In [8]:

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

In [4]:

```
df=pd.read_csv('hotel_booking.csv')
```

Exploratory Data Analysis and Data Cleaning

In [23]:

```
df.head()
```

Out[23]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_dat
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 [9]:

```
df.shape
```

Out[9]:

```
(119390, 36)
```

In [10]:

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

In [12]:

```
df.columns
```

Out[12]:

```
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 [13]:
```
df.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 [14]:
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

In [16]:

```
df.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 [18]:

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

Out[18]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_ty
count	119390	119390	119390	118902	119390	119390	1193
unique	2	12	5	177	8	5	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	
freq	79330	13877	92310	48590	56477	97870	859

In [20]:

```python
for i in df.describe(include='object').columns:
    print(i)
    print(df[i].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 [22]:
```
df.isnull().sum()
```

Out[22]:
```
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 [24]:
```
#df.drop(['agent','company'], axis=1,inplace=True)
```

In [25]:
```
df.dropna(inplace=True)
```

In [26]:
```
df.isnull().sum()
```

Out[26]:
```
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 [27]:
```
df.describe()
```

Out[27]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month
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 [28]:
```
df=df[df['adr']<5000]
```

In [29]:

```
df.describe()
```

Out[29]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000
mean	0.371347	104.312018	2016.157657	27.166674	15.800802
std	0.483167	106.903570	0.707462	13.589966	8.780321
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

Data Analysis and Visualization

In [32]:

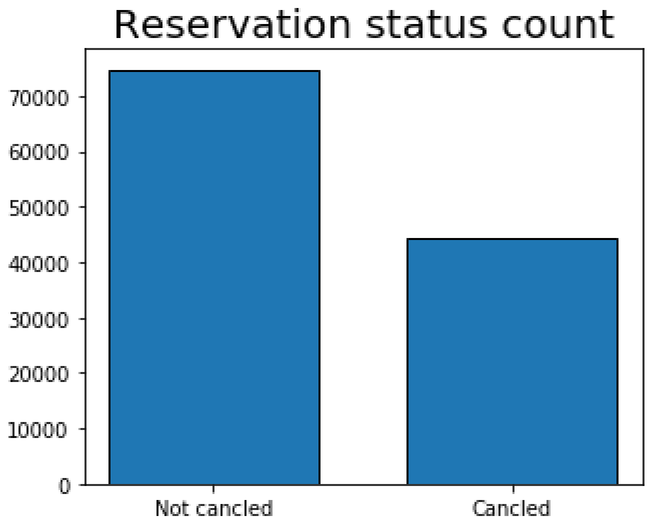
```
cancelled_percentage = df['is_canceled'].value_counts(normalize=True)
cancelled_percentage
```

Out[32]:

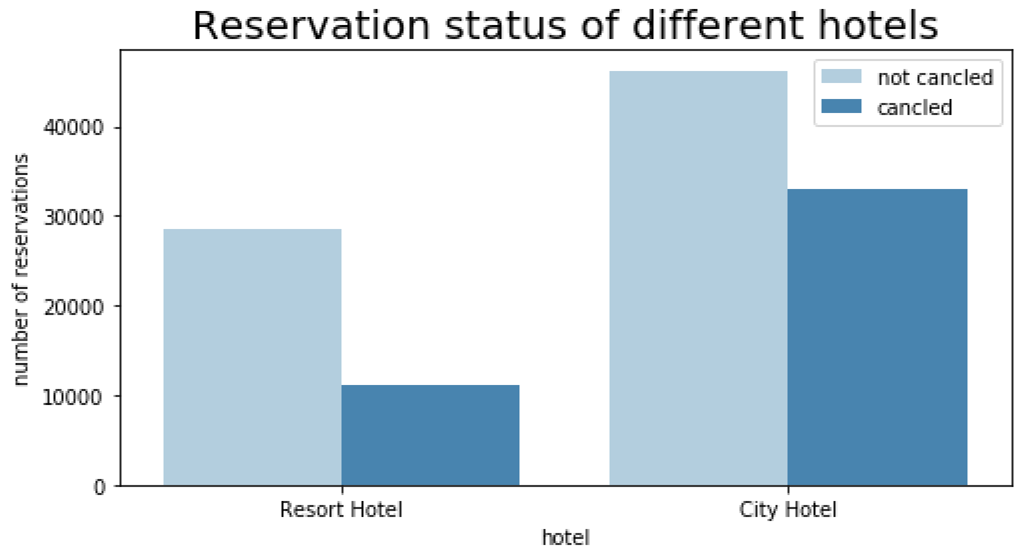
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

In [37]:

```
plt.figure(figsize=(5,4))
plt.title('Reservation status count',size=20)
plt.bar(['Not cancled','Cancled'],df['is_canceled'].value_counts(),edgecolor
='k',width=0.7)
plt.show()
```



```
In [36]: plt.figure(figsize=(8,4))
         ax1 = sns.countplot(x='hotel',hue='is_canceled',data=df,palette='Blues')
         legend_labels,_=ax1.get_legend_handles_labels()
         plt.title('Reservation status of different hotels',size=20)
         plt.xlabel('hotel')
         plt.ylabel('number of reservations')
         plt.legend(['not cancled','cancled'])
         plt.show()
```



```
In [38]: resort_hotel = df[df['hotel']=='Resort Hotel']
         resort_hotel['is_canceled'].value_counts(normalize=True)

Out[38]: 0    0.72025
         1    0.27975
         Name: is_canceled, dtype: float64
```

```
In [39]: city_hotel = df[df['hotel']=='City Hotel']
         city_hotel['is_canceled'].value_counts(normalize=True)

Out[39]: 0    0.582918
         1    0.417082
         Name: is_canceled, dtype: float64
```

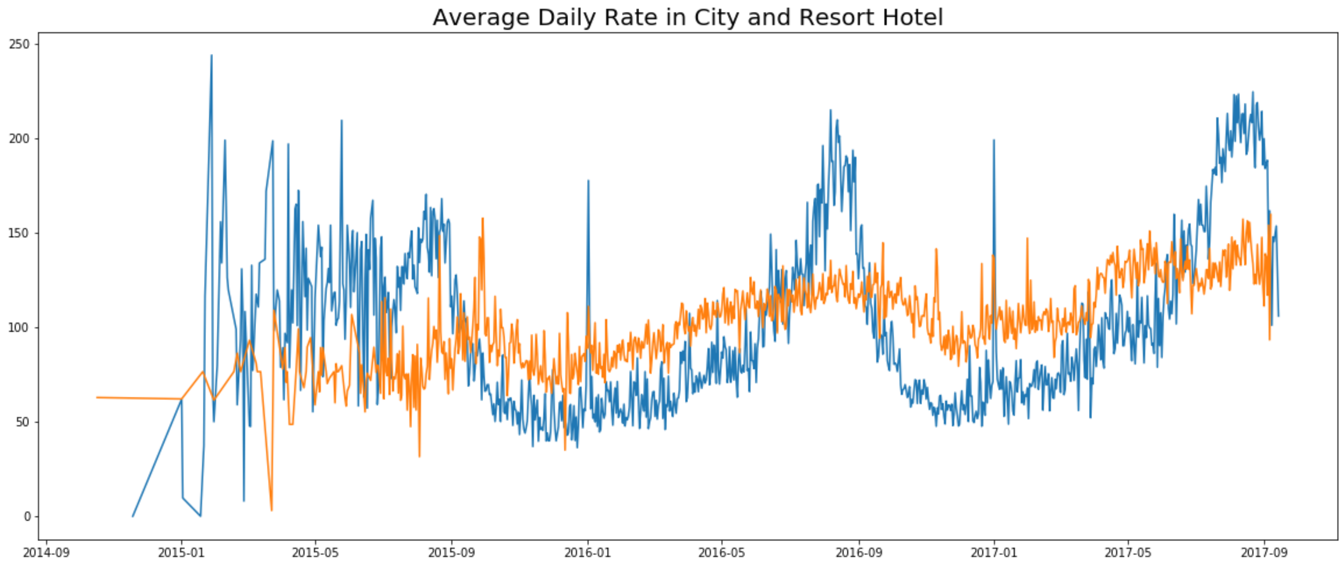
```
In [40]: resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
         city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

In [41]:

```
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')
```

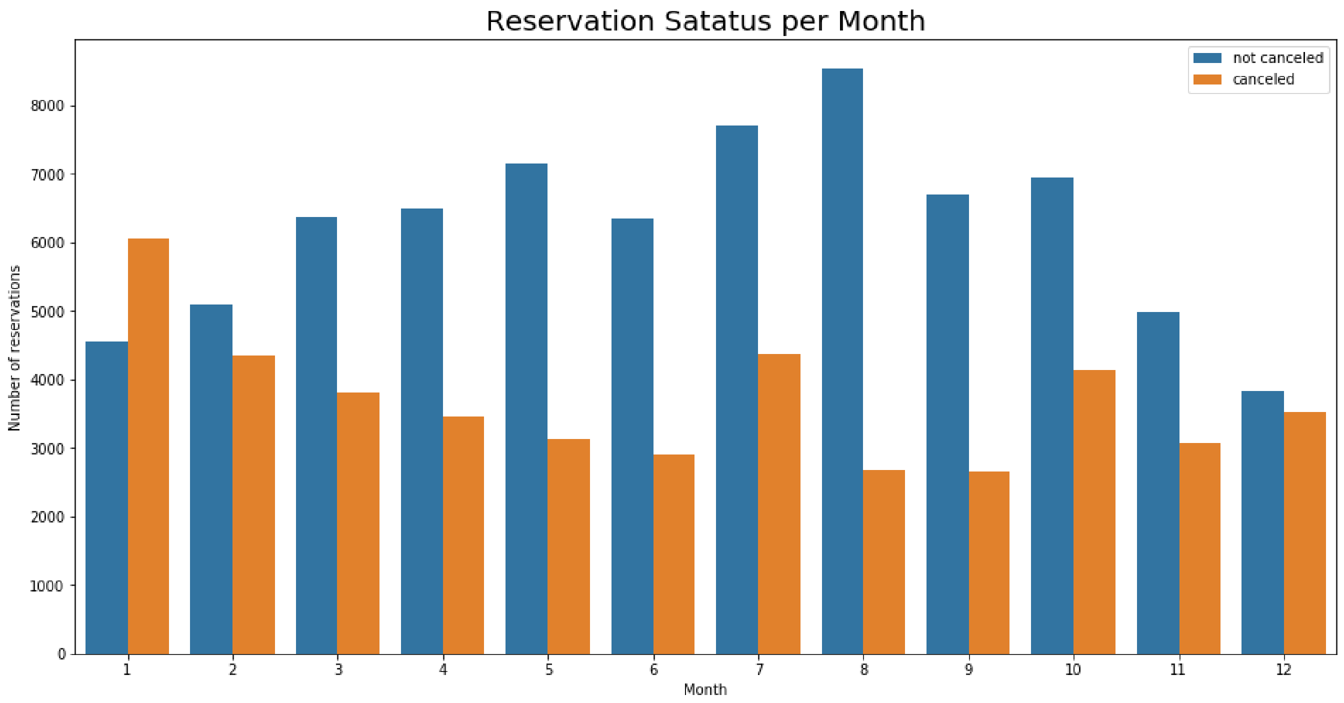
Out[41]:

```
[<matplotlib.lines.Line2D at 0x14d87ba5e88>]
```



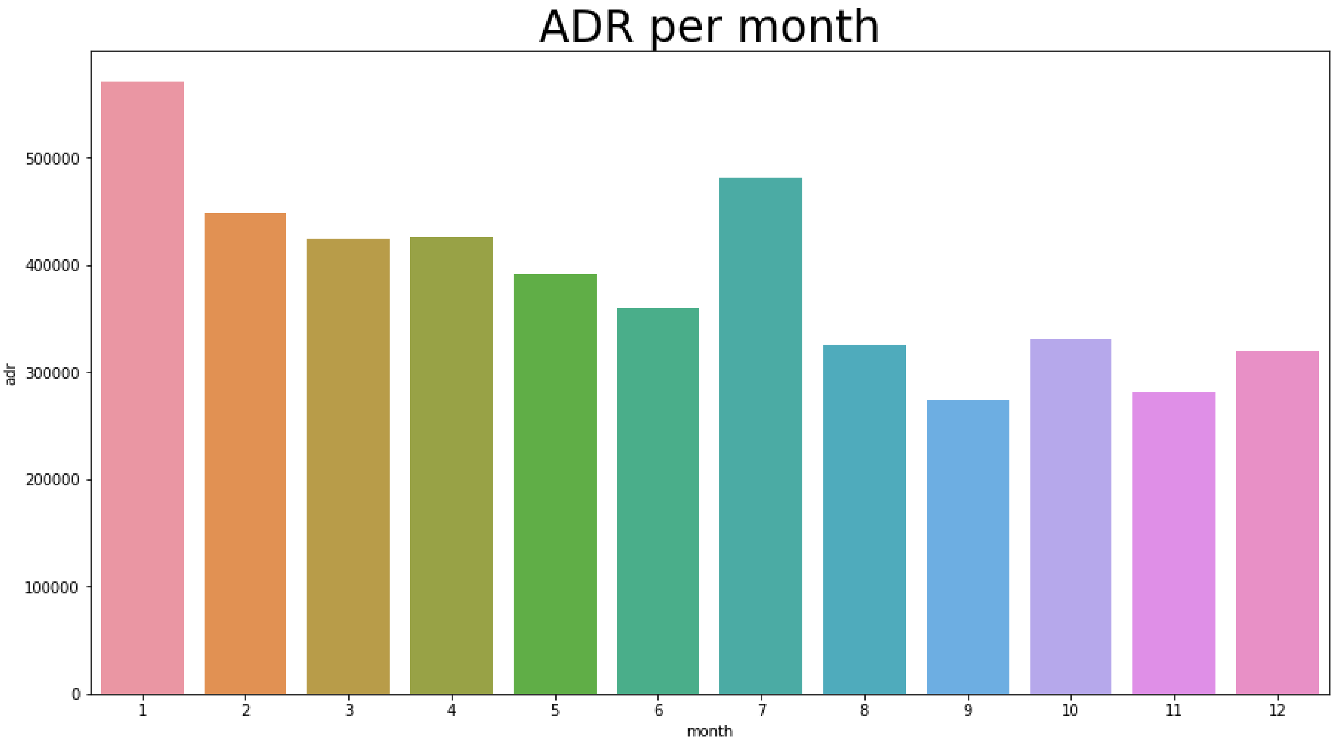
In [45]:

```
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 Satatus per Month',size=20)
plt.xlabel('Month')
plt.ylabel('Number of reservations')
plt.legend(['not canceled','canceled'])
plt.show()
```

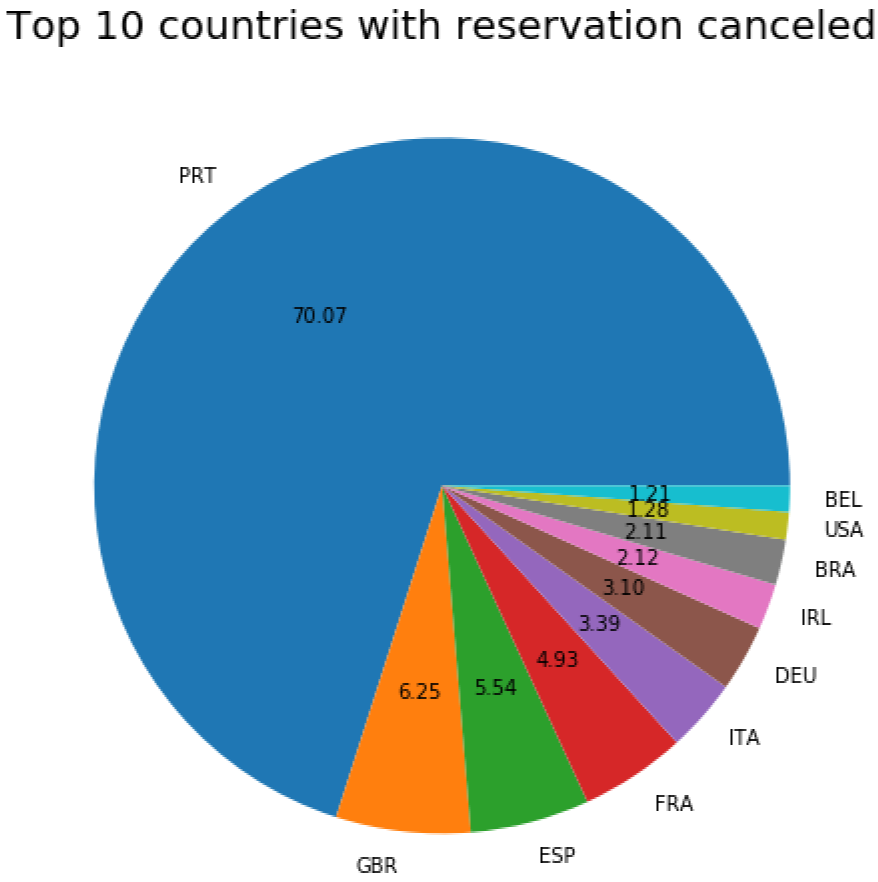


In [46]:

```
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']].sum().reset_index())
plt.show()
```



```
In [52]: cancelled_data = df[df['is_canceled']==1]
         top_10_country = cancelled_data['country'].value_counts()[:10]
         plt.figure(figsize = (8,8))
         plt.title('Top 10 countries with reservation canceled',size=20)
         plt.pie(top_10_country,labels= top_10_country.index,autopct='%.2f')
         plt.show()
```



```
In [53]: df['market_segment'].value_counts()
```

```
Out[53]: Online TA        56402
         Offline TA/TO    24159
         Groups           19806
         Direct           12448
         Corporate         5111
         Complementary      734
         Aviation           237
         Name: market_segment, dtype: int64
```

```
In [54]: df['market_segment'].value_counts(normalize=True)
```

```
Out[54]: 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 [55]:
```
cancelled_data['market_segment'].value_counts(normalize=True)
```

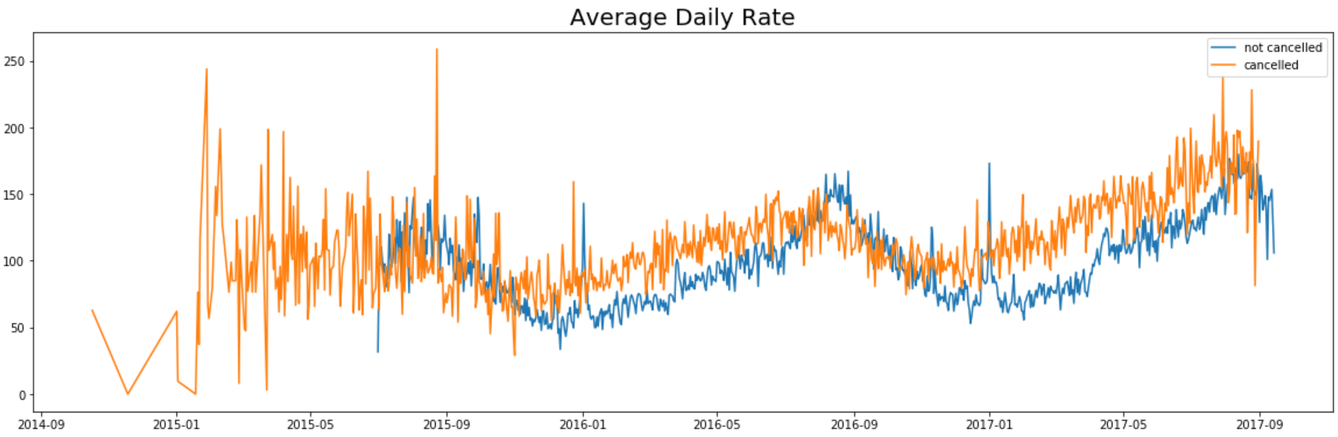
Out[55]:
```
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: market_segment, dtype: float64
```

In [61]:
```
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']].mean()
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('Average Daily Rate',size=20)
plt.plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr['adr'],label='not cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label='cancelled')
plt.legend()
```

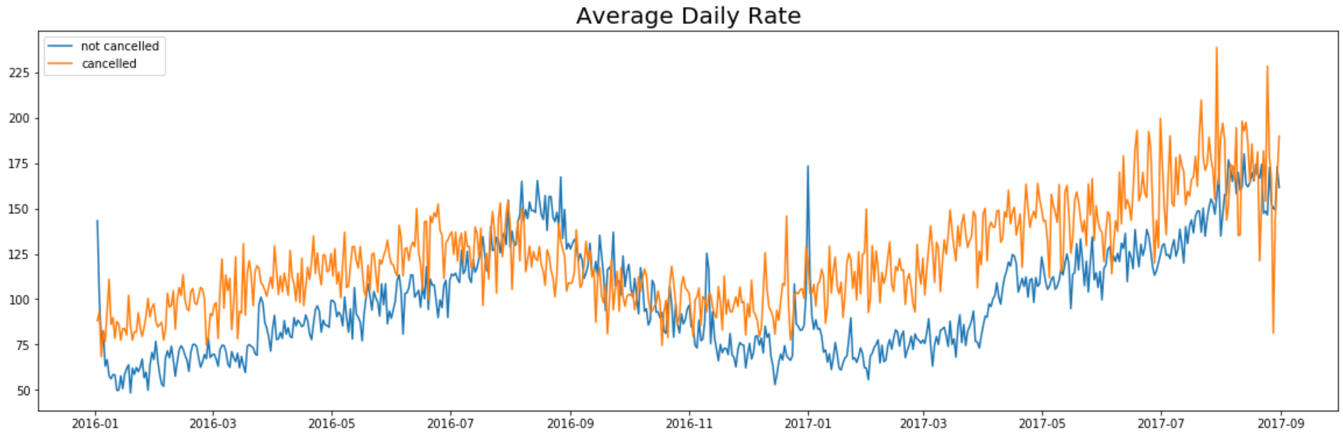
Out[61]:
```
<matplotlib.legend.Legend at 0x14d87a91c08>
```



In [62]:
```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016')&(cancelled_df_adr['reservation_status_date']<'2017-09')]
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016')&(not_cancelled_df_adr['reservation_status_date']<'2017-09')]
```

```
In [63]: plt.figure(figsize=(20,6))
         plt.title('Average Daily Rate',size=20)
         plt.plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr
         ['adr'],label='not cancelled')
         plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],la
         bel='cancelled')
         plt.legend()
```

```
Out[63]: <matplotlib.legend.Legend at 0x14d879aafc8>
```



Conclusion

1. Cancellation retes rise as the price does. In order to prevent cancellation of reservation, hotels could work on their pricing strategies and try to lower the rates for specific hotels based on locations. they can also provide some discount to the consumers.

2. As the ratio of the cancellation and not cancellation of the resort hotel is higher in resort hotel than city hotels.So the hotels should provide a reasonable discount on the room price on weekends or on holidays.

3. In the month of January, hotels can start campaings or making with a resonable amount to increase their revenue as the cancellation is the highest in this month.

4. They can also increase the quality of their hotels and their service mainly in portugal to reduce the cencellation rate.