In [5]:

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

In [7]:

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

In [8]:

```
df['reservation_status_data'] = pd.to_datetime(df['reservation_status_date'])
```

In [9]:

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

Out[9]:

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

In [10]:

```
for col in df.describe(include = 'object').columns:
    print(col)
    print(df[col].unique())
    print('-'*50)
```

```
'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' 'G
roups'
 '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']
--------------------------------------------------
```

In [11]:

```
df.isnull().sum()
```

Out[11]:

```
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
reservation_status_data             0
dtype: int64
```

In [12]:

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

In [13]:

```
df.isnull().sum()
```

Out[13]:

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

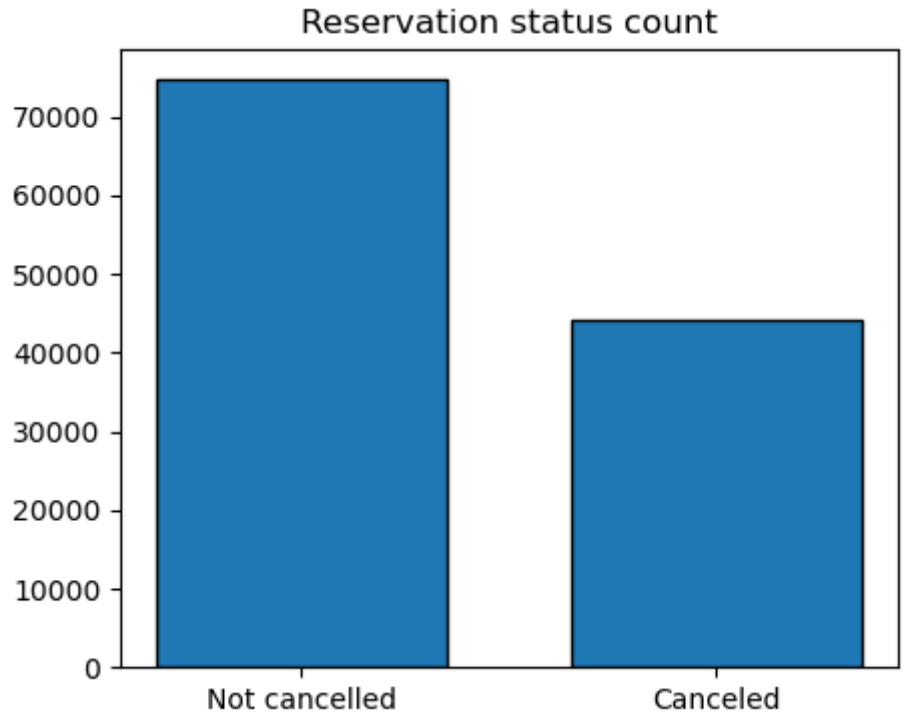
In [14]:

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

In [15]:

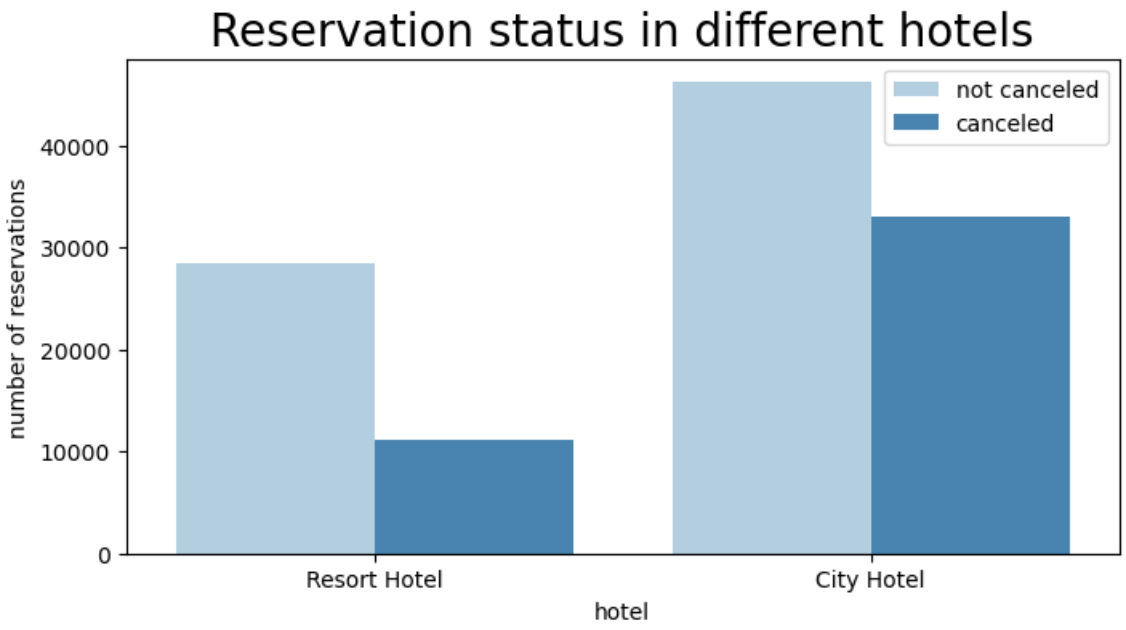
```
cancelled_per = df['is_canceled'].value_counts(normalize = True)
print(cancelled_per)

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

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



In [16]:

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



In [17]:

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

Out[17]:

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

In [18]:

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

Out[18]:

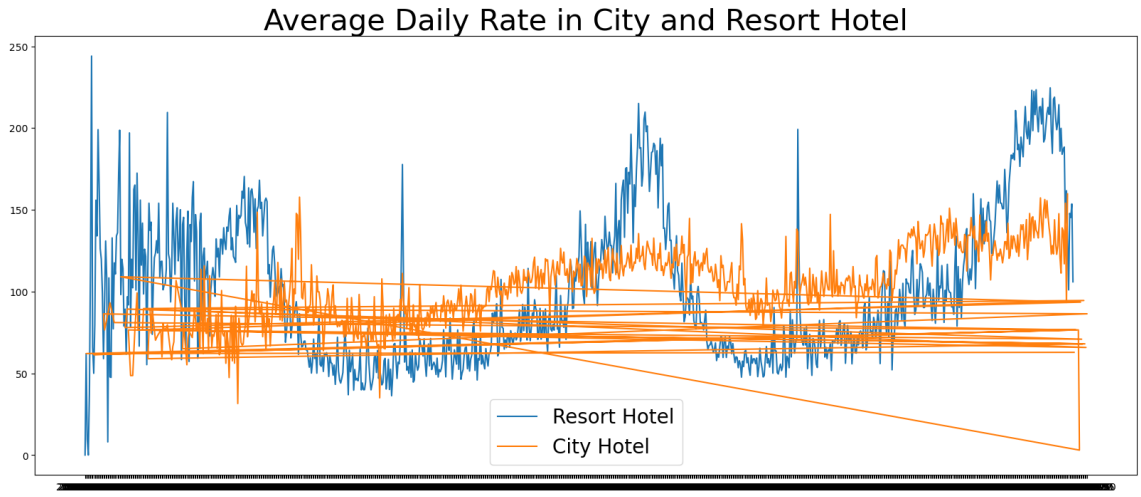
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [21]:

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

In [22]:

```
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()
```

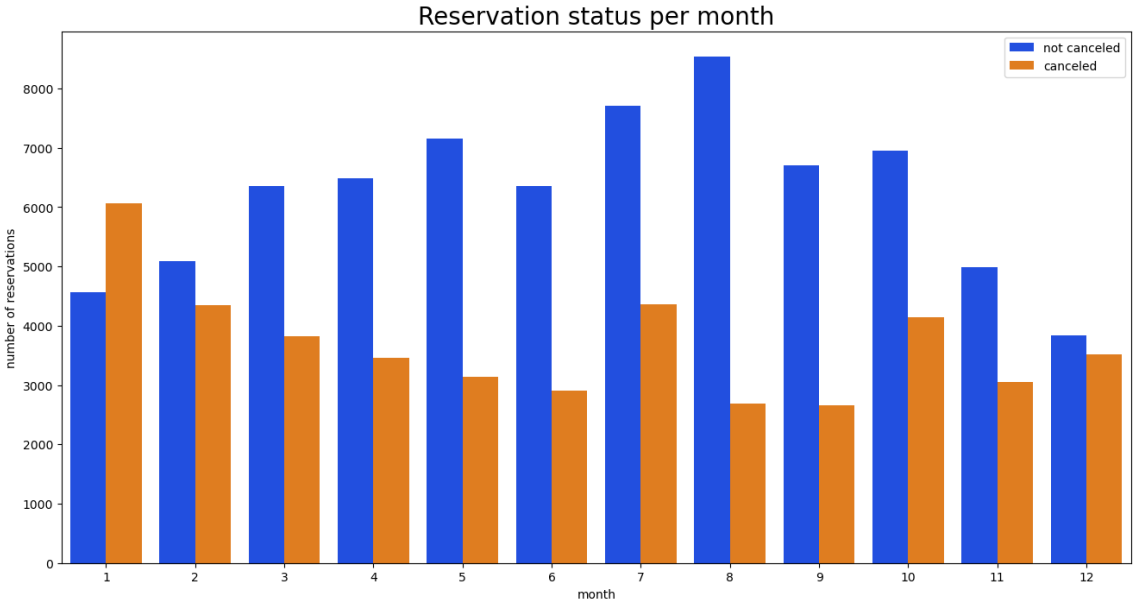


In [29]:

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


df['month'] = df['reservation_status_date'].dt.month
```

In [30]:

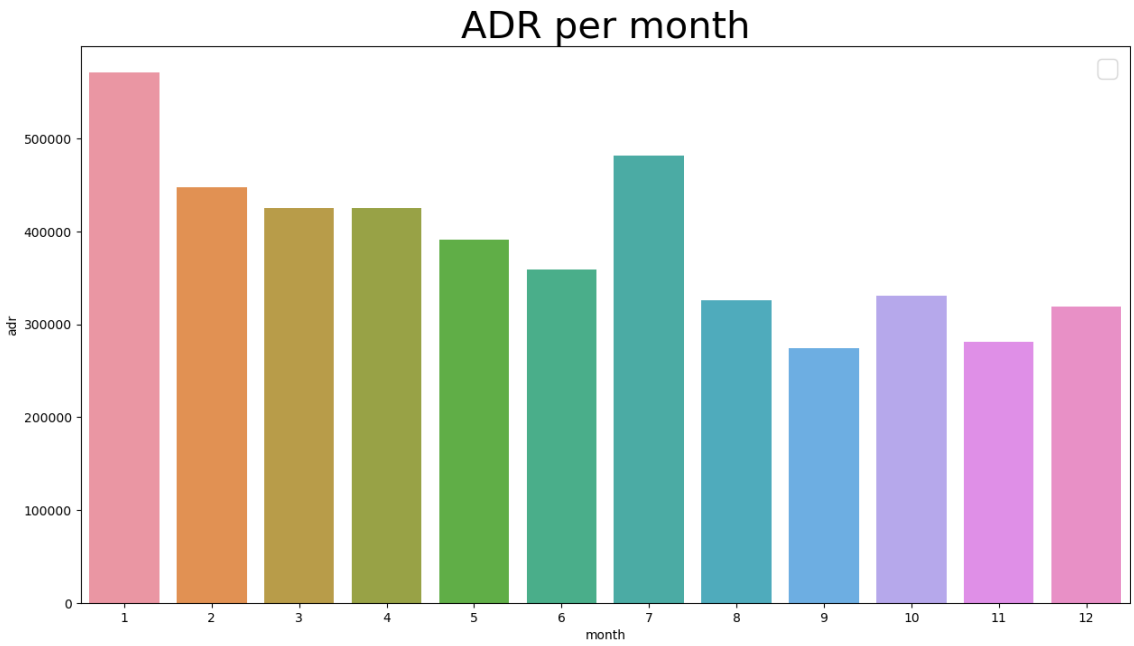
```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
ax1 = sns.countplot(x='month', hue='is_canceled', data=df, palette='bright')
legend_labels, _ = ax1.get_legend_handles_labels()
ax1.legend(legend_labels, bbox_to_anchor=(1, 1))
plt.title('Reservation status per month', size=20)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.legend(['not canceled','canceled'])
plt.show()
```



In [31]:

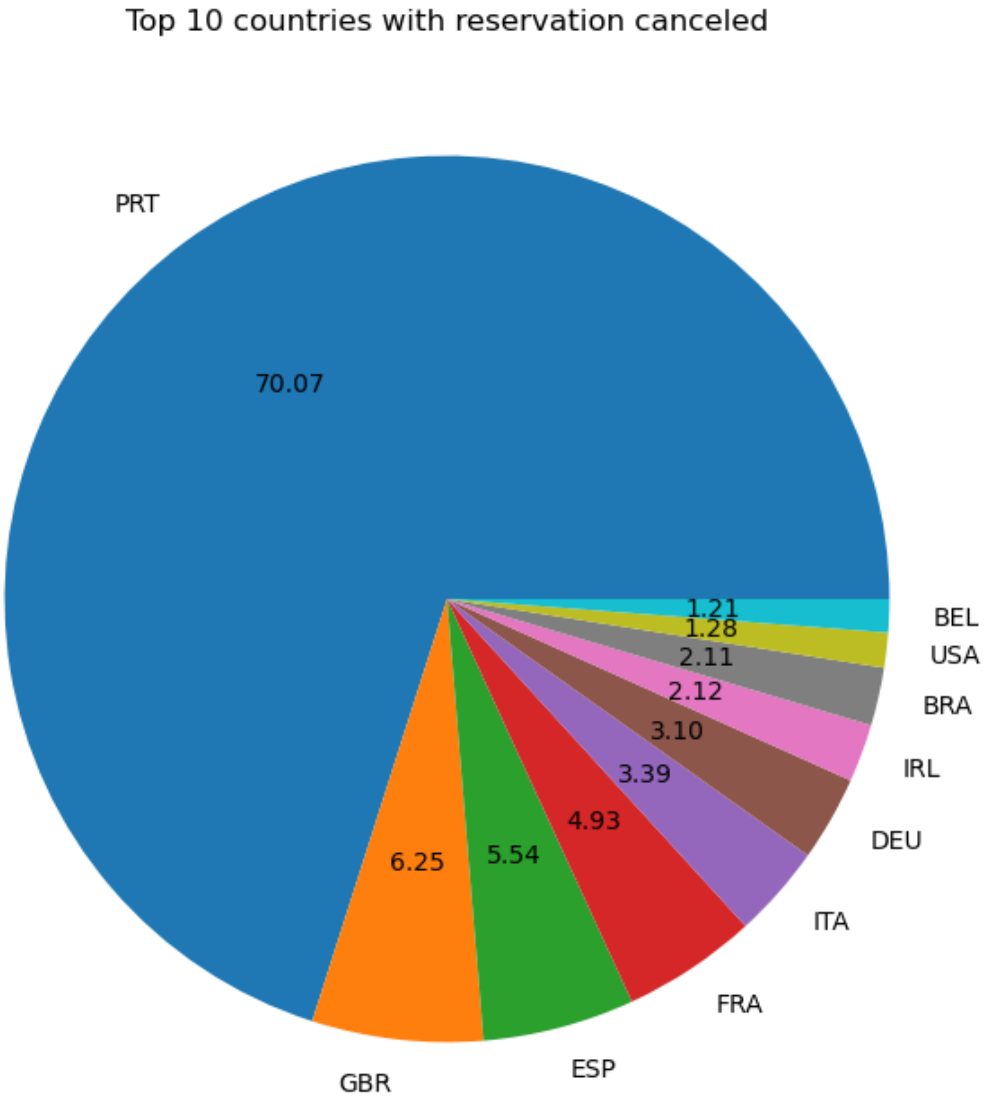
```
plt.figure(figsize = (15,8))
plt.title('ADR per month',fontsize = 30)
sns.barplot('month','adr',data = df[df['is_canceled'] == 1].groupby('month')[['adr']
plt.legend(fontsize = 20)
plt.show()
```

```
No artists with labels found to put in legend.  Note that artists whose
label start with an underscore are ignored when legend() is called with
no argument.
```



In [32]:

```
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')
plt.pie(top_10_country, autopct = '%.2f', labels = top_10_country.index)
plt.show()
```



In [33]:

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

Out[33]:

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

In [34]:

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

Out[34]:

```
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 [35]:

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

Out[35]:

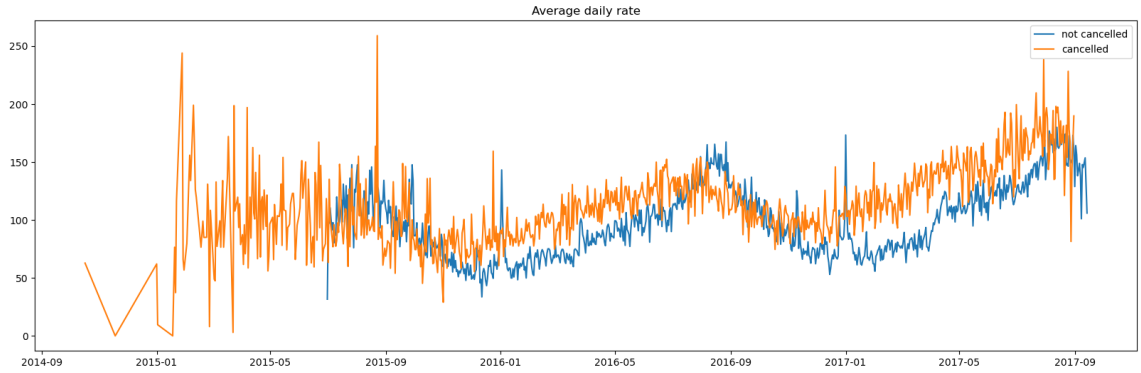
```
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: market_segment, dtype: float64
```

In [ ]:

```

```

In [53]:

```
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('Average 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()
plt.show()
```

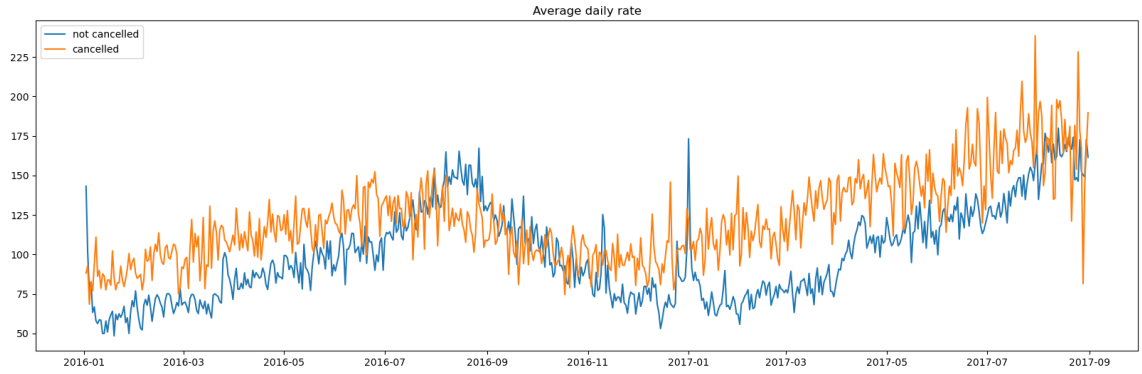


In [55]:

```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] > '
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status
```

In [56]:

```
plt.figure(figsize = (20,6))
plt.title('Average 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()
plt.show()
```



In [ ]: