Importing libraries

In [16]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading the dataset

In [17]:
```
df= pd.read_csv(r"C:\Users\X483151\OneDrive - Old Mutual\Desktop\dashboards\Data Portfolio\archive (2)\hotel_booking.csv")
```

Exploratory data analysis and Data cleaning

In [18]:
```
df.head()
```

Out[18]:

	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	...
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...

5 rows × 36 columns

In [19]:
```
df.tail(10)
```

Out[19]:

	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
119380	City Hotel	0	44	2017	August	35	31	1	3	2
119381	City Hotel	0	188	2017	August	35	31	2	3	2
119382	City Hotel	0	135	2017	August	35	30	2	4	3
119383	City Hotel	0	164	2017	August	35	31	2	4	2
119384	City Hotel	0	21	2017	August	35	30	2	5	2
119385	City Hotel	0	23	2017	August	35	30	2	5	2
119386	City Hotel	0	102	2017	August	35	31	2	5	3
119387	City Hotel	0	34	2017	August	35	31	2	5	2
119388	City Hotel	0	109	2017	August	35	31	2	5	2
119389	City Hotel	0	205	2017	August	35	29	2	7	2

10 rows × 36 columns

In [20]:
```
df['reservation_status_date']= pd.to_datetime(df['reservation_status_date'])
```

In [21]:
```
df.describe(include='object')
```

Out[21]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166

In [22]:
```
for col in df.describe( include= 'object'):
    print(col)
    print(df[col].unique())
```

```
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 [23]:
```
df.isnull().sum()
```

Out[23]:
```
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 [24]:
```
df.drop(['company','agent'], axis=1 , inplace= True)
df.dropna(inplace= True)
```

In [25]:
```
df.isnull().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                          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
dtype: int64
```

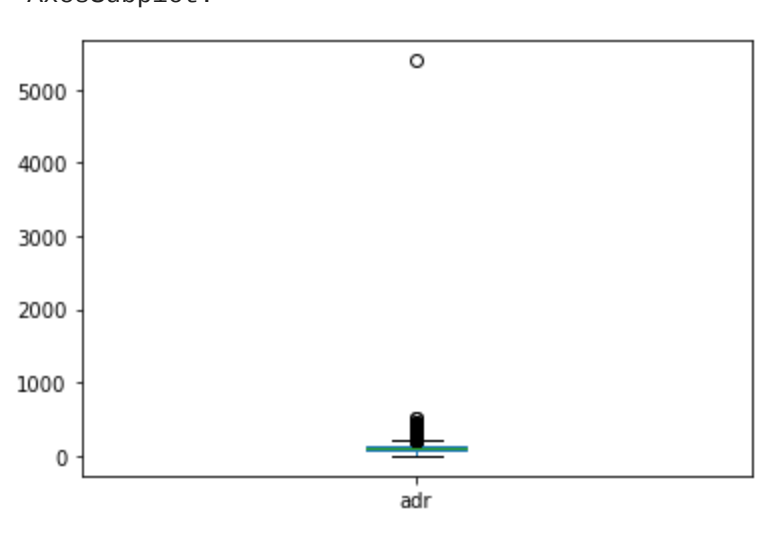
In [26]:
```
df.describe()
```

Out[26]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.00000
mean	0.371352	104.311435	2016.157656	27.166555	15.800880	0.928897	2.502145	1.858391	0.10420
std	0.483168	106.903309	0.707459	13.589971	8.780324	0.996216	1.900168	0.578576	0.39917
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.00000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.00000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.00000
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.00000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	41.000000	55.000000	10.00000

In [27]:
```
df['adr'].plot(kind='box')
```

Out[27]: <AxesSubplot:>



In [28]:
```
df=df[df['adr']<5000]
```

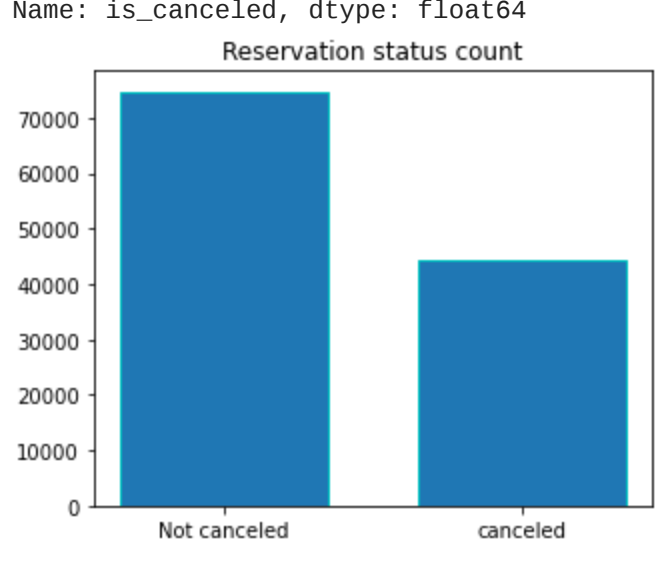
Data analysis and Visualization

In [32]:
```
cancelled_perc= df['is_canceled'].value_counts(normalize = True)
cancelled_perc
```

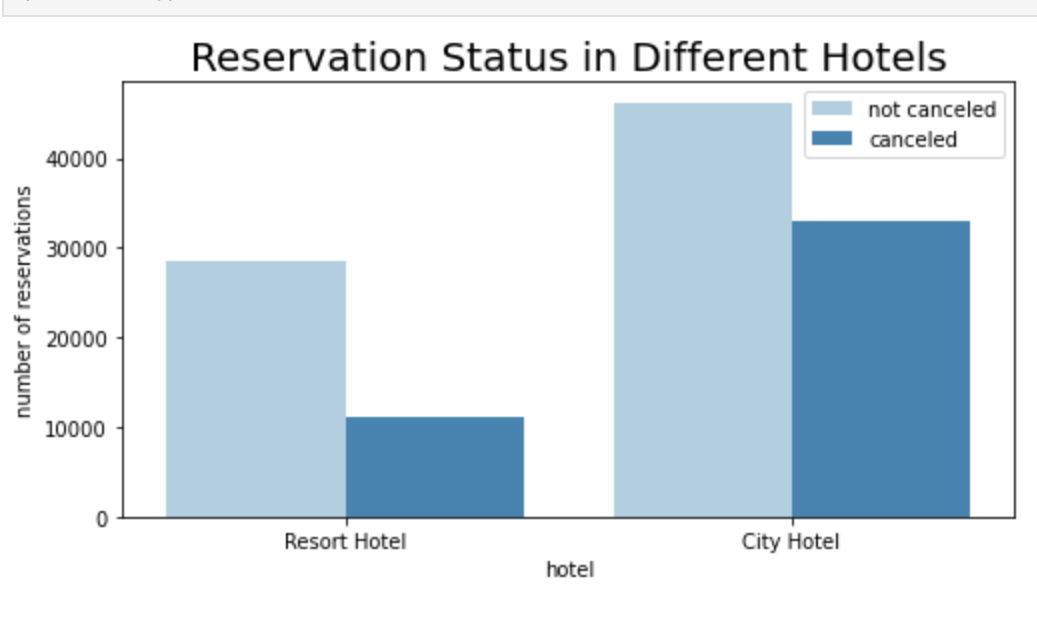
Out[32]:
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

In [33]:
```
cancelled_perc= df['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)

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

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



In [42]:
```
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(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 [48]:
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

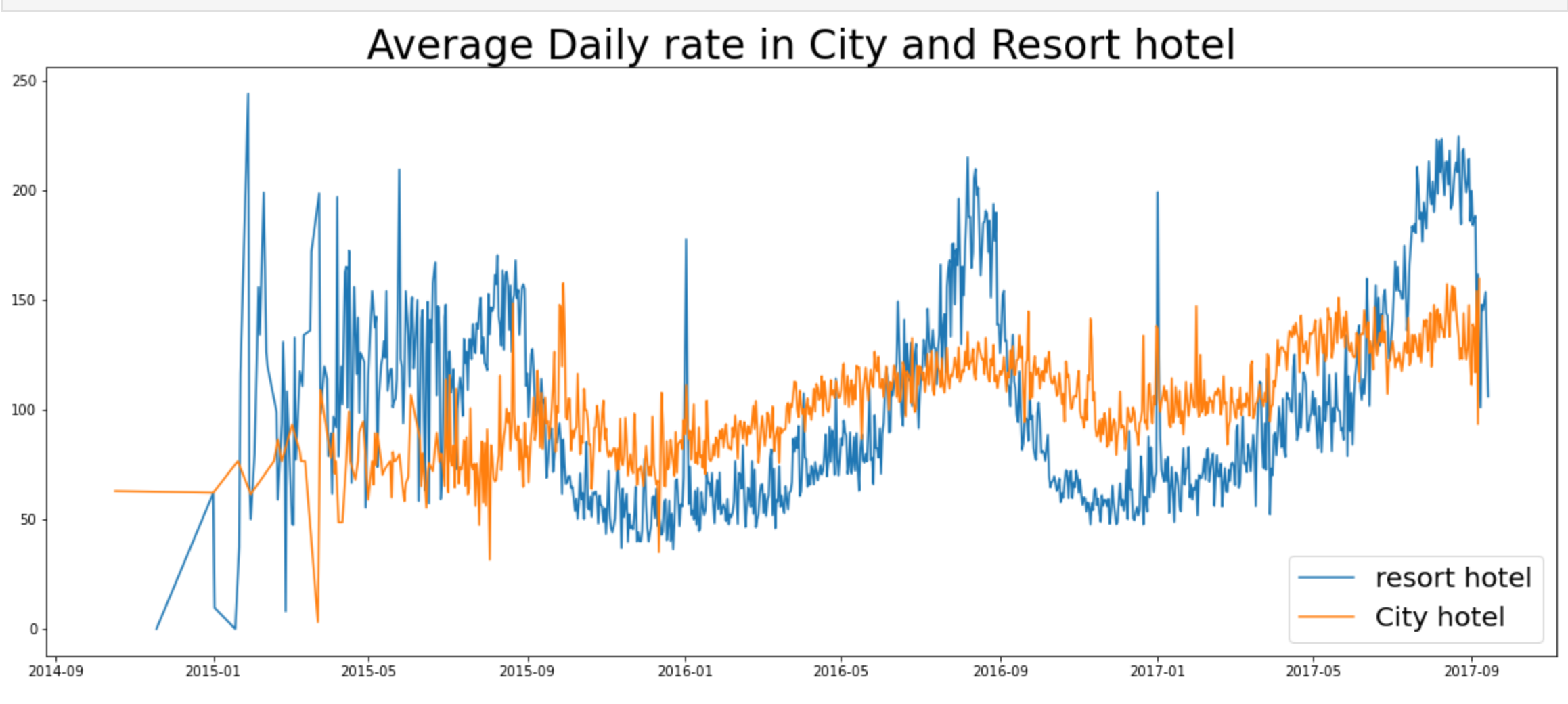
Out[48]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

In [50]:
```
City_hotel = df[df['hotel'] == 'City Hotel']
City_hotel['is_canceled'].value_counts(normalize = True)
```

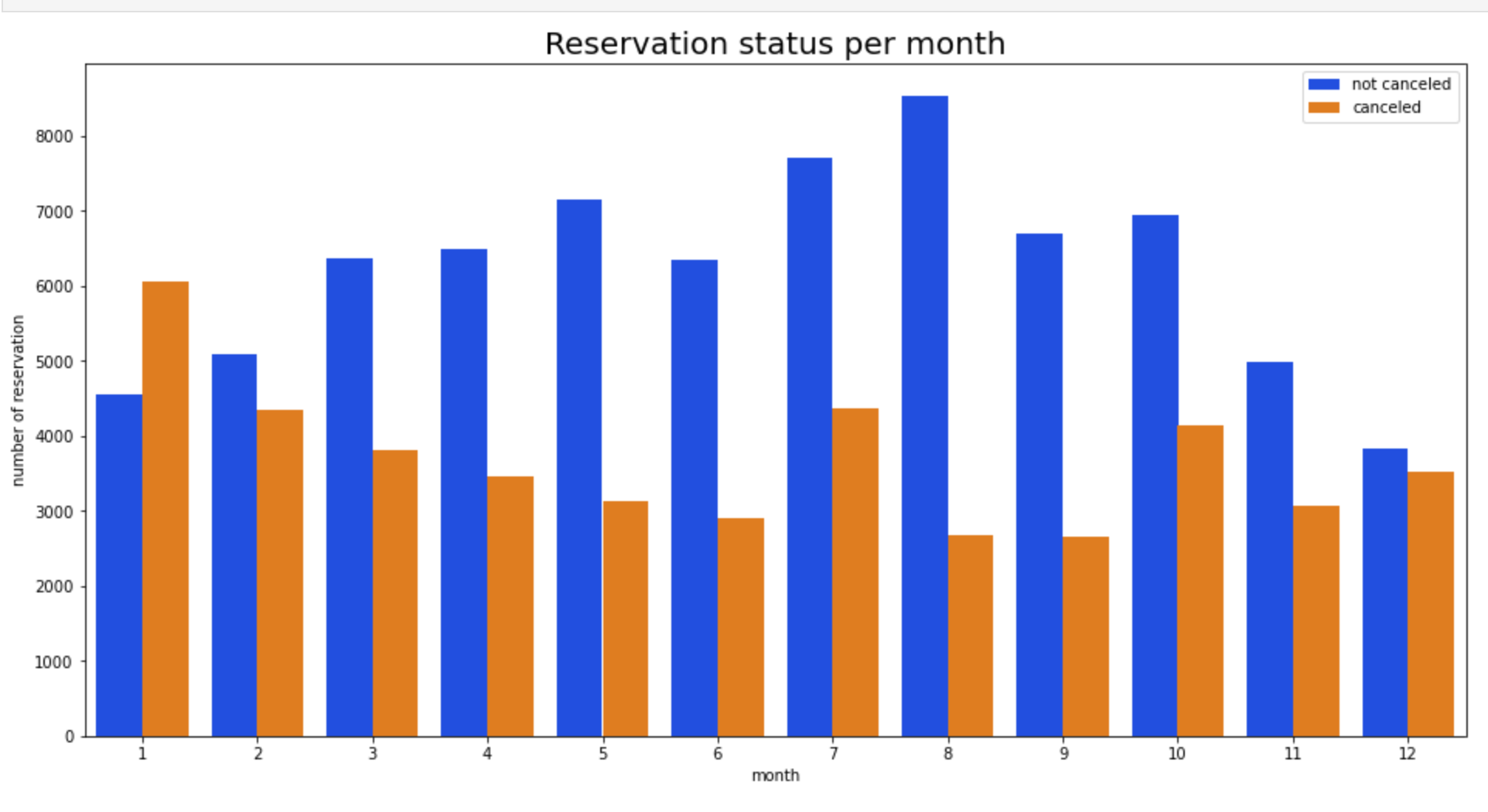
Out[50]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [55]:
```
resort_hotel= resort_hotel.groupby('reservation_status_date')[['adr']].mean()
City_hotel= City_hotel.groupby('reservation_status_date')[['adr']].mean()
```

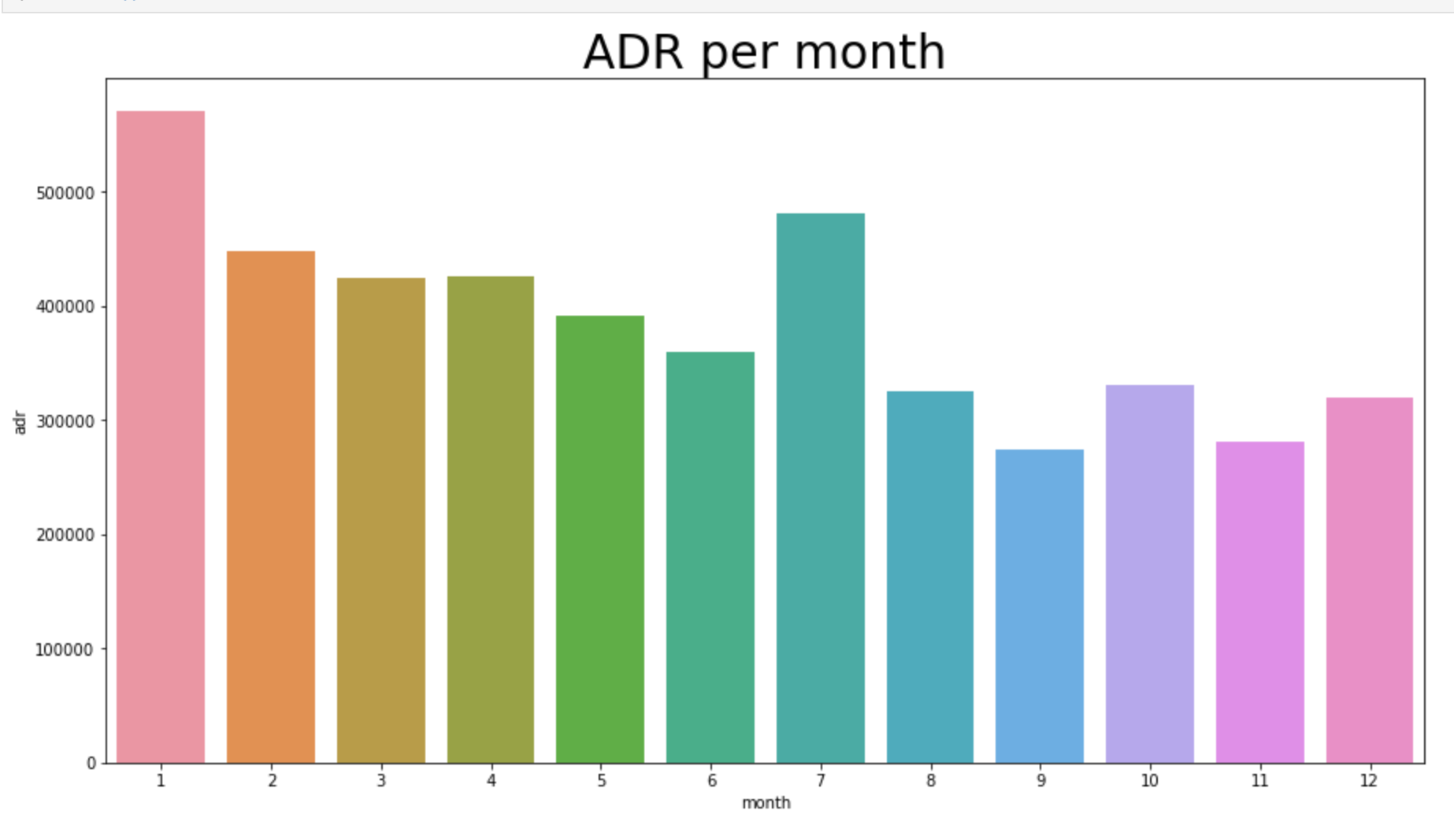
In [56]:
```
plt.figure(figsize=(20,8))
plt.title('Average Daily rate in City and Resort hotel', fontsize=30)
plt.plot(resort_hotel.index, resort_hotel['adr'] , label= 'resort hotel')
plt.plot(City_hotel.index, City_hotel['adr'] , label= 'City hotel')
plt.legend(fontsize=20)
plt.show()
```



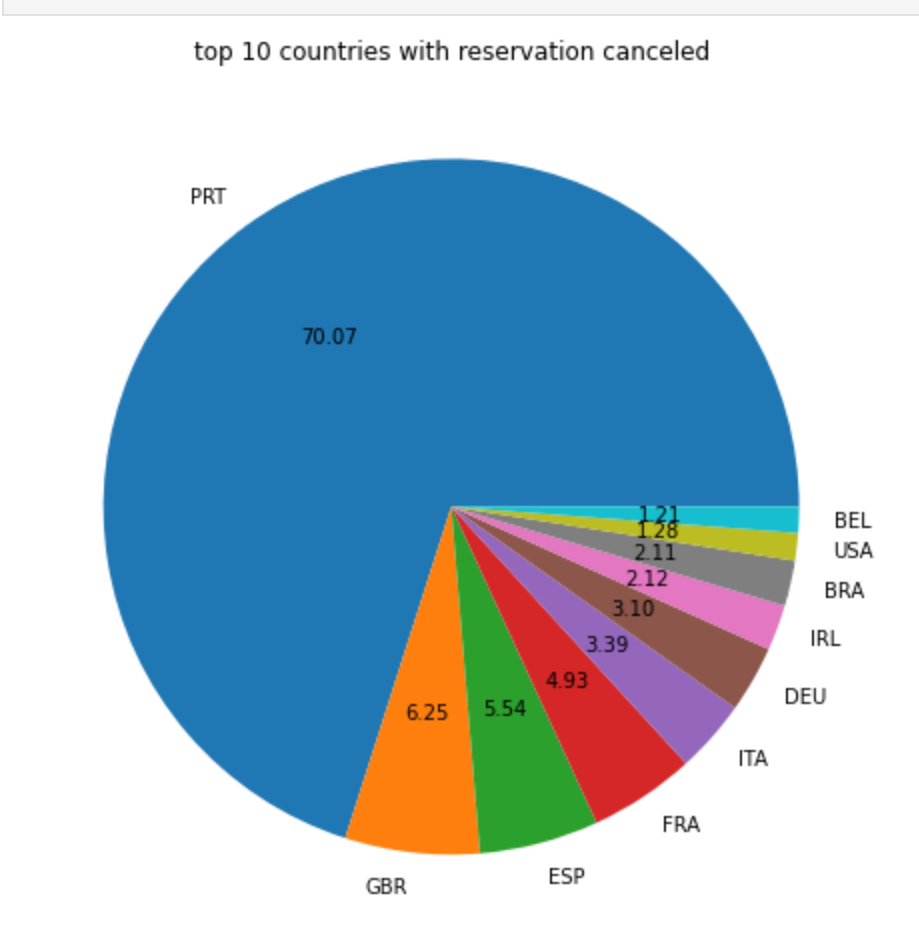
In [57]:
```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize= (16,8))
ax1=sns.countplot(x= 'month', hue= 'is_canceled', data=df, palette='bright')
legend_labels,_=ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per month', size=20)
plt.xlabel('month')
plt.ylabel('number of reservation')
plt.legend(['not canceled','canceled'])
plt.show()
```



In [59]:
```
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 [62]:
```
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 [64]:
```
df['market_segment'].value_counts()
```

Out[64]:
```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

In [65]:
```
df['market_segment'].value_counts(normalize=True)
```

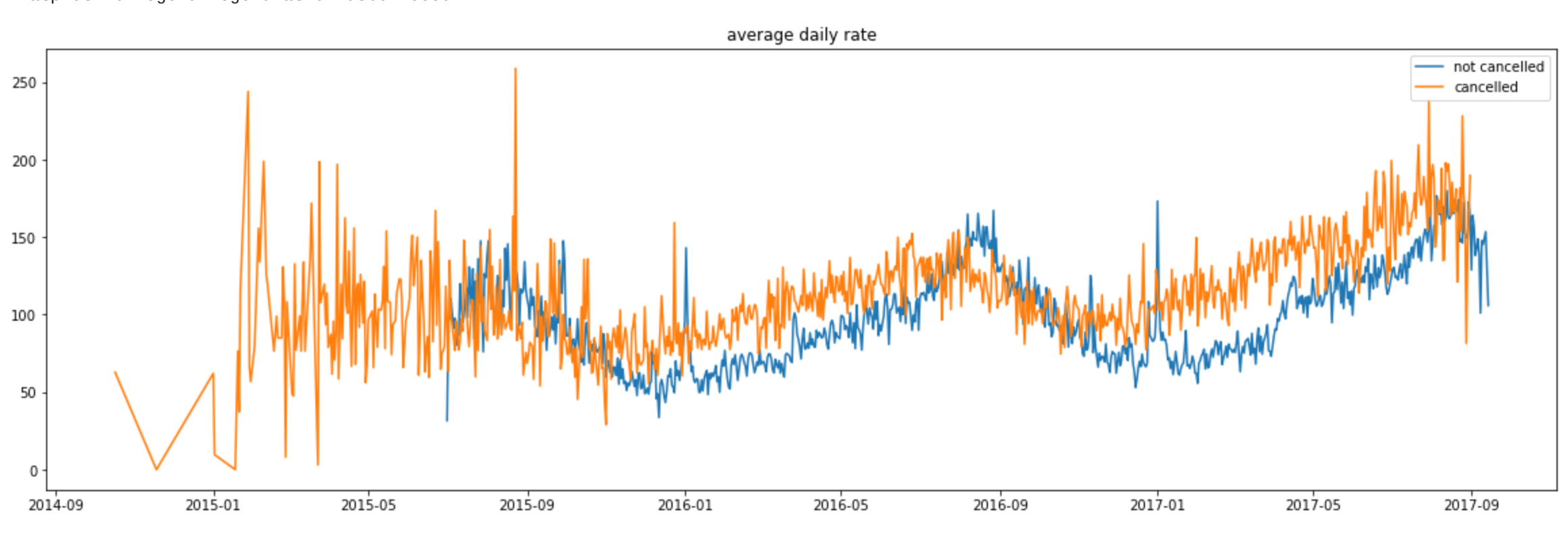
Out[65]:
```
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: market_segment, dtype: float64
```

In [66]:
```
cancelled_data['market_segment'].value_counts(normalize=True)
```

Out[66]:
```
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
```

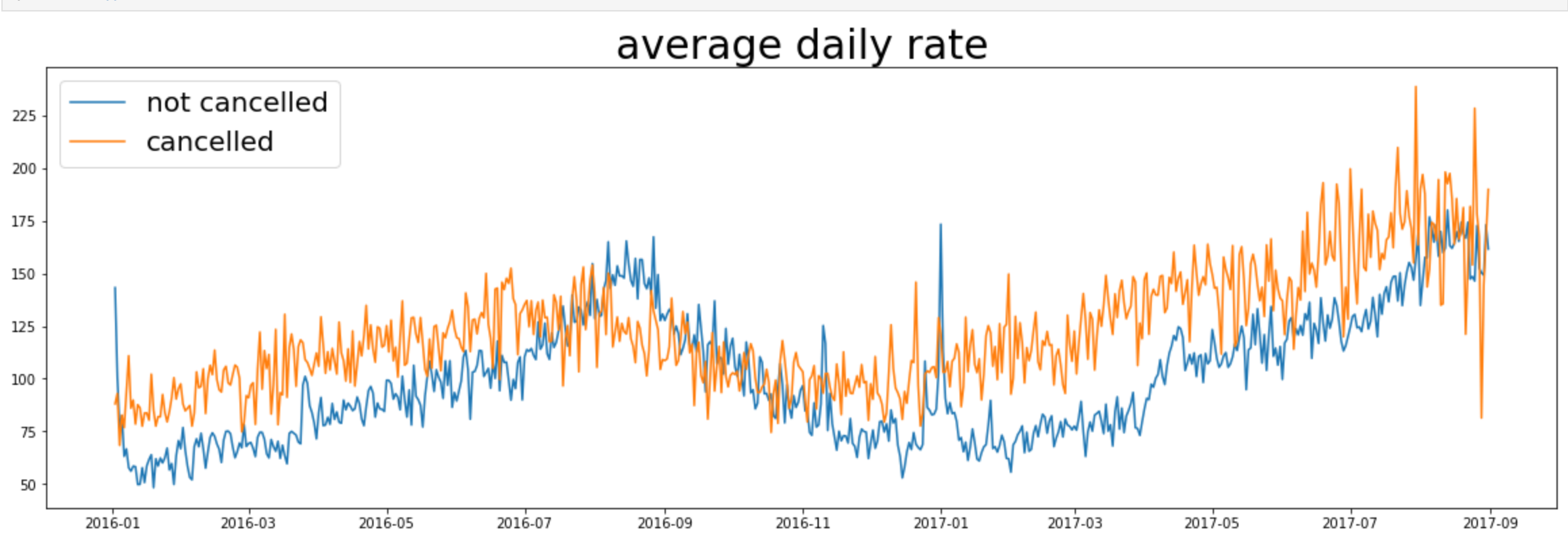
In [68]:
```
cancelled_df_adr= cancelled_data.groupby('reservation_status_date')[['adr']].mean()
cancelled_df_adr.reset_index(inplace=True)
cancelled_df_adr.sort_values('reservation_status_date', inplace=True)

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')
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[68]: <matplotlib.legend.Legend at 0x1d3b677cee0>



In [70]:
```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] > '2016-01-01') & (cancelled_df_adr['reservation_status_date'] < '
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date'] > '2016-01-01') & (not_cancelled_df_adr['reservation_s
```

In [74]:
```
plt.figure(figsize=(20,6))
plt.title('average daily rate', fontsize=30)
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(fontsize=20)
plt.show()
```



In [ ]: