32]:
```
# column country and children having some missing values so, we drop it.

df.dropna(inplace=True)
```

In [33]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 118898 entries, 0 to 119389
Data columns (total 30 columns):
 #   Column                          Non-Null Count   Dtype         
---  ------                          --------------   -----         
 0   hotel                           118898 non-null  object        
 1   is_canceled                     118898 non-null  int64         
 2   lead_time                       118898 non-null  int64         
 3   arrival_date_year               118898 non-null  int64         
 4   arrival_date_month              118898 non-null  object        
 5   arrival_date_week_number        118898 non-null  int64         
 6   arrival_date_day_of_month       118898 non-null  int64         
 7   stays_in_weekend_nights         118898 non-null  int64         
 8   stays_in_week_nights            118898 non-null  int64         
 9   adults                          118898 non-null  int64         
 10  children                        118898 non-null  float64       
 11  babies                          118898 non-null  int64         
 12  meal                            118898 non-null  object        
 13  country                         118898 non-null  object        
 14  market_segment                  118898 non-null  object        
 15  distribution_channel            118898 non-null  object        
 16  is_repeated_guest               118898 non-null  int64         
 17  previous_cancellations          118898 non-null  int64         
 18  previous_bookings_not_canceled  118898 non-null  int64         
 19  reserved_room_type              118898 non-null  object        
 20  assigned_room_type              118898 non-null  object        
 21  booking_changes                 118898 non-null  int64         
 22  deposit_type                    118898 non-null  object        
 23  days_in_waiting_list            118898 non-null  int64         
 24  customer_type                   118898 non-null  object        
 25  adr                             118898 non-null  float64       
 26  required_car_parking_spaces     118898 non-null  int64         
 27  total_of_special_requests       118898 non-null  int64         
 28  reservation_status              118898 non-null  object        
 29  reservation_status_date         118898 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(16), object(11)
memory usage: 28.1+ MB
```

In [34]:

```
df.describe().T
```

Out[34]:

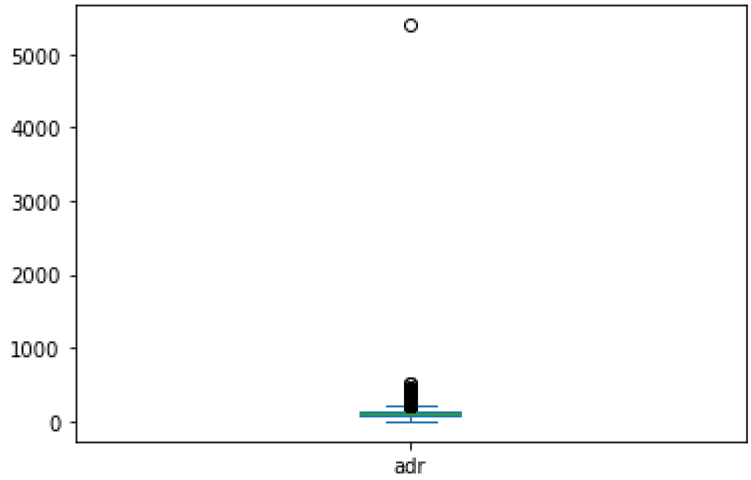
	count	mean	std	min	25%	50%	75%
is_canceled	118898.0	0.371352	0.483168	0.00	0.0	0.0	1.0
lead_time	118898.0	104.311435	106.903309	0.00	18.0	69.0	161.0
arrival_date_year	118898.0	2016.157656	0.707459	2015.00	2016.0	2016.0	2017.0
arrival_date_week_number	118898.0	27.166555	13.589971	1.00	16.0	28.0	38.0
arrival_date_day_of_month	118898.0	15.800880	8.780324	1.00	8.0	16.0	23.0
stays_in_weekend_nights	118898.0	0.928897	0.996216	0.00	0.0	1.0	2.0
stays_in_week_nights	118898.0	2.502145	1.900168	0.00	1.0	2.0	3.0
adults	118898.0	1.858391	0.578576	0.00	2.0	2.0	2.0
children	118898.0	0.104207	0.399172	0.00	0.0	0.0	0.0
babies	118898.0	0.007948	0.097380	0.00	0.0	0.0	0.0
is_repeated_guest	118898.0	0.032011	0.176029	0.00	0.0	0.0	0.0
previous_cancellations	118898.0	0.087142	0.845869	0.00	0.0	0.0	0.0
previous_bookings_not_canceled	118898.0	0.131634	1.484672	0.00	0.0	0.0	0.0
booking_changes	118898.0	0.221181	0.652785	0.00	0.0	0.0	0.0
days_in_waiting_list	118898.0	2.330754	17.630452	0.00	0.0	0.0	0.0
adr	118898.0	102.003243	50.485862	-6.38	70.0	95.0	126.0
required_car_parking_spaces	118898.0	0.061885	0.244172	0.00	0.0	0.0	0.0
total_of_special_requests	118898.0	0.571683	0.792678	0.00	0.0	0.0	1.0

In [36]:

```
# Here adr [average_daily_rate] column has outlier

df["adr"].plot(kind="box")
```

Out[36]:

```
<AxesSubplot:>
```



In [37]:

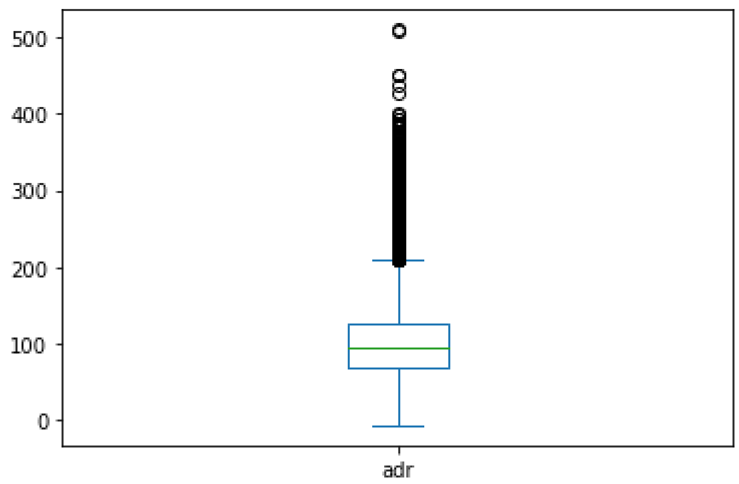
```
# droping outlier
df=df[df["adr"]<5000]
```

In [38]:

```
df["adr"].plot(kind="box")
```

Out[38]:

```
<AxesSubplot:>
```



In [39]:

```python
df.describe().T
```

Out[39]:

	count	mean	std	min	25%	50%	75%
is_canceled	118897.0	0.371347	0.483167	0.00	0.0	0.0	1.0
lead_time	118897.0	104.312018	106.903570	0.00	18.0	69.0	161.0
arrival_date_year	118897.0	2016.157657	0.707462	2015.00	2016.0	2016.0	2017.0
arrival_date_week_number	118897.0	27.166674	13.589966	1.00	16.0	28.0	38.0
arrival_date_day_of_month	118897.0	15.800802	8.780321	1.00	8.0	16.0	23.0
stays_in_weekend_nights	118897.0	0.928905	0.996217	0.00	0.0	1.0	2.0
stays_in_week_nights	118897.0	2.502157	1.900171	0.00	1.0	2.0	3.0
adults	118897.0	1.858390	0.578578	0.00	2.0	2.0	2.0
children	118897.0	0.104208	0.399174	0.00	0.0	0.0	0.0
babies	118897.0	0.007948	0.097381	0.00	0.0	0.0	0.0
is_repeated_guest	118897.0	0.032011	0.176030	0.00	0.0	0.0	0.0
previous_cancellations	118897.0	0.087143	0.845872	0.00	0.0	0.0	0.0
previous_bookings_not_canceled	118897.0	0.131635	1.484678	0.00	0.0	0.0	0.0
booking_changes	118897.0	0.221175	0.652784	0.00	0.0	0.0	0.0
days_in_waiting_list	118897.0	2.330774	17.630525	0.00	0.0	0.0	0.0
adr	118897.0	101.958683	48.091199	-6.38	70.0	95.0	126.0
required_car_parking_spaces	118897.0	0.061885	0.244173	0.00	0.0	0.0	0.0
total_of_special_requests	118897.0	0.571688	0.792680	0.00	0.0	0.0	1.0

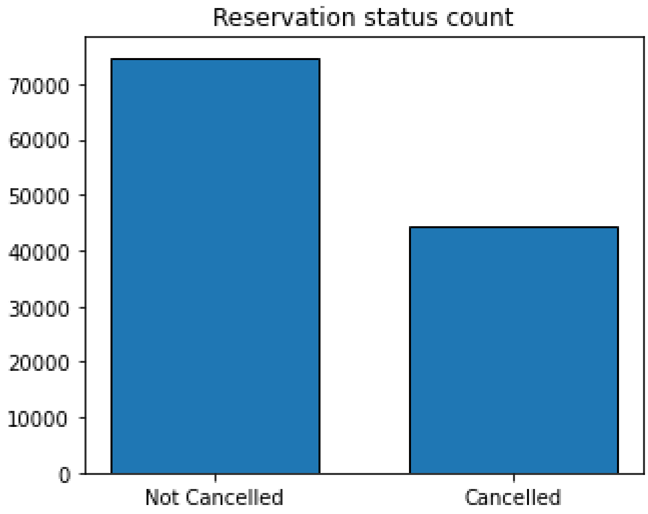
Data Analysis and Visualizations

In [49]:

```python
# Checking the cancelled percentage

cancelled_perc= np.round((df["is_canceled"].value_counts()/df["is_canceled"].count
print(cancelled_perc)
```

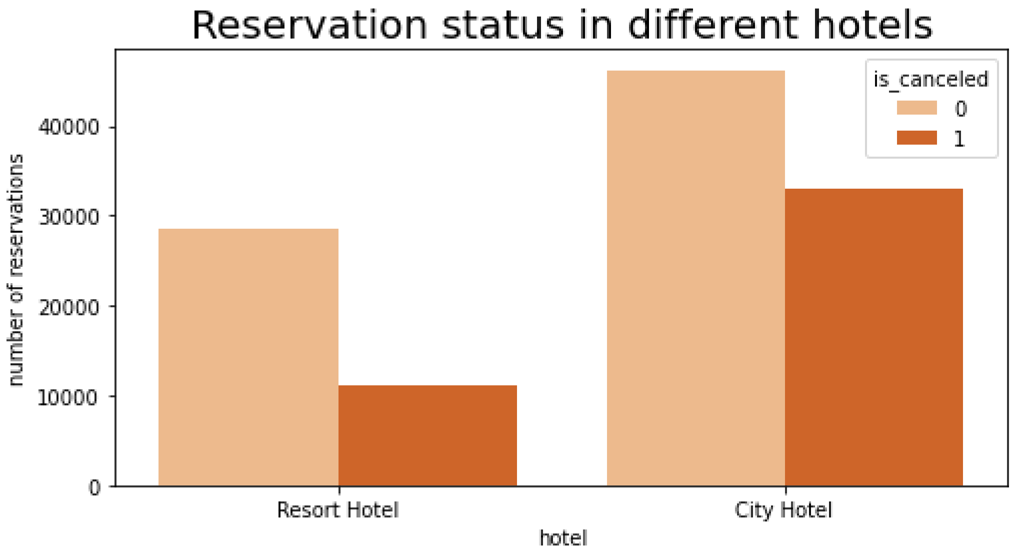
```
### 1- Cancellation     and   0- Not Canceled

## Here cancelation percentage is around 37% which is quite more and not managable
## then it will manageable.

plt.figure(figsize=(5,4))
plt.title("Reservation status count")
plt.bar(["Not Cancelled","Cancelled"],df["is_canceled"].value_counts(),edgecolor= '
plt.show()
```

```
0    62.87
1    37.13
Name: is_canceled, dtype: float64
```



Cancellation percentage based on hotel

In [53]:
```
plt.figure(figsize=(8,4))
ax1=sns.countplot(x="hotel",hue="is_canceled",data=df,palette='Oranges')

plt.title("Reservation status in different hotels",size=20)
plt.xlabel("hotel")
plt.ylabel("number of reservations")
plt.show()
```



In [57]:
```
# cancellation percentage for resort hotel

resort_hotel= df[df["hotel"]== "Resort Hotel"]
```

```
cancel_perc= np.round((resort_hotel["is_canceled"].value_counts(normalize=True))*1
cancel_perc
```

Out[57]:

```
0    72.02
1    27.98
Name: is_canceled, dtype: float64
```

In [59]:

```
# cancellation percentage for City Hotel

city_hotel=df[df["hotel"]=="City Hotel"]
cancel_perc= np.round((city_hotel["is_canceled"].value_counts(normalize=True))*100
cancel_perc
```

Out[59]:

```
0    58.0
1    42.0
Name: is_canceled, dtype: float64
```

In [60]:

```
## City Hotel having high cancel percent which is 42 %  comparision to Resort Hotel
```

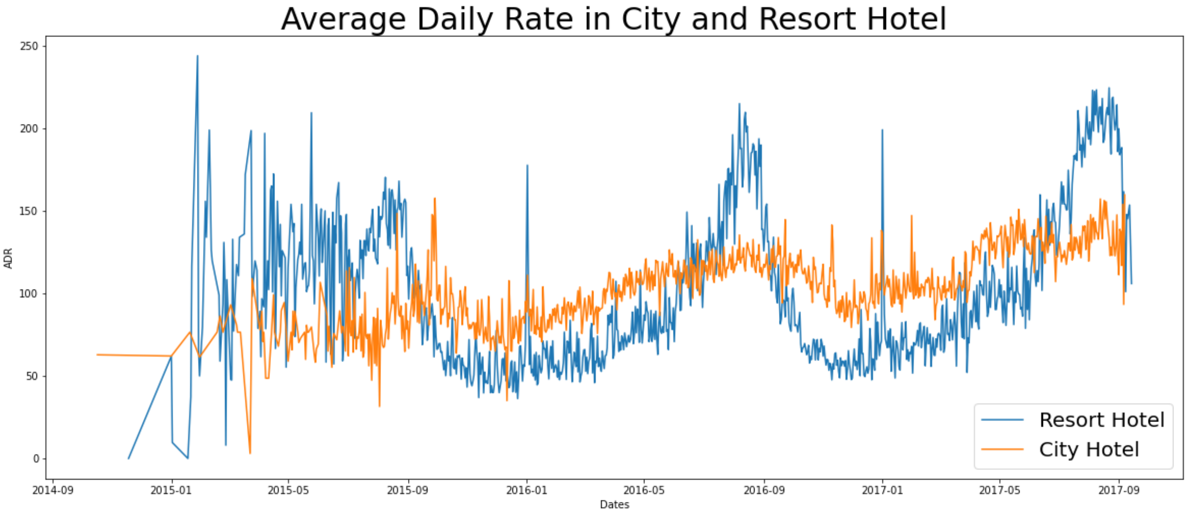
Checking the Price Effect on Cancellation

In [62]:

```
resort_hotel=resort_hotel.groupby("reservation_status_date")[["adr"]].mean()
city_hotel= city_hotel.groupby("reservation_status_date")[["adr"]].mean()
```

In [72]:

```
plt.figure(figsize=(20,8))
plt.title("Average Daily Rate in City and Resort Hotel", fontsize=30)
plt.plot(resort_hotel.index,resort_hotel["adr"],label="Resort Hotel")
plt.plot(city_hotel.index,city_hotel["adr"],label="City Hotel")
plt.xlabel("Dates")
plt.ylabel("ADR")
plt.legend(fontsize=20)
plt.show()
```

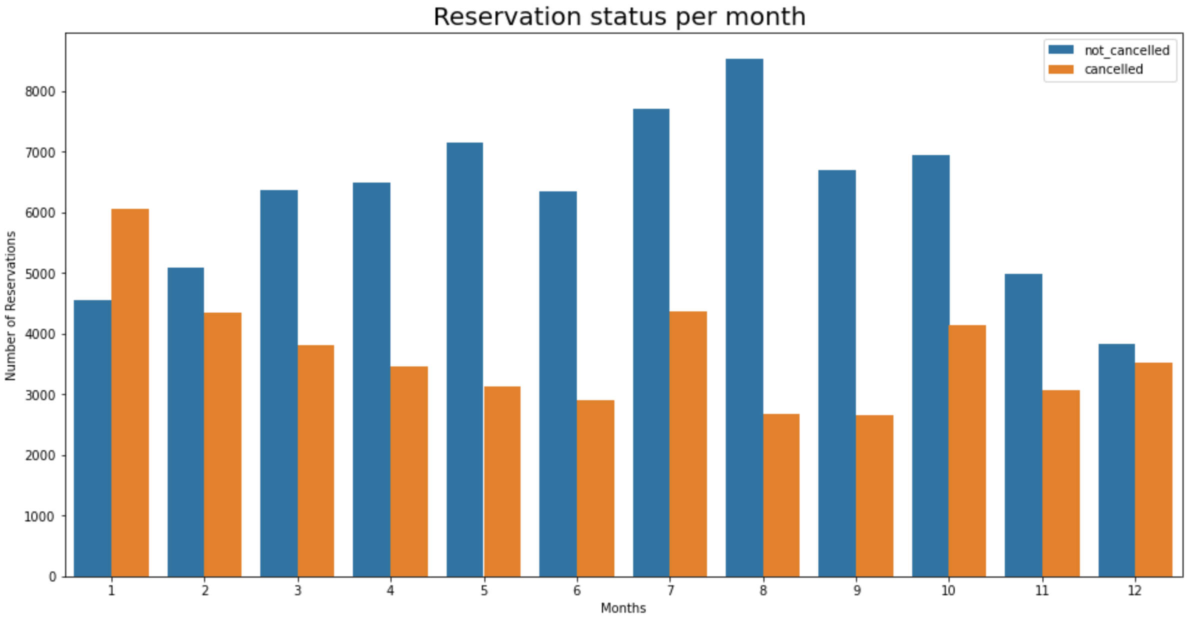


Here City hotel line is in mid of the Resort hotel as it states that the price of City hotel is less than City hotel price and spikes in the lines shows that it is due weekends and seasonal rates.

Checking Reservation and Cancellation Rate based on Months

In [77]:

```
df["month"]= df["reservation_status_date"].dt.month
plt.figure(figsize=(16,8))
ax1= sns.countplot(x="month",hue="is_canceled",data=df)
plt.title("Reservation status per month ",size=20)
plt.xlabel("Months")
plt.ylabel("Number of Reservations")
plt.legend(["not_cancelled","cancelled"])
plt.show()
```



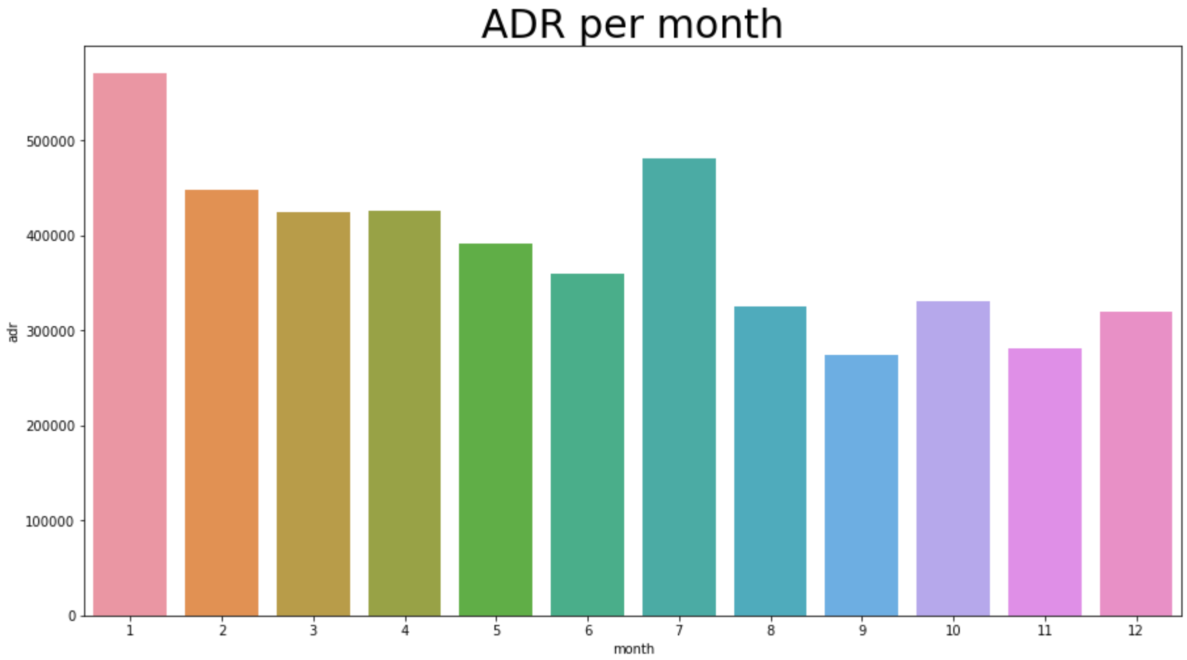
1.) In January there is maximum cancellation as comparision other followed by july and so on. 2.) In August there is minimum cancellation as comparision other followed by september and so on.

1.) In August there is maximum non-cancellation as comparision to others followed by july and so on. 2.) In December there is minimum non-cancellation as comparision to others followed by january and so on.

Checking effect of price on cancellation rate month wise

In [78]:

```
plt.figure(figsize=(15,8))
plt.title("ADR per month",fontsize=30)
sns.barplot("month","adr", data=df[df["is_canceled"]==1].groupby("month")[["adr"]]
plt.show()
```



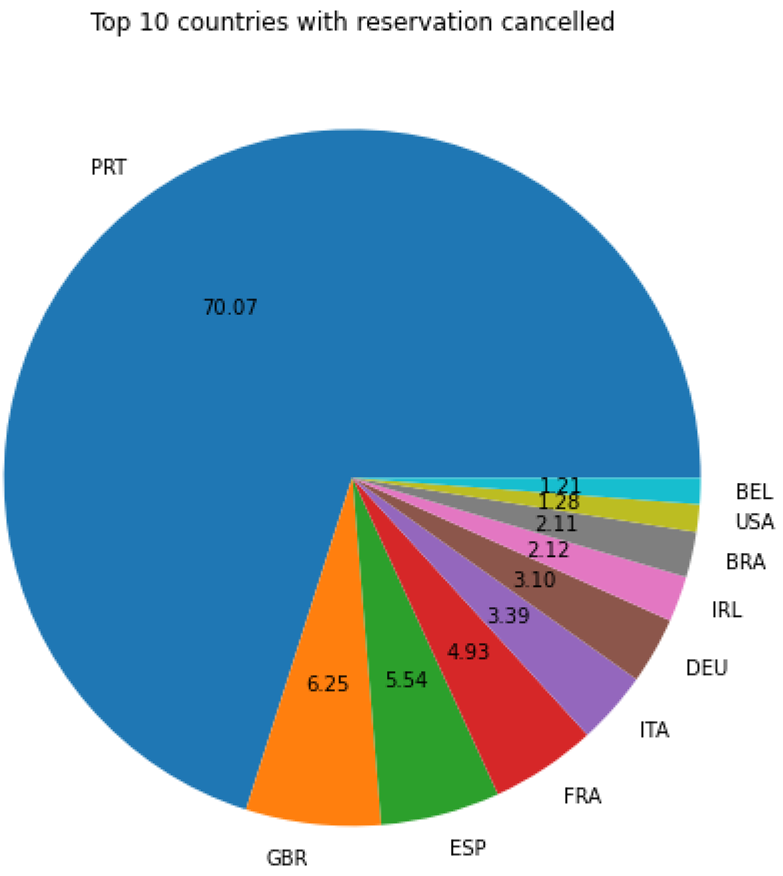
Here Adr in August is comparatively less followed by september and in august the cancellation is low.

In January the ADR is high and cancellation rate is high in that particular month.

So, it proves the Cancellation depends on prices of the hotel.

Checking Cancellation rate with respect to top 10 countries.

```
In [82]: cancelled_data= df[df["is_canceled"]==1]
         top_10_countries= cancelled_data["country"].value_counts()[:10]
         plt.figure(figsize=(8,8))
         plt.title("Top 10 countries with reservation cancelled")
         plt.pie(top_10_countries,autopct="%.2f",labels=top_10_countries.index)
         plt.show()
```



Here, PRT[Portugal] Country has maximum percent of cancellation rate.

Hotels Should work on price factors in PRT, Do promotional campagins adopting better system and facilites.

Checking the clients coming from which source.

```
In [86]: np.round((df["market_segment"].value_counts(normalize=True))*100,2)
```

Out[86]:

```
Online TA        47.44
Offline TA/TO    20.32
Groups           16.66
Direct           10.47
Corporate         4.30
Complementary     0.62
Aviation          0.20
Name: market_segment, dtype: float64
```

Checking cancellation with market_segment

In [92]:

```
np.round((cancelled_data["market_segment"].value_counts(normalize=True))*100,2)
```

Out[92]:

```
Online TA        46.97
Groups           27.40
Offline TA/TO    18.75
Direct            4.35
Corporate         2.22
Complementary     0.20
Aviation          0.12
Name: market_segment, dtype: float64
```

Here, Clients are mostly coming from Online TA and our assumption is that mostly clients are coming from Offline TA/TO.

Cancellation Rate is mostly on Online TA.

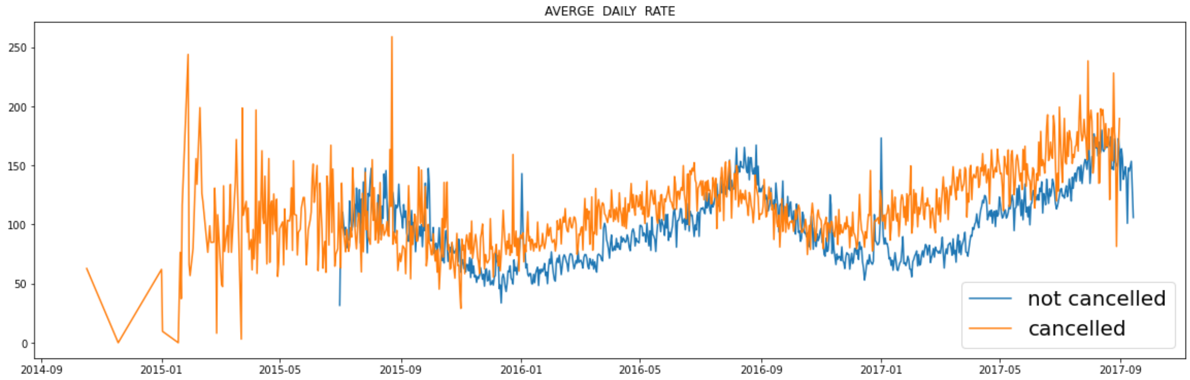
Online Regestration is 47.44% Online Cancellation is 46.97%

It Suggest Clients book hotels by viewing sites but they when they actual visit the hotel it might be not same when they see while booking. It may be the reason for high cancellation by Online TA.

Checking ADR for cancelled and non-cancelled

In [99]:

```
cancelled_df_adr= cancelled_data.groupby("reservation_status_date")[["adr"]].mean(
cancelled_df_adr.reset_index(inplace=True)
cancelled_df_adr.sort_values("reservation_status_date",inplace=True)

not_cancelled_data= df[df["is_canceled"]==0]

not_cancelled_df_adr= not_cancelled_data.groupby("reservation_status_date")[["adr"
not_cancelled_df_adr.reset_index(inplace=True)
not_cancelled_df_adr.sort_values("reservation_status_date",inplace=True)

plt.figure(figsize=(20,6))
plt.title("AVERGE  DAILY  RATE")

plt.plot(not_cancelled_df_adr["reservation_status_date"],not_cancelled_df_adr["adr
plt.plot(cancelled_df_adr["reservation_status_date"],cancelled_df_adr["adr"],label
plt.legend(fontsize=20)
plt.show()
```


AVERGE DAILY RATE
250
200
150
100
50
0
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
not cancelled
cancelled

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```