import Libraries

In [64]:

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

Import Dataset

In [65]:

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

Exploratory Data analysis and Data cleaning

In [66]:

```
df.head()
```

Out[66]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_c
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 [67]:

```
df.tail()
```

Out[67]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arri
119385	City Hotel	0	23	2017	August	35	
119386	City Hotel	0	102	2017	August	35	
119387	City Hotel	0	34	2017	August	35	
119388	City Hotel	0	109	2017	August	35	
119389	City Hotel	0	205	2017	August	35	

5 rows × 32 columns

In [68]:

```
df.shape
```

Out[68]:

```
(119390, 32)
```

In [69]:

```
df.columns
```

Out[69]:

```
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 [70]:

```
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 [71]:

```
df.dtypes
```

Out[71]:

```
hotel                              object
is_canceled                         int64
lead_time                           int64
arrival_date_year                   int64
arrival_date_month                 object
arrival_date_week_number            int64
arrival_date_day_of_month           int64
stays_in_weekend_nights             int64
stays_in_week_nights                int64
adults                              int64
children                          float64
babies                              int64
meal                               object
country                            object
market_segment                     object
distribution_channel               object
is_repeated_guest                   int64
previous_cancellations              int64
previous_bookings_not_canceled      int64
reserved_room_type                 object
assigned_room_type                 object
booking_changes                     int64
deposit_type                       object
agent                             float64
company                           float64
days_in_waiting_list                int64
customer_type                      object
adr                               float64
required_car_parking_spaces         int64
total_of_special_requests           int64
reservation_status                 object
reservation_status_date            object
dtype: object
```

In [72]:

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

In [73]:

```
df.head(2)
```

Out[73]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_c
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	

2 rows × 32 columns

In [74]:

```
df.describe()
```

Out[74]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_mont
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.00000
mean	0.370416	104.011416	2016.156554	27.165173	15.79824
std	0.482918	106.863097	0.707476	13.605138	8.78082
min	0.000000	0.000000	2015.000000	1.000000	1.00000
25%	0.000000	18.000000	2016.000000	16.000000	8.00000
50%	0.000000	69.000000	2016.000000	28.000000	16.00000
75%	1.000000	160.000000	2017.000000	38.000000	23.00000
max	1.000000	737.000000	2017.000000	53.000000	31.00000

In [75]:

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

Out[75]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_
count	119390	119390	119390	118902	119390	119390	11
unique	2	12	5	177	8	5	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	
freq	79330	13877	92310	48590	56477	97870	8

In [76]:

```
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 [77]:

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

Out[77]:

```
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 [78]:

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

In [79]:

```
df.dropna(inplace=True)
```

In [80]:

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

Out[80]:

```
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 [81]:

```
df.describe()
```

Out[81]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_mont
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.00000
mean	0.371352	104.311435	2016.157656	27.166555	15.80088
std	0.483168	106.903309	0.707459	13.589971	8.78032
min	0.000000	0.000000	2015.000000	1.000000	1.00000
25%	0.000000	18.000000	2016.000000	16.000000	8.00000
50%	0.000000	69.000000	2016.000000	28.000000	16.00000
75%	1.000000	161.000000	2017.000000	38.000000	23.00000
max	1.000000	737.000000	2017.000000	53.000000	31.00000

In [82]:

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

Data Analysis And Visualization

In [83]:

```
df.head(2)
```

Out[83]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	

2 rows × 30 columns

In [84]:

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

Out[84]:

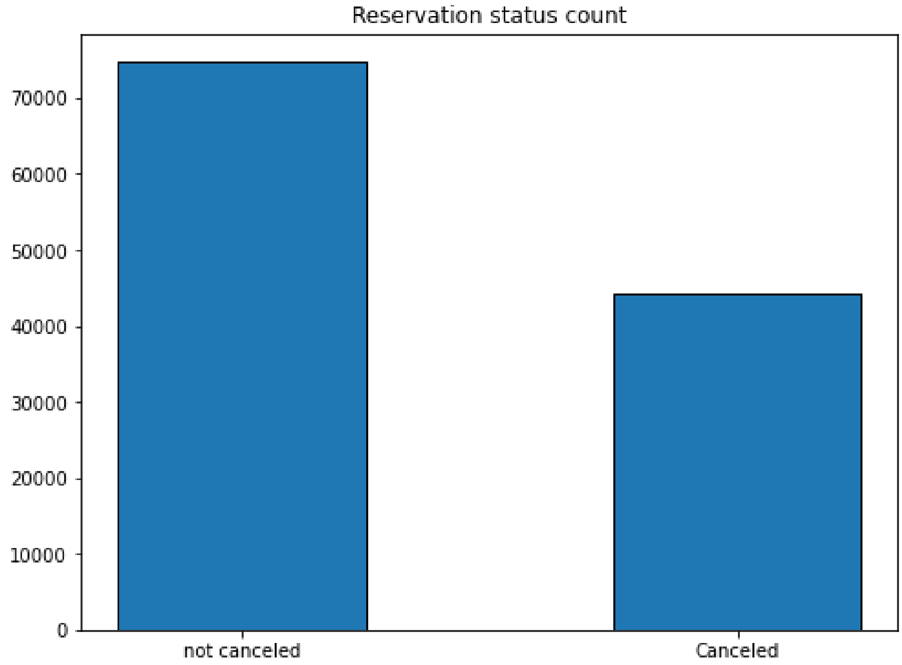
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

In [85]:

```
plt.figure(figsize=(8,6))
plt.title("Reservation status count")
plt.bar(['not canceled' , 'Canceled'],df['is_canceled'].value_counts(),edgecolor='k' , width=0.
```

Out[85]:

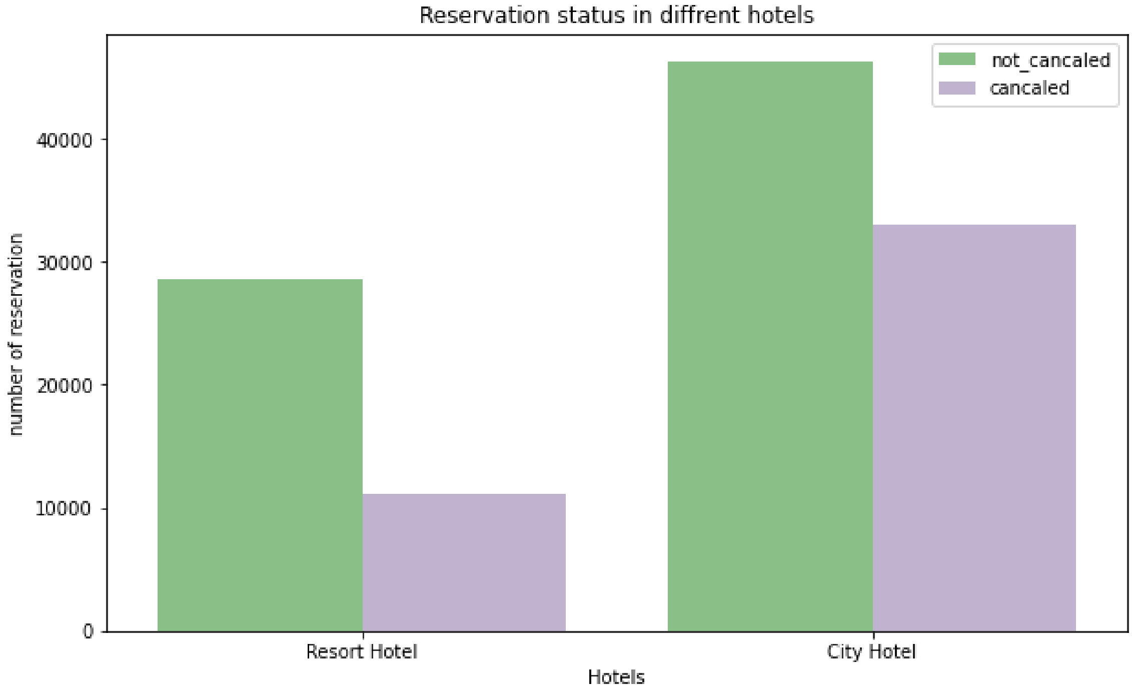
```
<BarContainer object of 2 artists>
```



In [86]:

```
plt.figure(figsize=(10,6))
sns.countplot(x="hotel",hue='is_canceled',data=df,palette='Accent')
plt.title("Reservation status in diffrent hotels")
plt.xlabel("Hotels")
plt.ylabel("number of reservation")
plt.legend(["not_cancaled", "cancaled"])

plt.show()
```



In [87]:

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

Out[87]:

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

In [88]:

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

Out[88]:

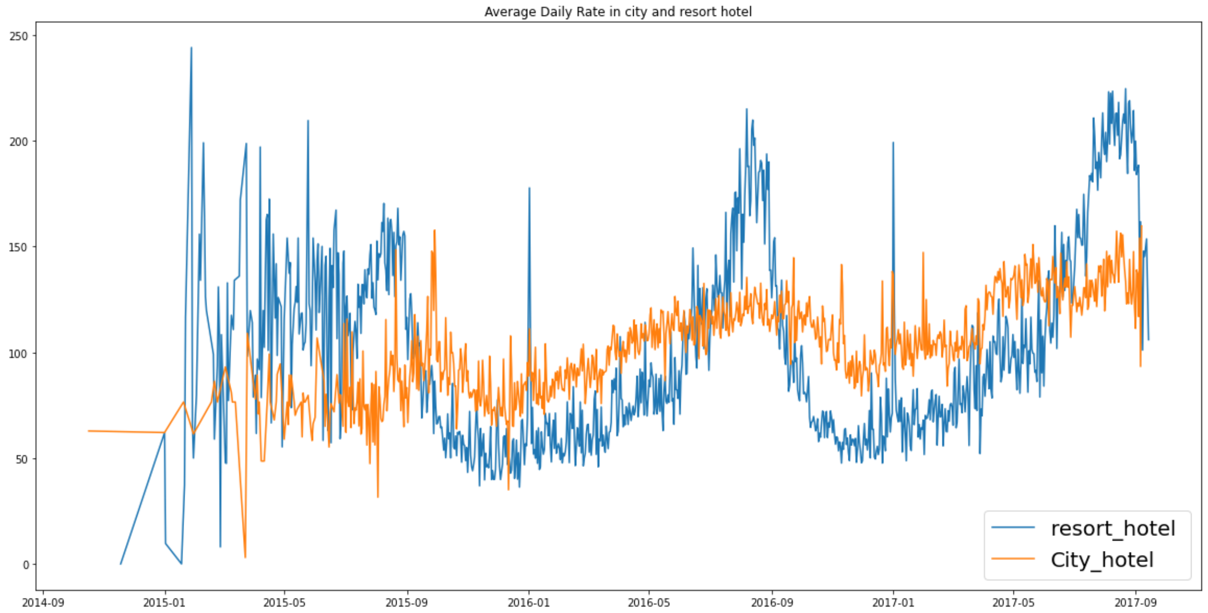
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [89]:

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

In [90]:

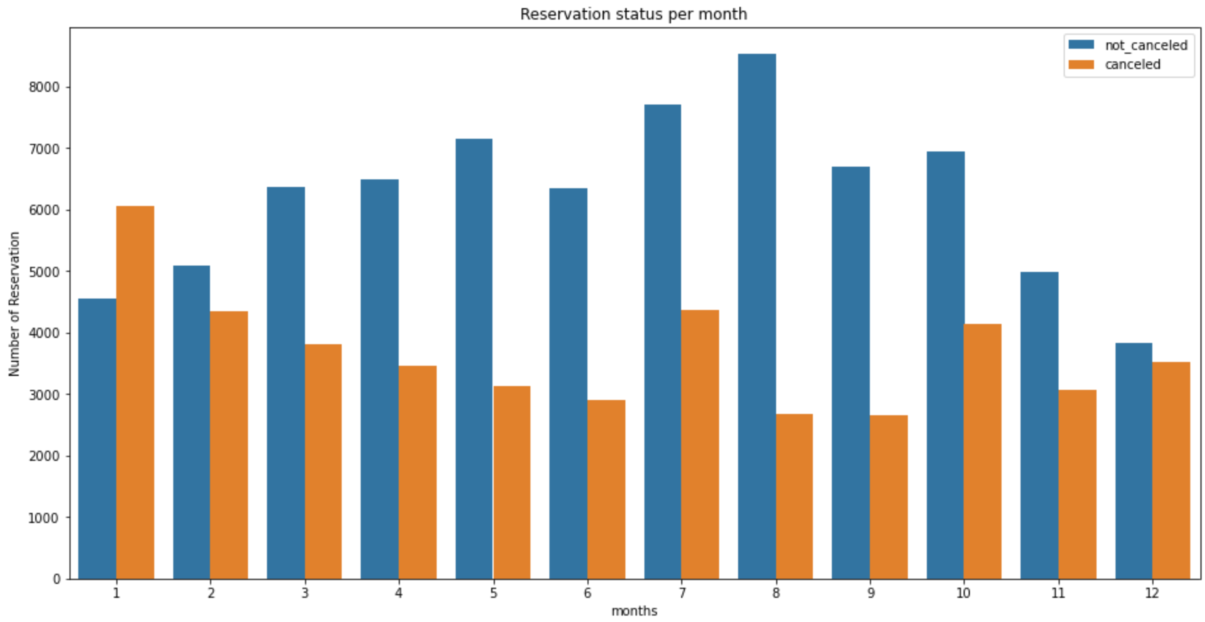
```
plt.figure(figsize=(20,10))
plt.title("Average Daily Rate in city and resort hotel")
plt.plot(resort_hotel.index,resort_hotel['adr'],label='resort_hotel ' )
plt.plot(City_hotel.index,City_hotel['adr'],label='City_hotel'   )
plt.legend(fontsize=20)
plt.show()
```



In [91]:

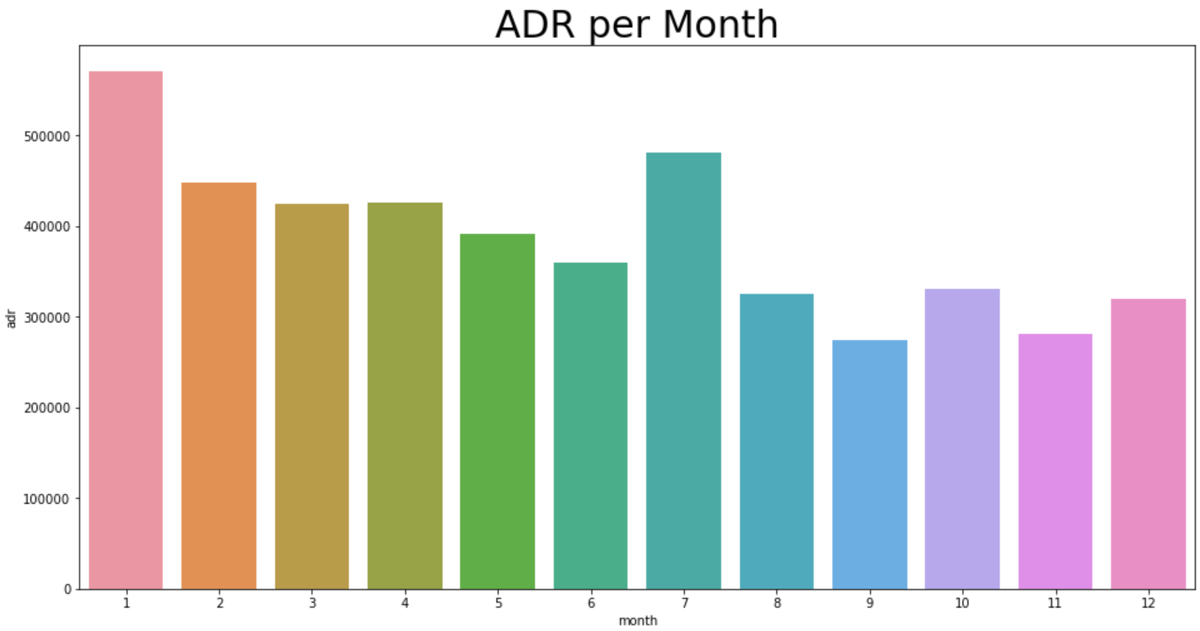
```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
sns.countplot(x='month' ,hue='is_canceled',data=df)

plt.title("Reservation status per month")
plt.ylabel("Number of Reservation")
plt.xlabel("months")
plt.legend(["not_canceled","canceled"])
plt.show()
```



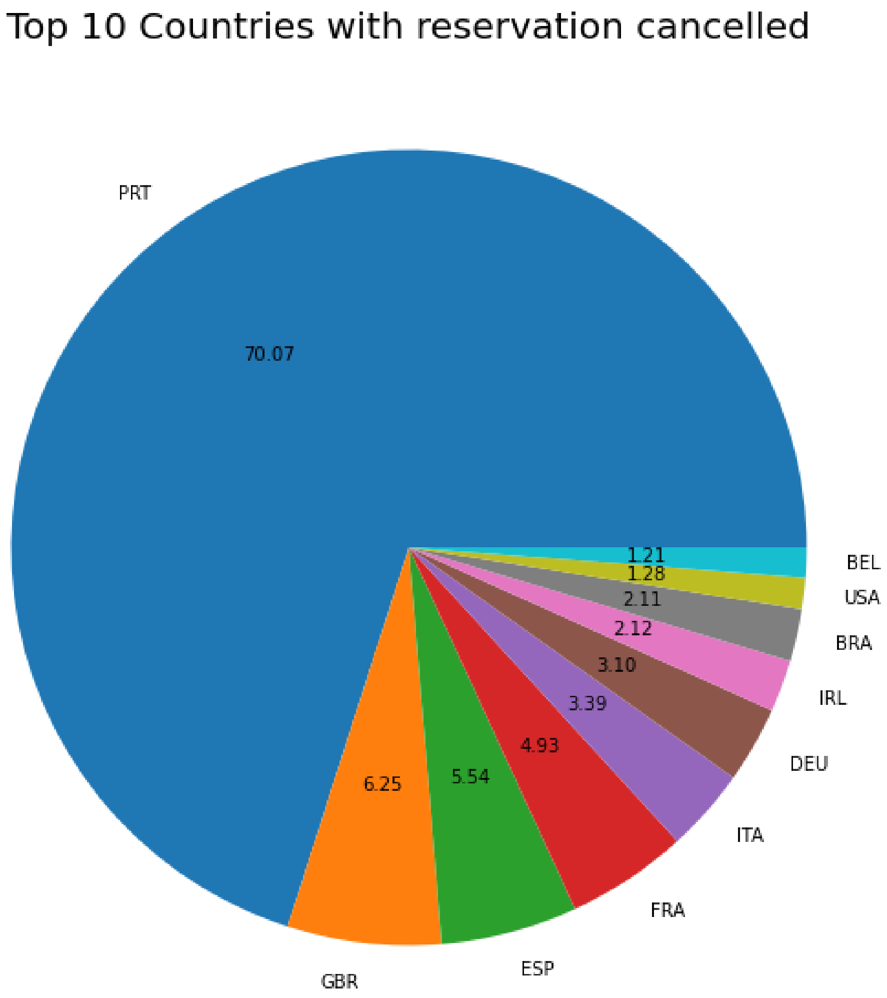
In [92]:

```
plt.figure(figsize=(16,8))
plt.title("ADR per Month", fontsize=30)
sns.barplot("month","adr",data=df[df['is_canceled']==1].groupby('month')[['adr']].sum().reset_i
plt.show()
```



In [93]:

```
plt.figure(figsize=(10,10))
cancel=df[df['is_canceled']==1]
top_10_country=cancel['country'].value_counts()[:10]
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.title("Top 10 Countries with reservation cancelled",fontsize=20)
plt.show()
```



In [94]:

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

Out[94]:

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

In [101]:

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

Out[101]:

```
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 [96]:

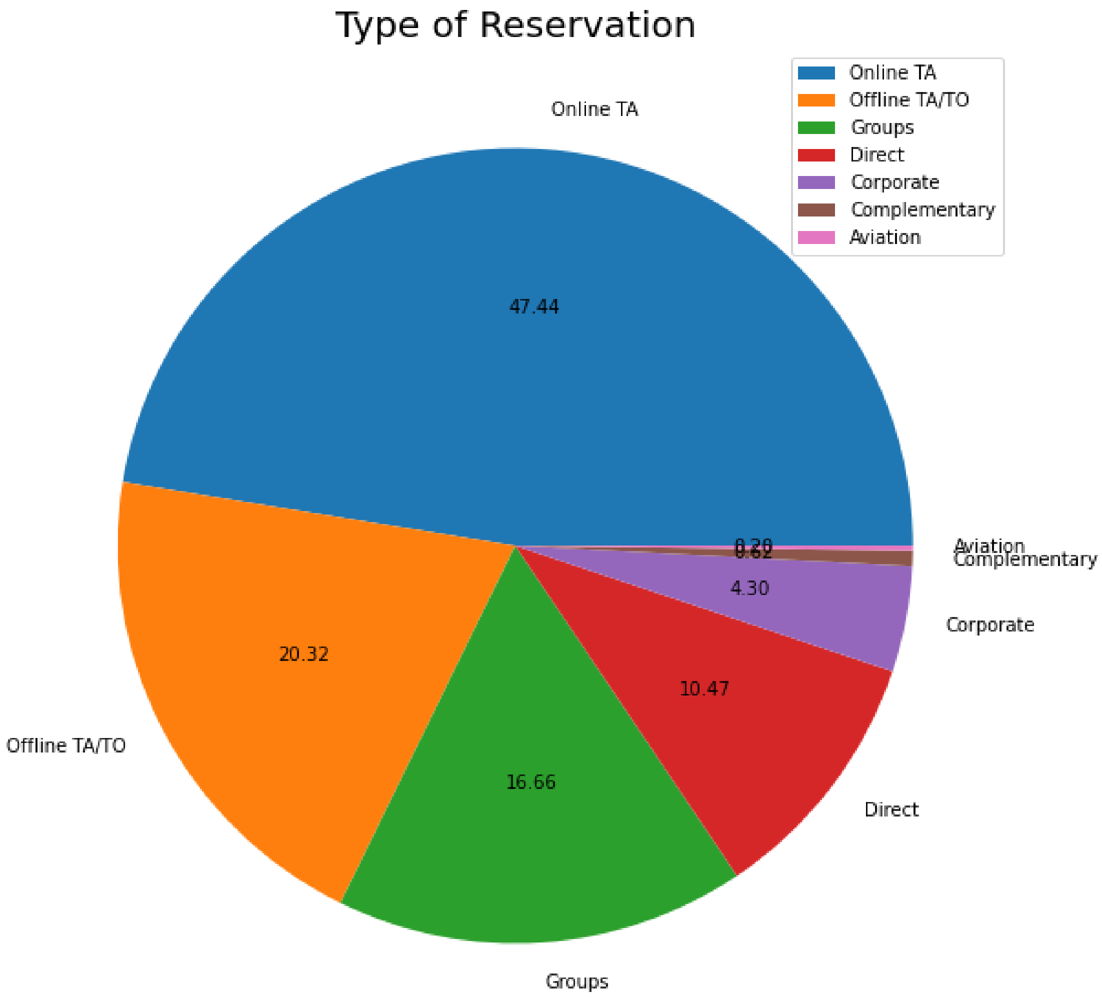
```
cancel['market_segment'].value_counts(normalize=True)
```

Out[96]:

```
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 [120]:

```
plt.figure(figsize=(10,10))
plt.pie(type_reservation,autopct='%.2f',labels=type_reservation.index,radius=1)
plt.title("Type of Reservation",fontsize=20)
plt.legend(loc='upper right')
plt.show()
```


Type of Reservation
Online TA
Offline TA/TO
Groups
Direct
Corporate
Complementary
Aviation
Online TA
Offline TA/TO
Groups
Direct
Corporate
Complementary
Aviation