Import Libraries

In [1]:

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

Import Data Set

In [4]:

```
data = pd.read_csv(r"E:\hotel_booking.csv\hotel_booking.csv")
```

In [5]:

```
data.head()
```

Out[5]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_
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 [6]:

```
data.tail()
```

Out[6]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_da
119385	City Hotel	0	23	2017	August	35	
119386	City Hotel	0	102	2017	August	35	
119387	City Hotel	0	34	2017	August	35	
119388	City Hotel	0	109	2017	August	35	
119389	City Hotel	0	205	2017	August	35	

5 rows × 36 columns

Analysis and Cleaning

In [7]:

```
data.head(10)
```

Out[7]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of
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	Resort Hotel	0	14	2015	July	27	
6	Resort Hotel	0	0	2015	July	27	
7	Resort Hotel	0	9	2015	July	27	
8	Resort Hotel	1	85	2015	July	27	
9	Resort Hotel	1	75	2015	July	27	

10 rows × 36 columns

In [8]:

```
data.tail(10)
```

Out[8]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_da
119380	City Hotel	0	44	2017	August	35	
119381	City Hotel	0	188	2017	August	35	
119382	City Hotel	0	135	2017	August	35	
119383	City Hotel	0	164	2017	August	35	
119384	City Hotel	0	21	2017	August	35	
119385	City Hotel	0	23	2017	August	35	
119386	City Hotel	0	102	2017	August	35	
119387	City Hotel	0	34	2017	August	35	
119388	City Hotel	0	109	2017	August	35	
119389	City Hotel	0	205	2017	August	35	

10 rows × 36 columns

In [9]:

```
data.shape
```

Out[9]:

```
(119390, 36)
```

Data Cleaning / Removing

In [10]:

```
# Removing personal information in data :-

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

In [11]:

```
data.head()
```

Out[11]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of
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 × 32 columns

In [12]:

```
data.shape
```

Out[12]:

```
(119390, 32)
```

In [13]:

```
data.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'],
      dtype='object')
```

In [14]:

```
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 [15]:

```
data['reservation_status_date'] = pd.to_datetime(data['reservation_status_date'])
```

In [16]:

```
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
```

In [17]:

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

Out[17]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assi
count	119390	119390	119390	118902	119390	119390	119390	
unique	2	12	5	177	8	5	10	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	
freq	79330	13877	92310	48590	56477	97870	85994	

In [18]:

```
for col in data.describe(include='object').columns:
    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']
__________________________________________________
```

In [19]:

```
data.isnull().sum()
```

Out[19]:

```
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 [20]:

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

In [21]:

```
data.isnull().sum()
```

Out[21]:

```
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 [22]:

```
data.describe()
```

Out[22]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	
mean	0.371352	104.311435	2016.157656	27.166555	15.800880	
std	0.483168	106.903309	0.707459	13.589971	8.780324	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	

In [23]:

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

Data Analysis and Visualization

In [24]:

```
cancelled_perc = data['is_canceled'].value_counts(normalize = True)
```

In [25]:

```
cancelled_perc
```

Out[25]:

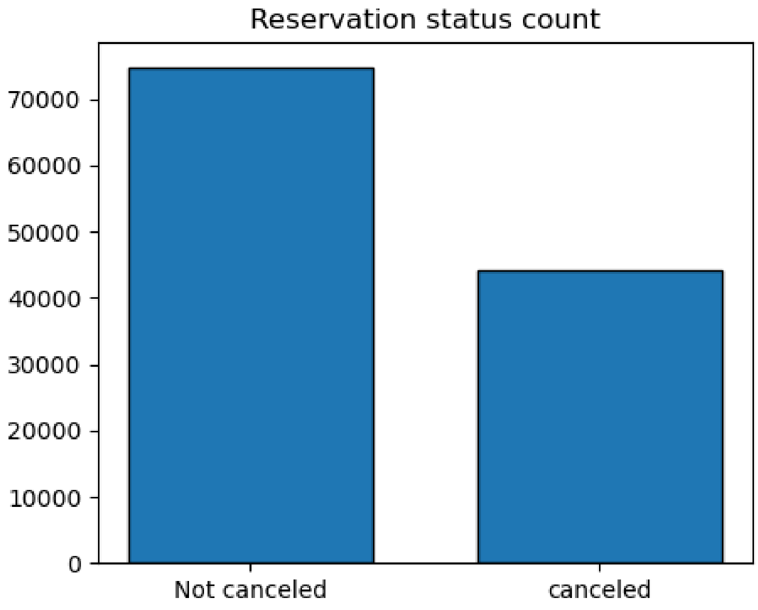
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

Analysis and Findings

In [26]:

```
cancelled_perc = data['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)

plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['Not canceled', 'canceled'],data['is_canceled'].value_counts(),edgecolor = 'k', width = 0.7)
plt.show()
```

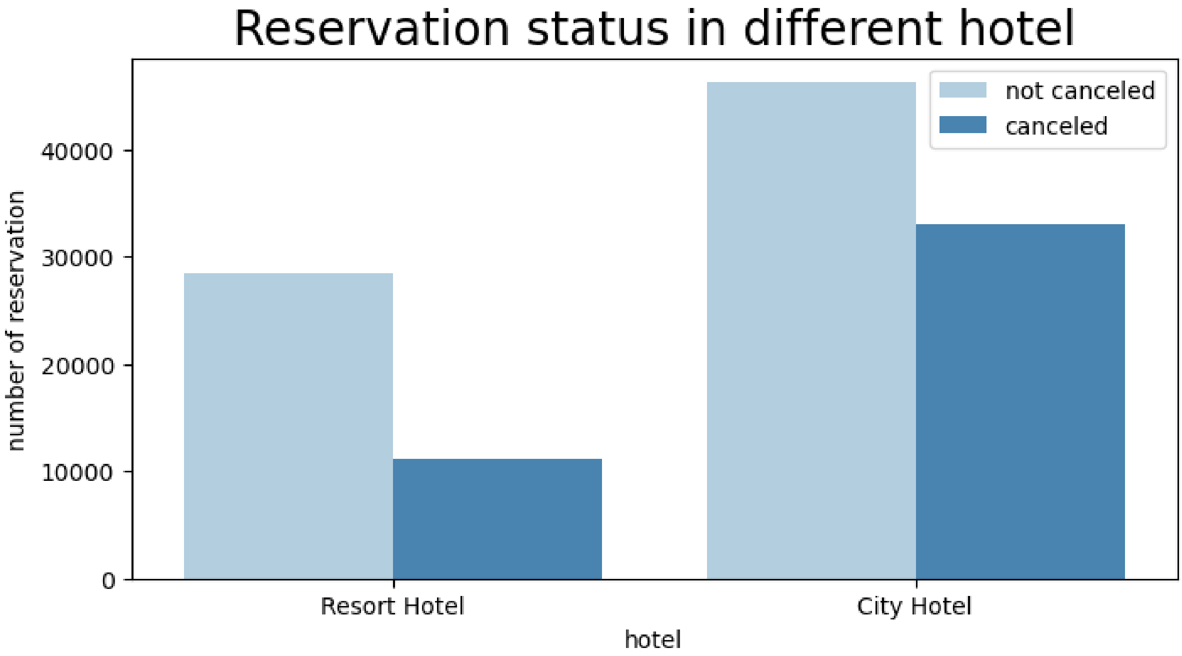
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```



The accompanying bar graph shows the percentage of reservations that are canceled and those that are not . it is obvious that there are stil a significant number of reservations that have not been canceled . There are still 37% of clients who canceled their reservation, which has a significant impact on the hotels earnings.

In [28]:

```
plt.figure(figsize=(8,4))
ax1 = sns.countplot(x ='hotel', hue = 'is_canceled', data = data, palette = 'Blues')
legend_labels,_=ax1. get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status in different hotel', size = 20)
plt.xlabel('hotel')
plt.ylabel('number of reservation')
plt.legend(['not canceled','canceled'])
plt.show()
```



In [29]:

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

Out[29]:

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

In [30]:

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

Out[30]:

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

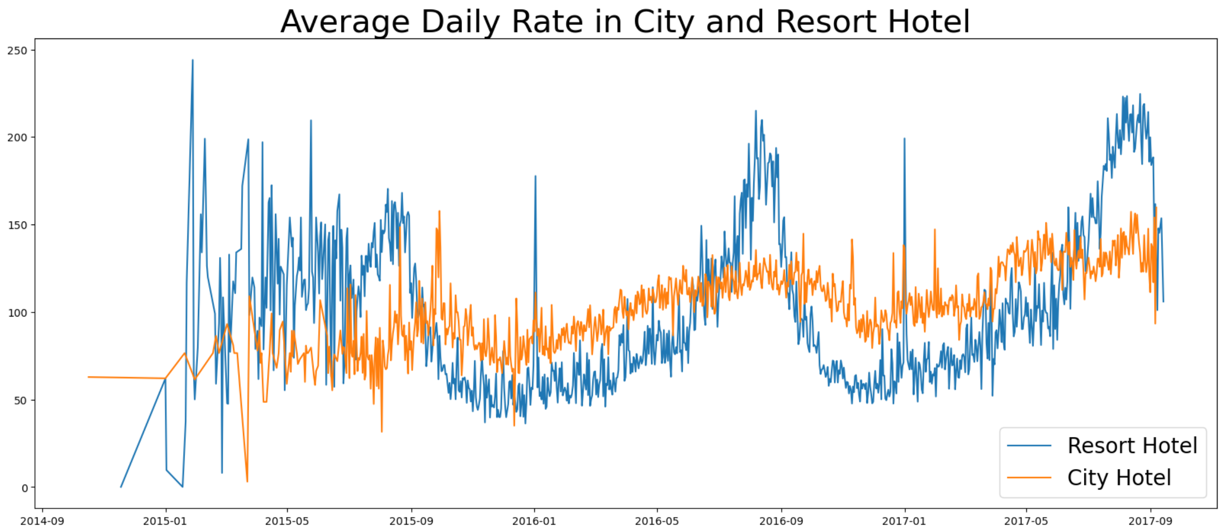
In [31]:

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

In comaprision to resort hotels. city hotels have more bookings. its possible the resort hotels are more expensive than those in cities.

In [33]:

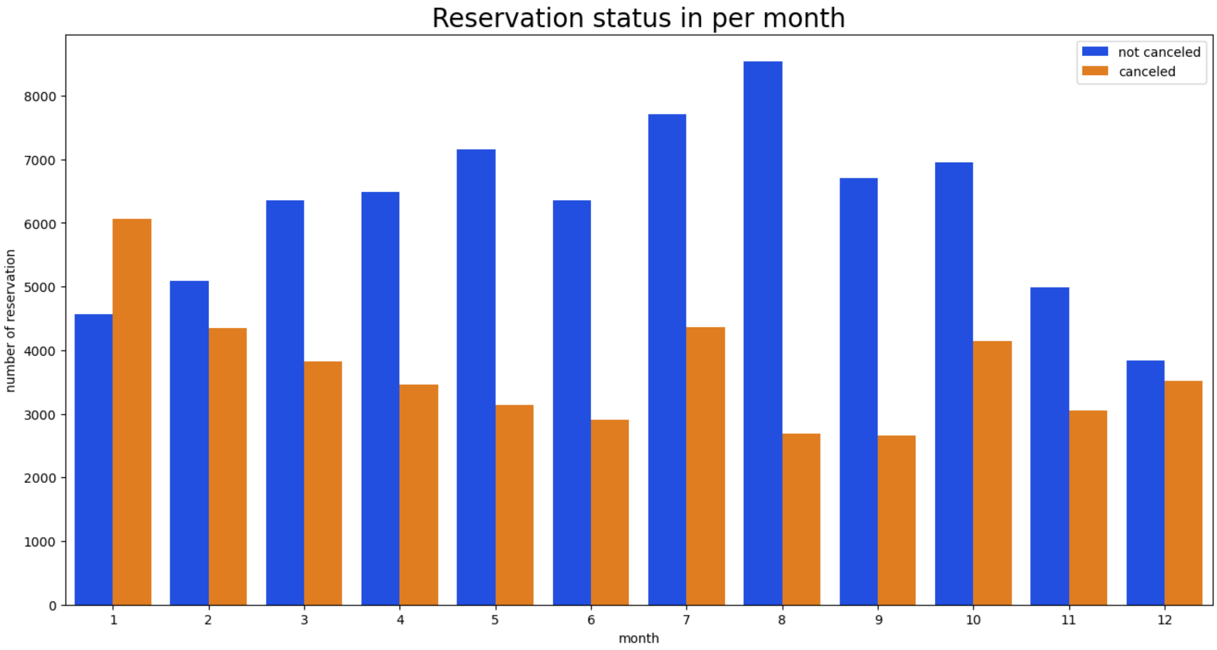
```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate in City and Resort Hotel', fontsize = 30)
plt.plot(resort_hotel.index,resort_hotel['adr'], label ='Resort Hotel')
plt.plot(city_hotel.index, city_hotel['adr'],label='City Hotel')
plt.legend(fontsize = 20)
plt.show()
```



The line graph above shows that , on certain days, the average daily rate for a city hotel is less than that of a resort hotel, and on other days, it is even less it goes without saying the weekends and holidays mays ee a rise in resort hotel rates.

In [38]:

```
data['month'] = data['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
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 in per month', size = 20)
plt.xlabel('month')
plt.ylabel('number of reservation')
plt.legend(['not canceled','canceled'])
plt.show()
```



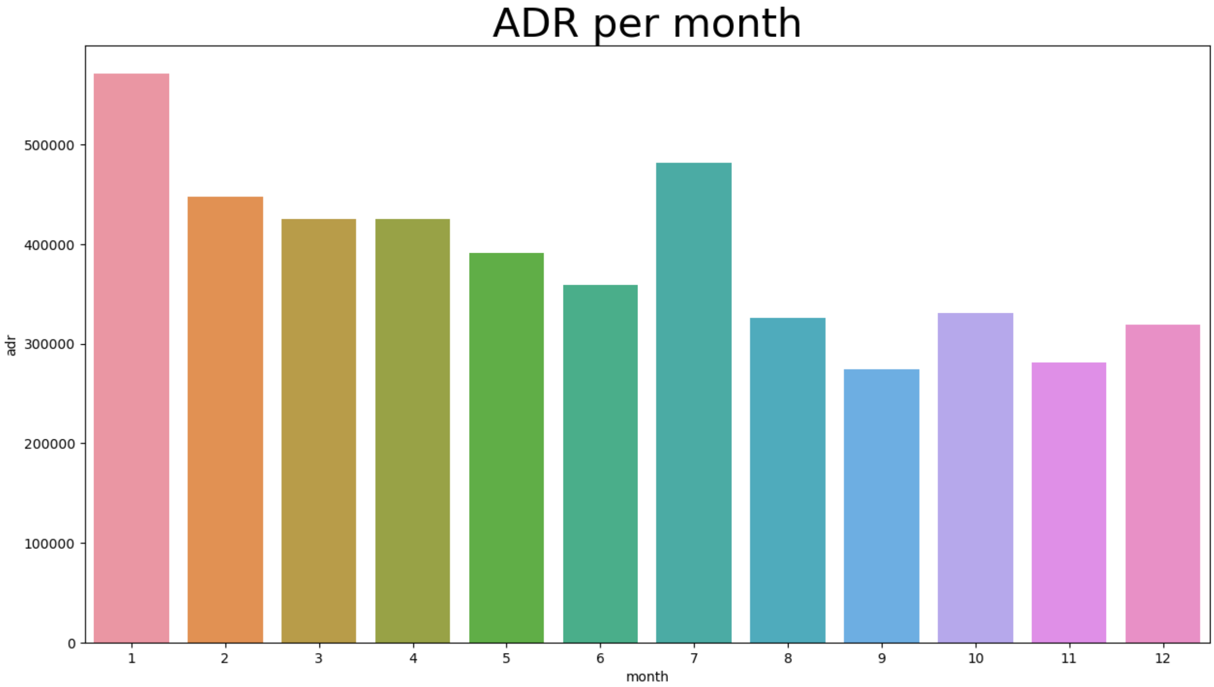
We have decided the grouped bar graph to analyze the month with the highest and lowest reservation levels according to reservation status. As can be seen , both thenumber of confirmed reservations and the number of canceled reservations are largest in the month of August. where as January is the month with the most canceled reservations.

In [39]:

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

Out[39]:

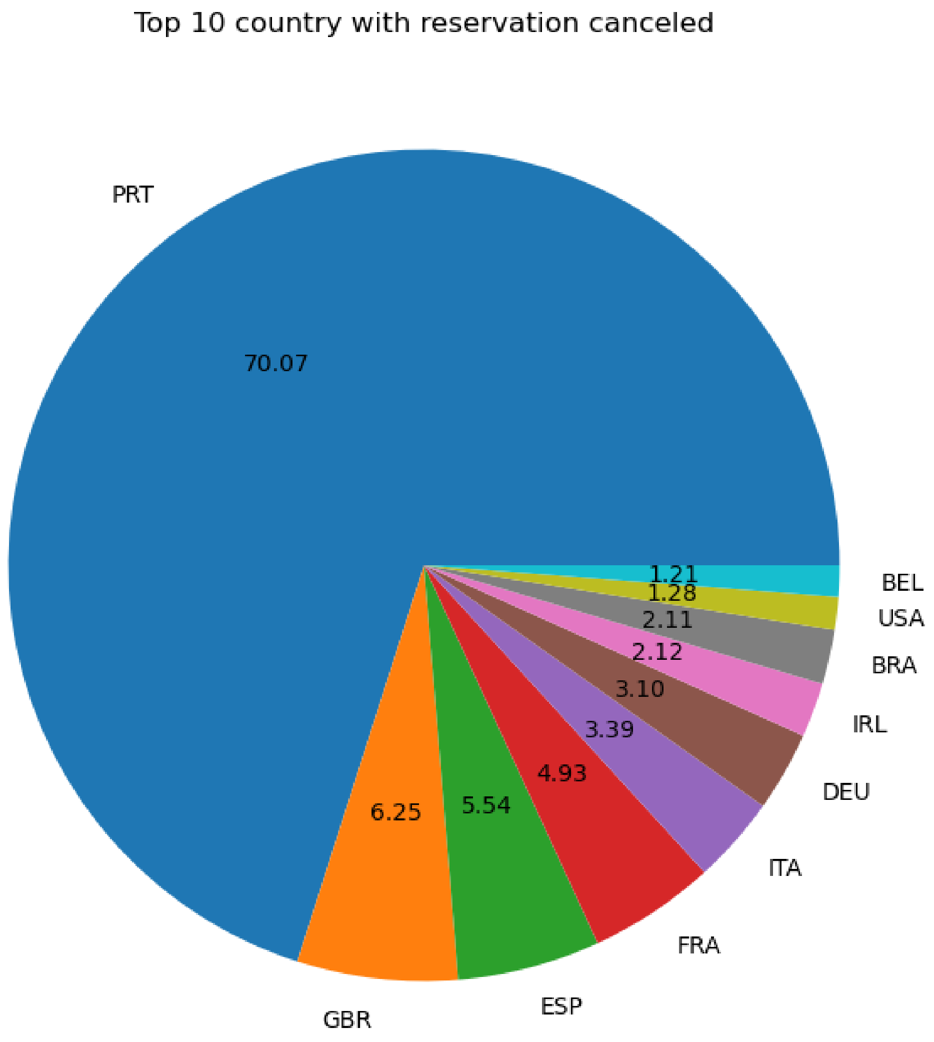
```
<AxesSubplot:title={'center':'ADR per month'}, xlabel='month', ylabel='adr'>
```



This bar graph demonstrate that cancelletions are most common when prices are greatest and are the least common when they are lowest. Therefore, the cost of the accomodation is soley responsible for the cancellation.

In [42]:

```
cancelled_data = data[data['is_canceled']==1]
top_10_country = cancelled_data['country'].value_counts()[:10]
plt.figure(figsize=(10,8))
plt.title('Top 10 country with reservation canceled')
plt.pie(top_10_country, autopct = '%.2f', labels = top_10_country.index)
plt.show()
```



Now let's see which country has the highest reservation canceled . The top country is Portugal with the highest number of cancellations.

Let's check the area from where guests are visiying the hotels and making reservations. Is it coming from direct or Goups , Online or Offline Travel Agents? Around 46% of the clients come from online travel agencies, where as 27% come from groups. Only 4% of c;ients book hotels directly by visiting them and making reservations.

In [43]:

```
data['market_segment'].value_counts()
```

Out[43]:

```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

In [44]:

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

Out[44]:

```
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: market_segment, dtype: float64
```

In [45]:

```
cancelled_data['market_segment'].value_counts(normalize = True)
```

Out[45]:

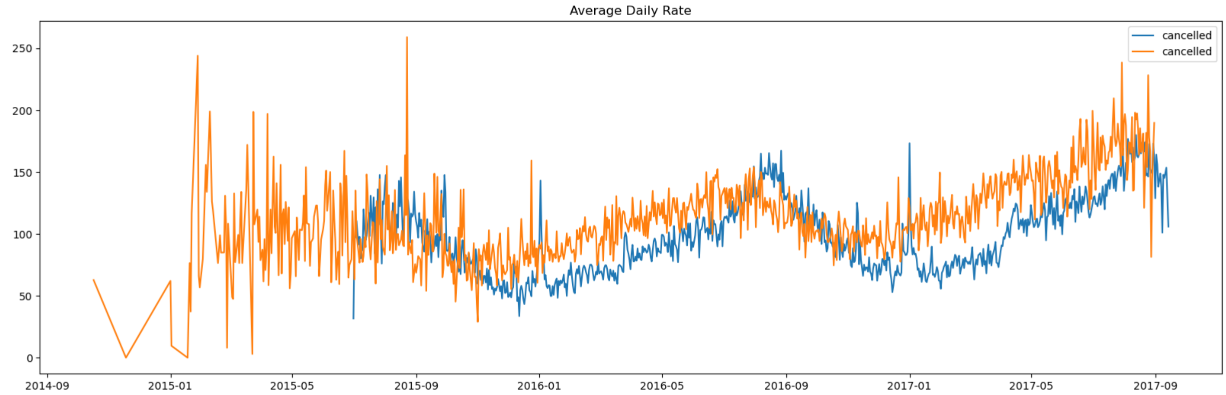
```
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: market_segment, dtype: float64
```

In [51]:

In [52]:

In [55]:

```
# Group cancelled reservation by reservation_status_date and calculate the average adr

cancelled_data_adr = cancelled_data.groupby('reservation_status_date')['adr'].mean().reset_index()
cancelled_data_adr.sort_values('reservation_status_date', inplace = True)

# Filter not cancelled reservation and calculate the average adr

not_cancelled_data = data[data['is_canceled']==0]
not_cancelled_data_adr = not_cancelled_data.groupby('reservation_status_date')['adr'].mean().reset_index()
not_cancelled_data_adr.sort_values('reservation_status_date', inplace =True)



#  Plot the average daily rate for both czncelled and not cancelled reservations

plt.figure(figsize=(20,6))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_data_adr['reservation_status_date'], not_cancelled_data_adr['adr'],label='cancelled')
plt.plot(cancelled_data_adr['reservation_status_date'], cancelled_data_adr['adr'],label='cancelled')
plt.legend()

#display the plot
plt.show()
```



In [57]:

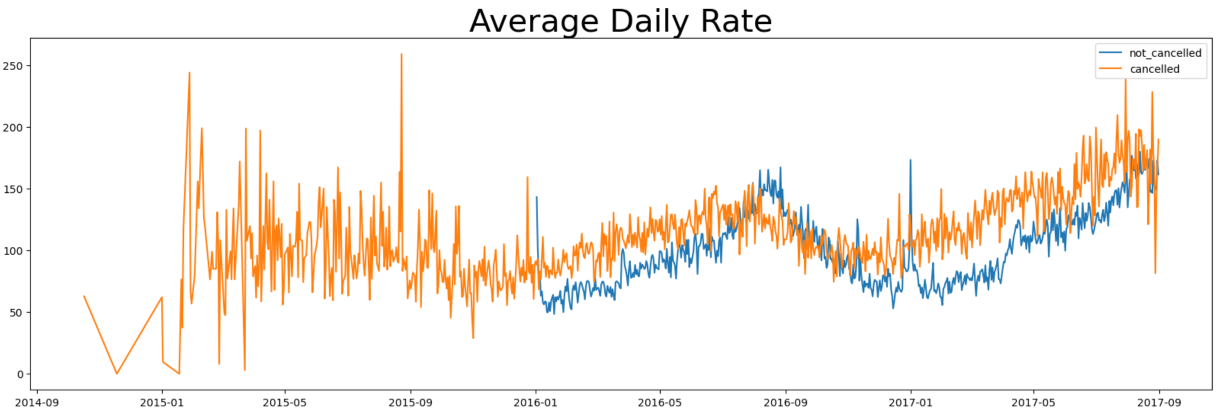
```
# Filter the "cancelled_data_adr" dataframe based on date condition

cancelled_data_adrv = cancelled_data_adr[(cancelled_data_adr['reservation_status_date']>'2016') &
                                         (cancelled_data_adr['reservation_status_date']< '2017-09')]
```

In [59]:

```
# Filter the data 'not_cancelled_data_adr' Dataframe based on date conditions

not_cancelled_data_adr = not_cancelled_data_adr[(not_cancelled_data_adr['reservation_status_date'] > '2016')
                                               (not_cancelled_data_adr['reservation_status_date'] < '2017-09
```

In [62]:

```
# Plot the average daily rate for both cancelled and not cancelled reservations

plt.figure(figsize=(20,6))
plt.title('Average Daily Rate', fontsize= 30)
plt.plot(not_cancelled_data_adr['reservation_status_date'],
         not_cancelled_data_adr['adr'],label = 'not_cancelled')

plt.plot(cancelled_data_adr['reservation_status_date'],
         cancelled_data_adr['adr'],label ='cancelled')
plt.legend()

# Display the plot
plt.show()
```



AS seen in the graph , reservation are cancelled when the average daily rate is higher than when it is not canceled . It clearly proves all the above analysis, that the hogher price leads to higher cancellation.

Suggestion :-

1. Cancellation rates rise as the price does . In order to prevent cancellation of reservations, hotels could work on their pricing strategies and try to lower rates for specific hotels based on locatiobn. They can also provide some discounts to the consumers.
2. As the ratio of the cancellation and not cancellation of the resort hotel higher in the resort hotel than the city hotels. So the hotels should provide a reasonable discount on the room prices on weekends or on holidays .
3. In the month of january , hotels can start campaign or marketinf with a reasonablr amount to increase their revenue as the cancelation is the highest in this month.
4. They can also increase the quality of their hotels and their services mainly in Portugal to reduce the cancellation rate.
5. They can change minimum amount of booking cancellation.
6. Also hotels can provide a cupons for previous customer to discount on next visit.

In [ ]:

In [ ]: