# Importing Libraries

In [91]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

# Loading the dataset

In [92]:
```
df = pd.read_csv('hotel_bookings 2.csv')
```

# Exploratory Data Analysis and Data Cleaning

In [93]:
```
df.head()
```

Out[93]:

| | hotel | is_canceled | lead_time | arrival_date_year | arrival_date_month | arrival_date_week_number | arrival_date_day_of_month | stays_in_ |
|---|---|---|---|---|---|---|---|---|
| **0** | Resort Hotel | 0 | 342 | 2015 | July | 27 | 1 | |
| **1** | Resort Hotel | 0 | 737 | 2015 | July | 27 | 1 | |
| **2** | Resort Hotel | 0 | 7 | 2015 | July | 27 | 1 | |
| **3** | Resort Hotel | 0 | 13 | 2015 | July | 27 | 1 | |
| **4** | Resort Hotel | 0 | 14 | 2015 | July | 27 | 1 | |

5 rows × 32 columns

In [94]:
```
df.tail()
```

Out[94]:

| | hotel | is_canceled | lead_time | arrival_date_year | arrival_date_month | arrival_date_week_number | arrival_date_day_of_month | stays |
|---|---|---|---|---|---|---|---|---|
| **119385** | City Hotel | 0 | 23 | 2017 | August | 35 | 30 | |
| **119386** | City Hotel | 0 | 102 | 2017 | August | 35 | 31 | |
| **119387** | City Hotel | 0 | 34 | 2017 | August | 35 | 31 | |
| **119388** | City Hotel | 0 | 109 | 2017 | August | 35 | 31 | |
| **119389** | City Hotel | 0 | 205 | 2017 | August | 35 | 29 | |

5 rows × 32 columns

In [95]:
```
df.shape
```

Out[95]:
```
(119390, 32)
```

In [96]:
```
df.columns
```

Out[96]:
```
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 [97]:
```
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 [98]:
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'], format='%d/%m/%Y')
```

In [99]:
```
df.describe(include = 'object')
```

Out[99]:

| | hotel | arrival_date_month | meal | country | market_segment | distribution_channel | reserved_room_type | assigned_room_type | d |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 119390 | 119390 | 119390 | 118902 | 119390 | 119390 | 119390 | 119390 | |
| **unique** | 2 | 12 | 5 | 177 | 8 | 5 | 10 | 12 | |
| **top** | City Hotel | August | BB | PRT | Online TA | TA/TO | A | A | |
| **freq** | 79330 | 13877 | 92310 | 48590 | 56477 | 97870 | 85994 | 74053 | |

In [100]:
```
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' 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 [101]:
```
df.isnull().sum()
```

Out[101]:
```
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 [102]:
```
df.drop(['company', 'agent'], axis = 1, inplace = True)    # Droping the column with many null values
df.dropna(inplace = True)   # Droping the null rows
```

In [103]:
```
df.isnull().sum()
```

Out[103]:
```
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
```

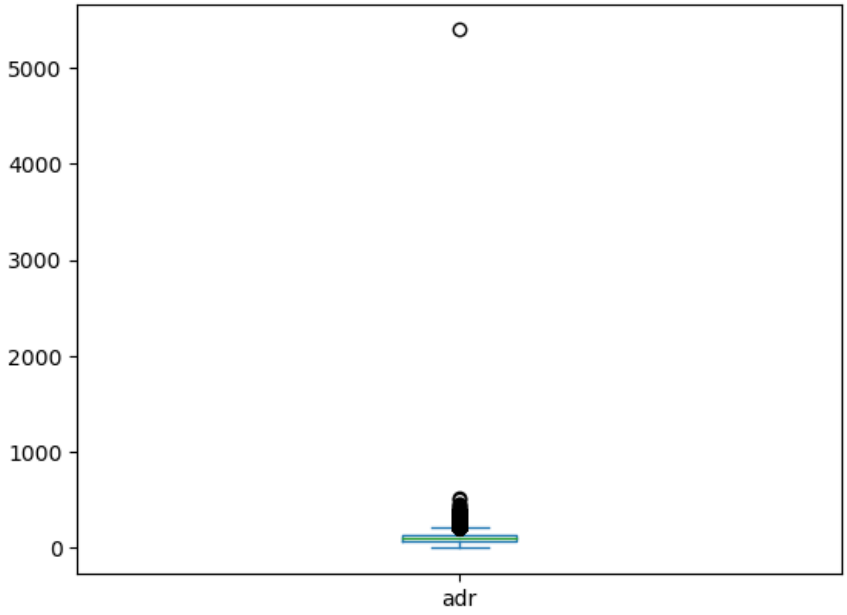
In [104]:
```
df.describe()
```

Out[104]:

| | is_canceled | lead_time | arrival_date_year | arrival_date_week_number | arrival_date_day_of_month | stays_in_weekend_nights |
|---|---|---|---|---|---|---|
| **count** | 118898.000000 | 118898.000000 | 118898.000000 | 118898.000000 | 118898.000000 | 118898.000000 |
| **mean** | 0.371352 | 104.311435 | 2016.157656 | 27.166555 | 15.800880 | 0.928897 |
| **min** | 0.000000 | 0.000000 | 2015.000000 | 1.000000 | 1.000000 | 0.000000 |
| **25%** | 0.000000 | 18.000000 | 2016.000000 | 16.000000 | 8.000000 | 0.000000 |
| **50%** | 0.000000 | 69.000000 | 2016.000000 | 28.000000 | 16.000000 | 1.000000 |
| **75%** | 1.000000 | 161.000000 | 2017.000000 | 38.000000 | 23.000000 | 2.000000 |
| **max** | 1.000000 | 737.000000 | 2017.000000 | 53.000000 | 31.000000 | 16.000000 |
| **std** | 0.483168 | 106.903309 | 0.707459 | 13.589971 | 8.780324 | 0.996216 |

In [105]:
```
df['adr'].plot(kind = 'box') # 1 point is greater than other points so this is outlier
```

Out[105]: <Axes: >




In [106]:
```
df = df[df['adr']<5000]
```

In [107]:
```
df.describe()
```

Out[107]:

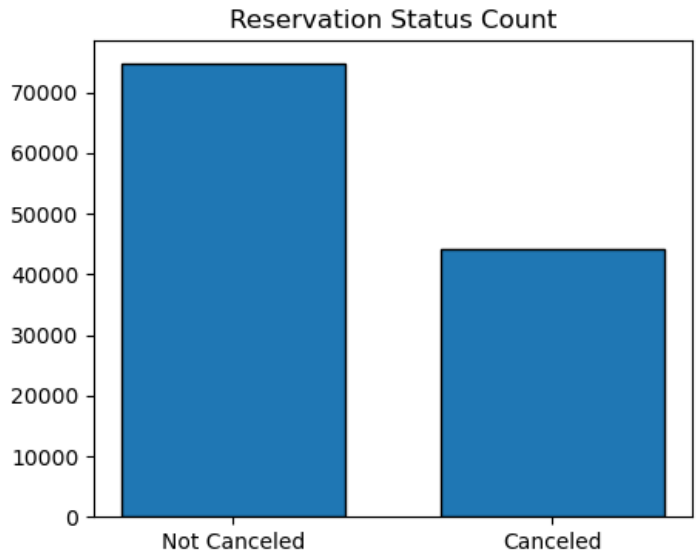
| | is_canceled | lead_time | arrival_date_year | arrival_date_week_number | arrival_date_day_of_month | stays_in_weekend_nights |
|---|---|---|---|---|---|---|
| **count** | 118897.000000 | 118897.000000 | 118897.000000 | 118897.000000 | 118897.000000 | 118897.000000 |
| **mean** | 0.371347 | 104.312018 | 2016.157657 | 27.166674 | 15.800802 | 0.928905 |
| **min** | 0.000000 | 0.000000 | 2015.000000 | 1.000000 | 1.000000 | 0.000000 |
| **25%** | 0.000000 | 18.000000 | 2016.000000 | 16.000000 | 8.000000 | 0.000000 |
| **50%** | 0.000000 | 69.000000 | 2016.000000 | 28.000000 | 16.000000 | 1.000000 |
| **75%** | 1.000000 | 161.000000 | 2017.000000 | 38.000000 | 23.000000 | 2.000000 |
| **max** | 1.000000 | 737.000000 | 2017.000000 | 53.000000 | 31.000000 | 16.000000 |
| **std** | 0.483167 | 106.903570 | 0.707462 | 13.589966 | 8.780321 | 0.996217 |

# Data Analysis and Visualizations

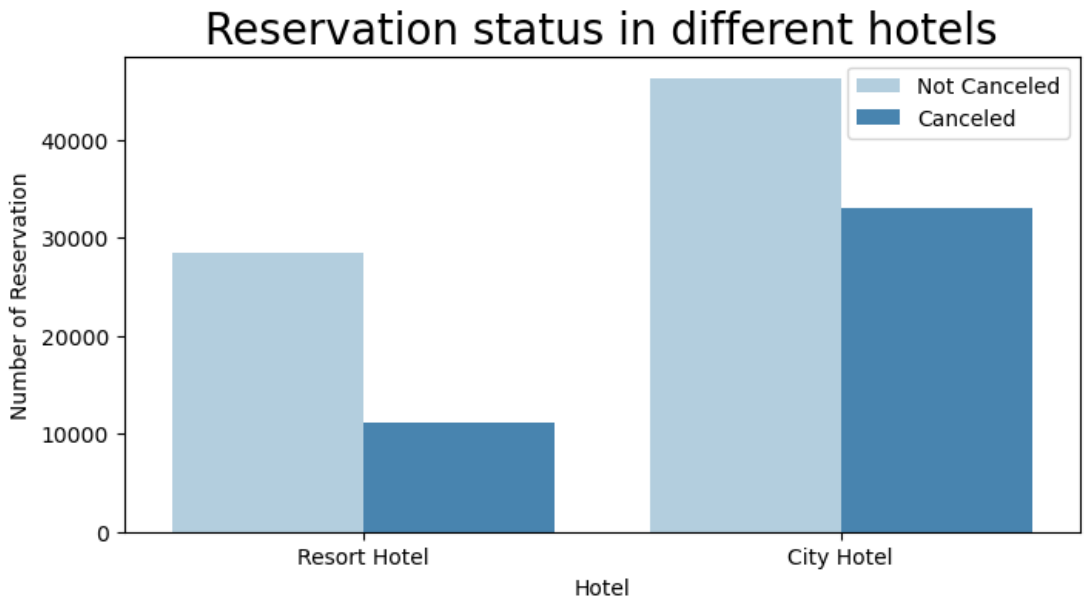
In [108]:
```
canceled_perc = df['is_canceled'].value_counts(normalize = True)    # normalize = True - Returns percentize
print(canceled_perc)


plt.figure(figsize = (5, 4))
plt.title("Reservation Status Count")
plt.bar(('Not Canceled', 'Canceled'), df['is_canceled'].value_counts(), edgecolor = 'k', width = 0.7)
```

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

Out[108]:
```
<BarContainer object of 2 artists>
```



In [109]:
```
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(('Not Canceled', 'Canceled'), bbox_to_anchor = (1, 1))
plt.title('Reservation status in different hotels', size = 20)
plt.xlabel('Hotel')
plt.ylabel('Number of Reservation')
plt.show()
```




In [110]:
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

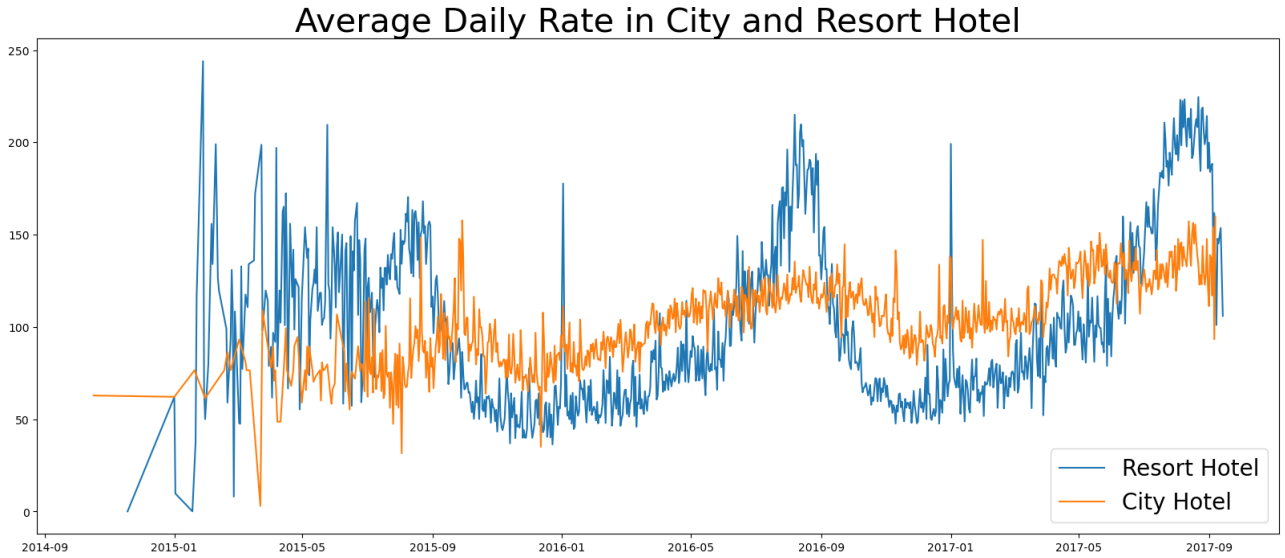
Out[110]:
```
is_canceled
0    0.72025
1    0.27975
Name: proportion, dtype: float64
```

In [111]:
```
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

Out[111]:
```
is_canceled
0    0.582918
1    0.417082
Name: proportion, dtype: float64
```

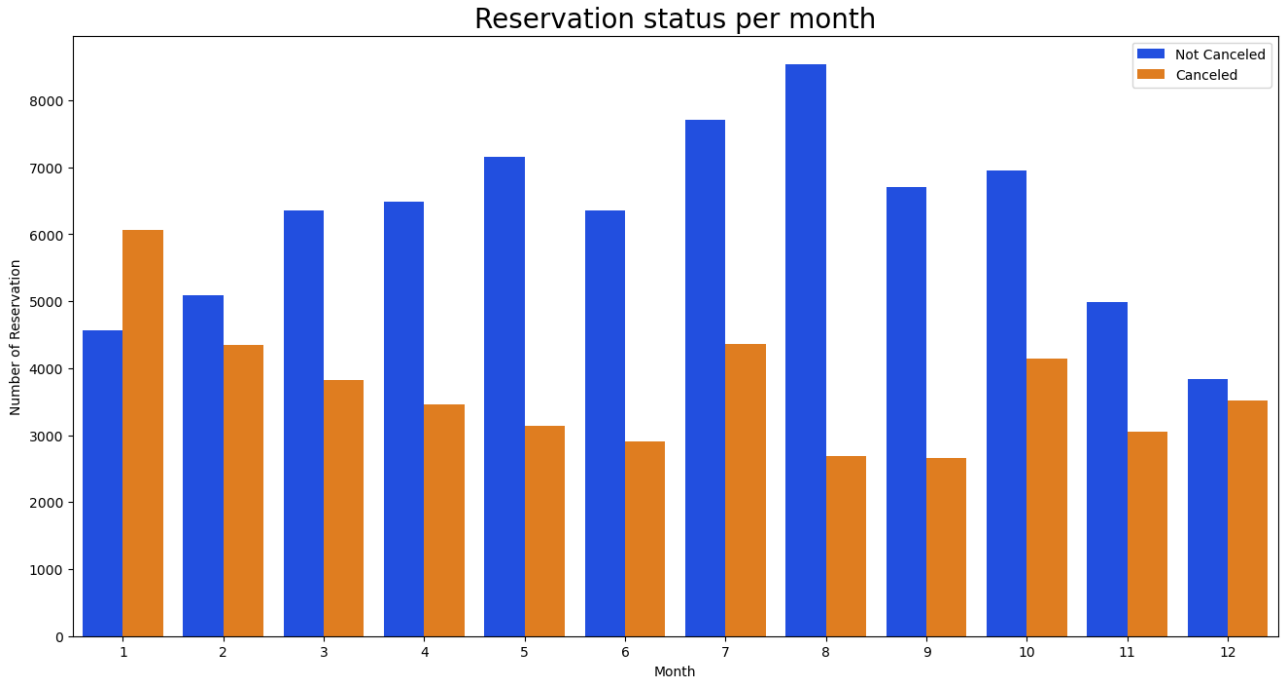
In [112]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

In [113]:
```
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()
```

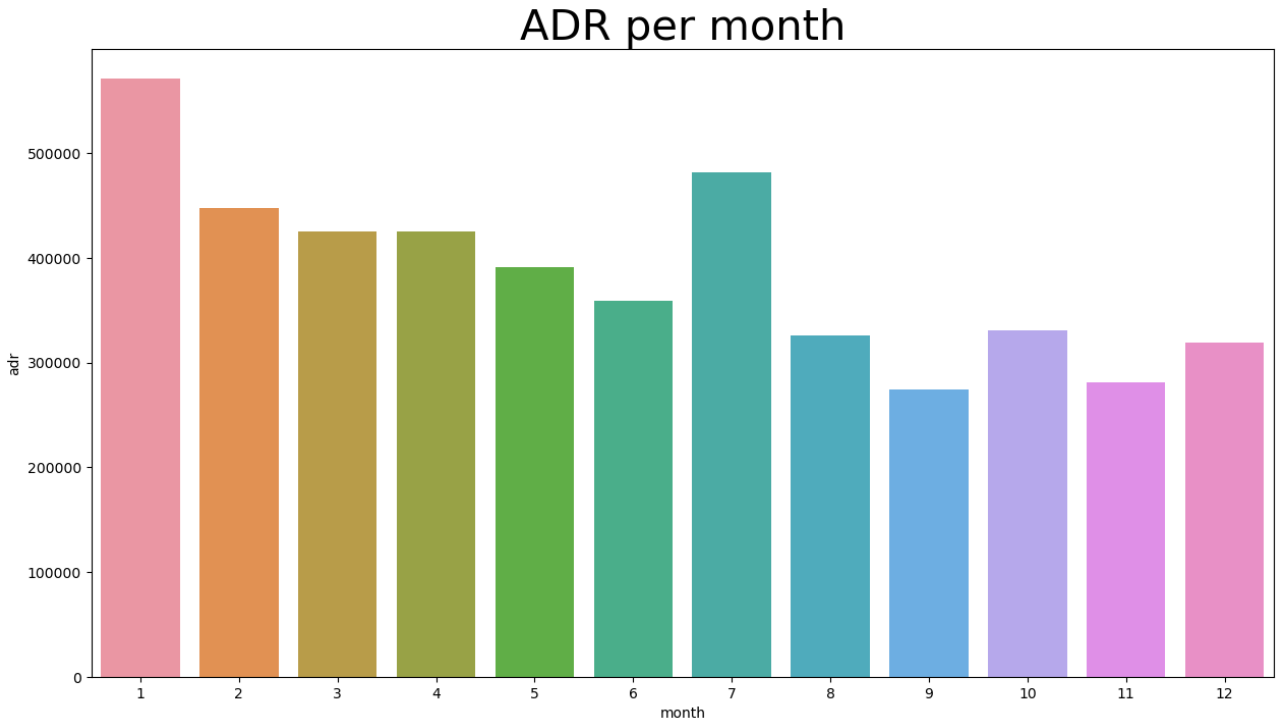



In [114]:
```
df['month'] = df['reservation_status_date'].dt.month
```

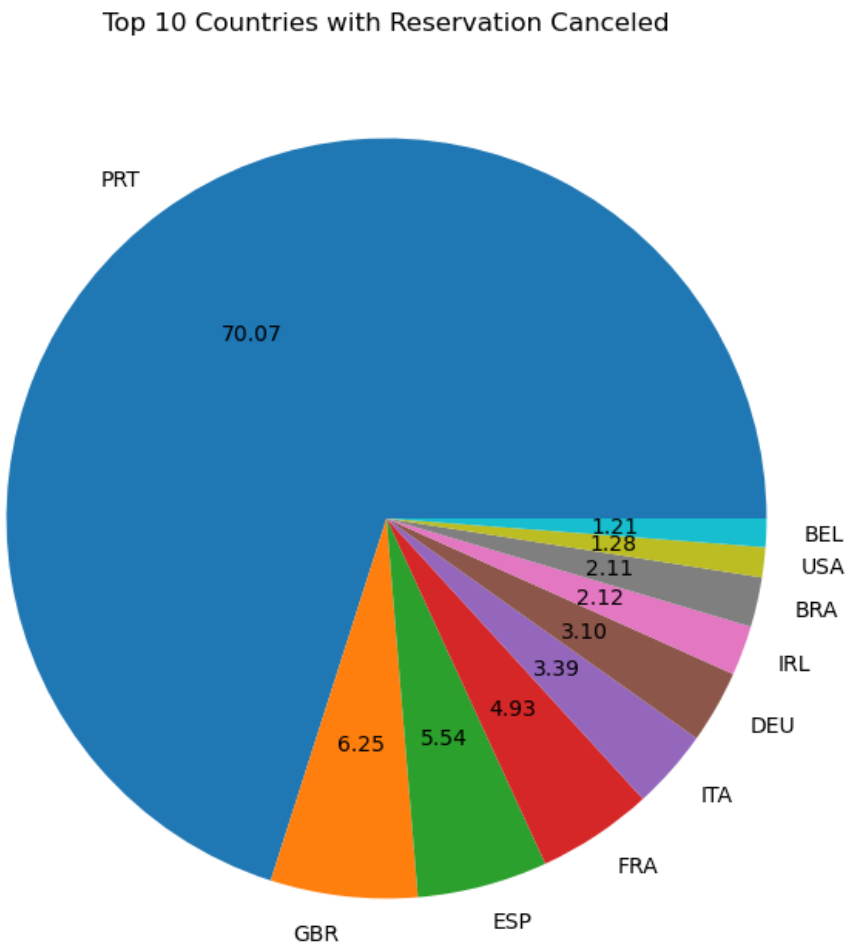
In [115]:
```
plt.figure(figsize = (16, 8))
ax1 = sns.countplot(x = 'month', hue = 'is_canceled', data = df, palette = 'bright')
legend_labels,_ = ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor = (1, 1))
plt.title('Reservation status per month', size = 20)
plt.xlabel('Month')
plt.ylabel('Number of Reservation')
plt.legend(['Not Canceled', 'Canceled'])
plt.show()
```



In [116]:
```
plt.figure(figsize = (15, 8))
plt.title('ADR per month', fontsize = 30)
sns.barplot(x = 'month', y = 'adr', data = df[df['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_i
plt.show()
```



In [117]:
```
canceled_data = df[df['is_canceled'] == 1]
top_10_country = canceled_data['country'].value_counts()[:10]  # Return country in decending order, [:10] - 1
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 [118]:
```
df['market_segment'].value_counts()
```

Out[118]:
```
market_segment
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: count, dtype: int64
```

In [119]:
```
df['market_segment'].value_counts(normalize = True)
```

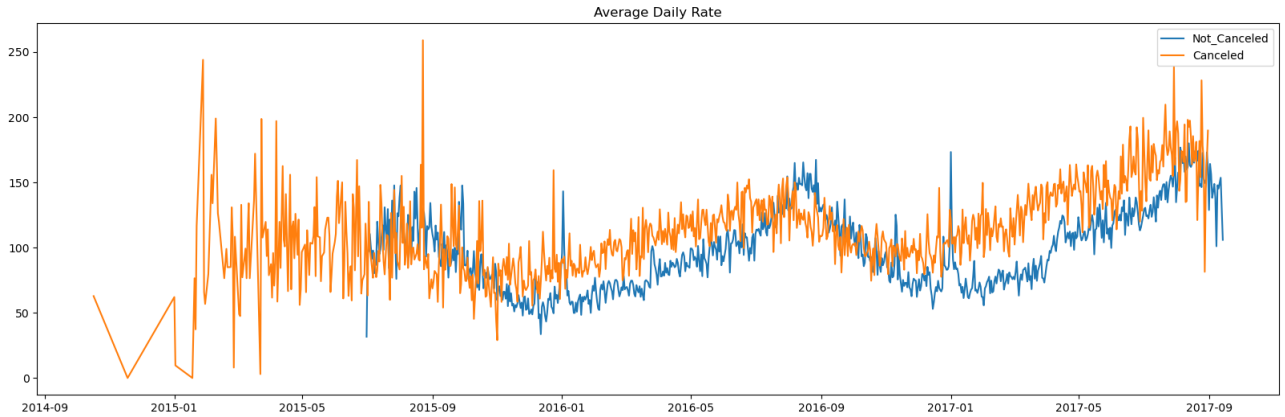
Out[119]:
```
market_segment
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: proportion, dtype: float64
```

In [120]:
```
canceled_data['market_segment'].value_counts(normalize = True)
```

Out[120]:
```
market_segment
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: proportion, dtype: float64
```

In [121]:
```
not_canceled_data = df[df['is_canceled'] == 0]
```

In [122]:
```
canceled_df_adr = canceled_data.groupby('reservation_status_date')[['adr']].mean()
canceled_df_adr.reset_index(inplace = True)
canceled_df_adr.sort_values('reservation_status_date', inplace = True)

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

plt.figure(figsize = (20, 6))
plt.title('Average Daily Rate')
plt.plot(not_canceled_df_adr['reservation_status_date'], not_canceled_df_adr['adr'], label = 'Not_Canceled')
plt.plot(canceled_df_adr['reservation_status_date'], canceled_df_adr['adr'], label = 'Canceled')
plt.legend()
plt.show()
```




In [123]:
```
canceled_df_adr = canceled_df_adr[(canceled_df_adr['reservation_status_date'] > '2016') & (canceled_df_adr['r
not_canceled_df_adr = not_canceled_df_adr[(not_canceled_df_adr['reservation_status_date'] > '2016') & (not_ca
```

### Filtered data from 2016 to sep, 2017

In [124]:
```
plt.figure(figsize = (20, 6))
plt.title('Average Daily Rate', fontsize = 30)
plt.plot(not_canceled_df_adr['reservation_status_date'], not_canceled_df_adr['adr'], label = 'Not_Canceled')
plt.plot(canceled_df_adr['reservation_status_date'], canceled_df_adr['adr'], label = 'Canceled')
plt.legend(fontsize = 15)
plt.show()
```