Importing Libraries

In [39]:

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

Loading the dataset

In [ ]:

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

Exploratory Data Analysis and Data Cleaning

In [51]:

```
df.head()
```

Out[51]:

	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 × 31 columns

In [52]:

```
df.tail()
```

Out[52]:

	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 × 31 columns

In [53]:

```
df.shape
```

Out[53]:

```
(118897, 31)
```

In [54]:

```
df.columns
```

Out[54]:

```
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',
       'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'month'],
      dtype='object')
```

In [55]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 118897 entries, 0 to 119389
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  int64
 2   lead_time                       118897 non-null  int64
 3   arrival_date_year               118897 non-null  int64
 4   arrival_date_month              118897 non-null  object
 5   arrival_date_week_number        118897 non-null  int64
 6   arrival_date_day_of_month       118897 non-null  int64
 7   stays_in_weekend_nights         118897 non-null  int64
 8   stays_in_week_nights            118897 non-null  int64
 9   adults                          118897 non-null  int64
 10  children                        118897 non-null  float64
 11  babies                          118897 non-null  int64
 12  meal                            118897 non-null  object
 13  country                         118897 non-null  object
 14  market_segment                  118897 non-null  object
 15  distribution_channel            118897 non-null  object
 16  is_repeated_guest               118897 non-null  int64
 17  previous_cancellations          118897 non-null  int64
 18  previous_bookings_not_canceled  118897 non-null  int64
 19  reserved_room_type              118897 non-null  object
 20  assigned_room_type              118897 non-null  object
 21  booking_changes                 118897 non-null  int64
 22  deposit_type                    118897 non-null  object
 23  days_in_waiting_list            118897 non-null  int64
 24  customer_type                   118897 non-null  object
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64
 27  total_of_special_requests       118897 non-null  int64
 28  reservation_status              118897 non-null  object
 29  reservation_status_date         118897 non-null  datetime64[ns]
 30  month                           118897 non-null  int64
dtypes: datetime64[ns](1), float64(2), int64(17), object(11)
memory usage: 29.0+ MB
```

In [56]:

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

In [57]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 118897 entries, 0 to 119389
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  int64
 2   lead_time                       118897 non-null  int64
 3   arrival_date_year               118897 non-null  int64
 4   arrival_date_month              118897 non-null  object
 5   arrival_date_week_number        118897 non-null  int64
 6   arrival_date_day_of_month       118897 non-null  int64
 7   stays_in_weekend_nights         118897 non-null  int64
 8   stays_in_week_nights            118897 non-null  int64
 9   adults                          118897 non-null  int64
 10  children                        118897 non-null  float64
 11  babies                          118897 non-null  int64
 12  meal                            118897 non-null  object
 13  country                         118897 non-null  object
 14  market_segment                  118897 non-null  object
 15  distribution_channel            118897 non-null  object
 16  is_repeated_guest               118897 non-null  int64
 17  previous_cancellations          118897 non-null  int64
 18  previous_bookings_not_canceled  118897 non-null  int64
 19  reserved_room_type              118897 non-null  object
 20  assigned_room_type              118897 non-null  object
 21  booking_changes                 118897 non-null  int64
 22  deposit_type                    118897 non-null  object
 23  days_in_waiting_list            118897 non-null  int64
 24  customer_type                   118897 non-null  object
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64
 27  total_of_special_requests       118897 non-null  int64
 28  reservation_status              118897 non-null  object
 29  reservation_status_date         118897 non-null  datetime64[ns]
 30  month                           118897 non-null  int64
dtypes: datetime64[ns](1), float64(2), int64(17), object(11)
memory usage: 29.0+ MB
```

In [58]:

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

Out[58]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reser
count	118897	118897	118897	118897	118897	118897	
unique	2	12	5	177	7	5	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	
freq	79301	13852	91862	48585	56402	97729	

In [59]:

```
for col in df.describe(include='object').columns:
    print(col)
    print(df[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' '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'
 'Aviation']
--------------------------------------------------
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
--------------------------------------------------
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'B' 'P']
--------------------------------------------------
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'L' 'K' 'P']
--------------------------------------------------
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
--------------------------------------------------
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
--------------------------------------------------
reservation_status
['Check-Out' 'Canceled' 'No-Show']
--------------------------------------------------
```

In [60]:

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

Out[60]:

```
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
month                             0
dtype: int64
```

In [64]:

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

In [62]:

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

Out[62]:

```
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
month                             0
dtype: int64
```

In [63]:

```
df.describe()
```

Out[63]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_
count	118897.000000	118897.000000	118897.000000	118897.000000	1188
mean	0.371347	104.312018	2016.157657	27.166674	
std	0.483167	106.903570	0.707462	13.589966	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	161.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

In [ ]:

```
#removing outliers
df=df[df['adr']<5000]
```

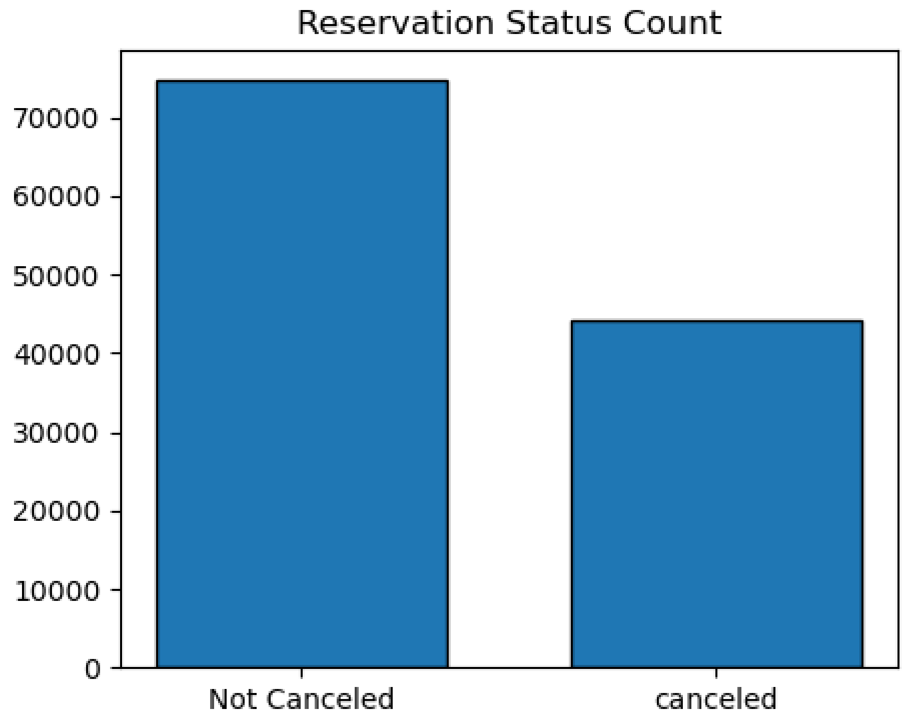
Data Analysis and Visualizations

In [40]:

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

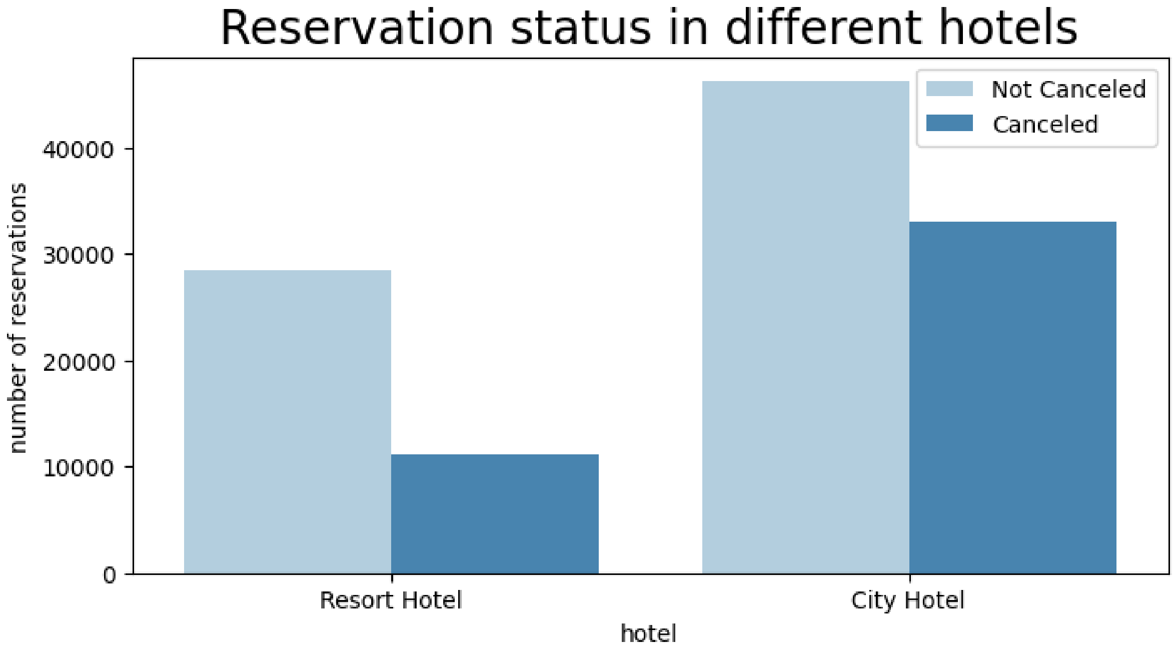
```
plt.figure(figsize = (5,4))
plt.title('Reservation Status Count')
plt.bar(['Not Canceled','canceled'],df['is_canceled'].value_counts(),edgecolor= 'k
plt.show()
```

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



In [50]:

```
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, ['Not Canceled', 'Canceled'], bbox_to_anchor=(1, 1))
plt.title('Reservation status in different hotels',size=20)
plt.xlabel('hotel')
plt.ylabel('number of reservations')
plt.show()
```



In [43]:

```
#Finding out no.of reservations and no.of cancelations in Resort Hotel
resort_hotel = df[df['hotel'] == 'Resort Hotel']
```

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

Out[43]:

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

In [44]:

```
#Finding out no.of reservations and no.of cancelations in City Hotel
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

Out[44]:

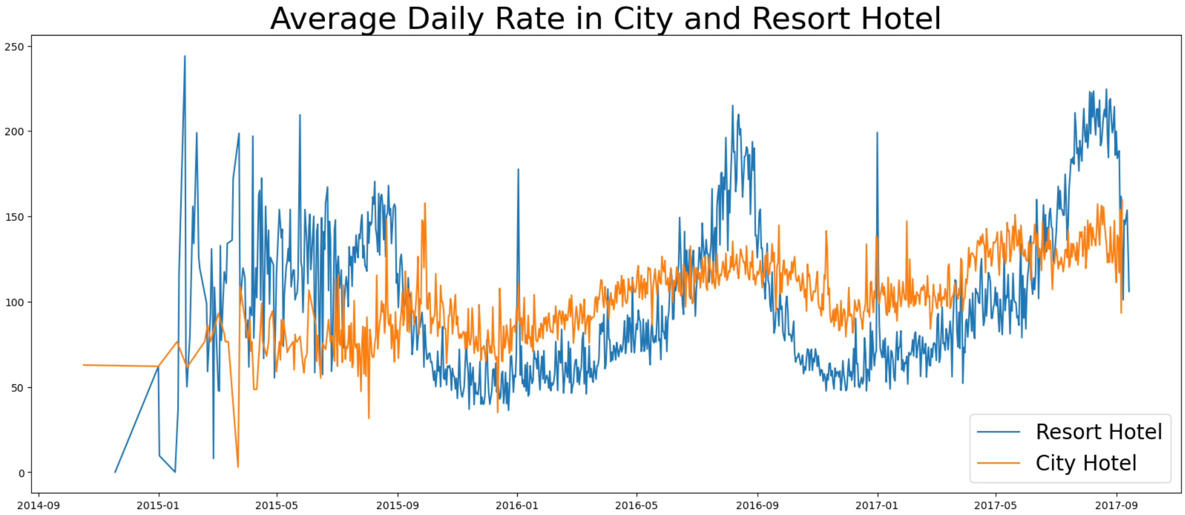
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [45]:

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

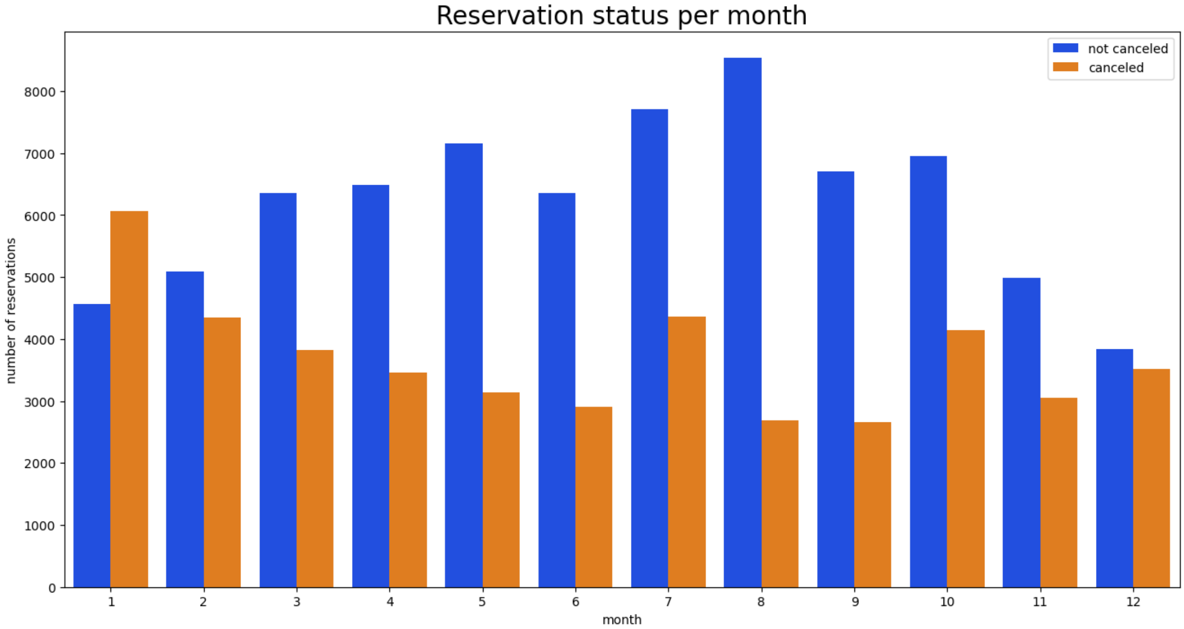
In [46]:

```
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()
```

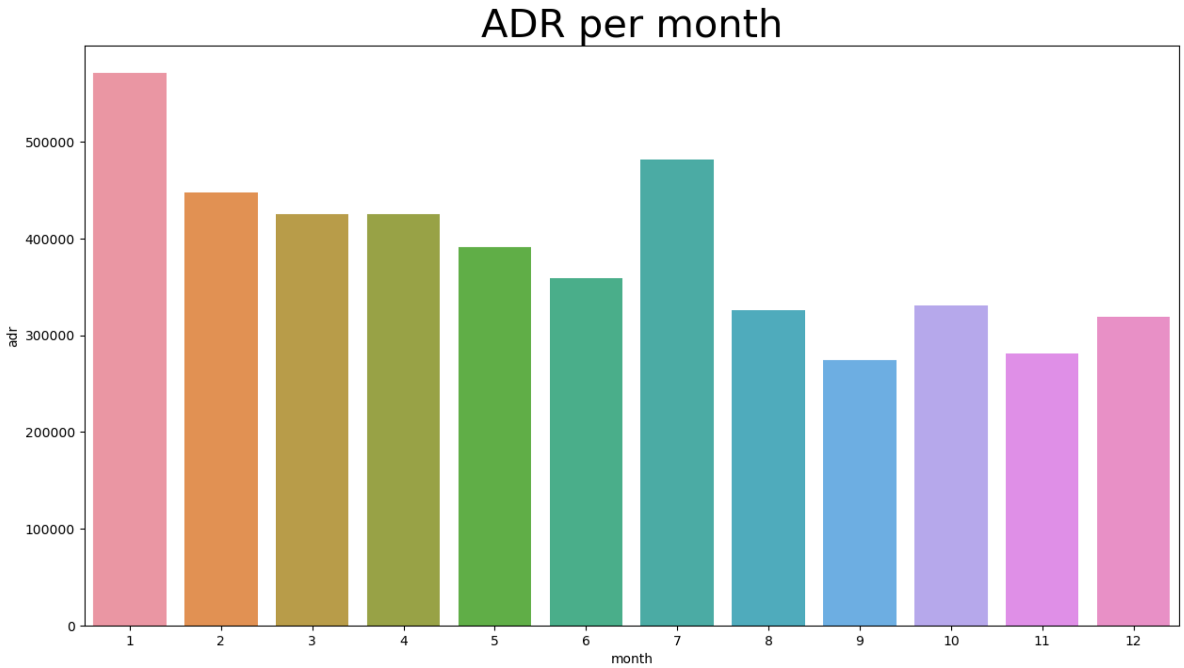


In [48]:

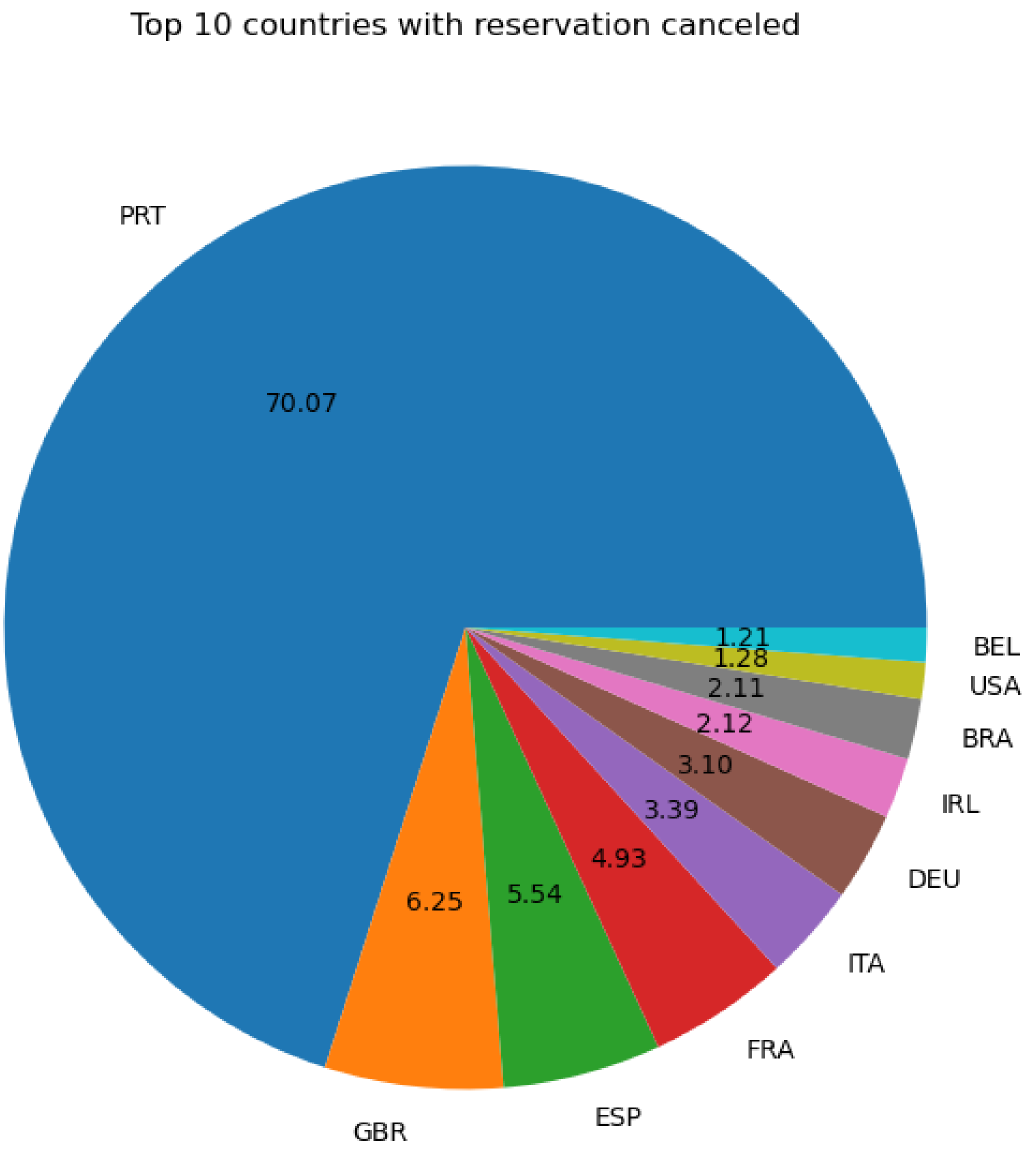
```
#Reservations per month
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize= (16,8))
ax1= sns.countplot(x='month',hue = 'is_canceled',data = df,palette= 'bright')
legend_labels,_=ax1.get_legend_handles_labels()
ax1.legend(legend_labels, ['Not Canceled', 'Canceled'], bbox_to_anchor=(1, 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 [66]:
```
#Average daily rate per each month
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()
```



In [69]:
```
#Top 10 countries with reservation canceled
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 [70]: df['market_segment'].value_counts()
```

```
Out[70]: Online TA        56402
         Offline TA/TO    24159
         Groups           19806
         Direct           12448
         Corporate         5111
         Complementary      734
         Aviation           237
         Name: market_segment, dtype: int64
```

```
In [71]: df['market_segment'].value_counts(normalize = True)
```

```
Out[71]: 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 [72]: cancelled_data['market_segment'].value_counts(normalize = True)
```

Out[72]:

```
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 [74]:

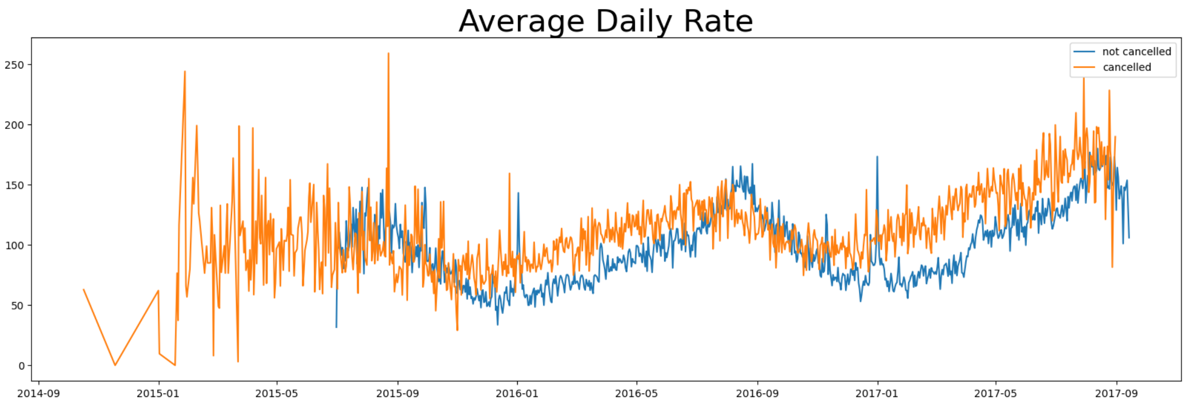
```
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[['adr']].mean
cancelled_df_adr.reset_index(inplace = True)
cancelled_df_adr.sort_values('reservation_status_date', inplace = True)

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',fontsize = 30)
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()
```

Out[74]:

```
<matplotlib.legend.Legend at 0x1c4a86ec5b0>
```



In [ ]: