data-analysis-hotel-booking

November 6, 2023

1 Importing Libraries

[53]:
```
import pandas as pd
import matplotlib as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
```

2 Loading the dataset

[54]:
```
df = pd.read_csv('hotel_booking.csv')
```

3 Exploratory Data Analysis and Data Cleaning

[55]:
```
df.head()
```

[55]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	\
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	

	arrival_date_week_number	arrival_date_day_of_month	\
0	27	1	
1	27	1	
2	27	1	
3	27	1	
4	27	1	

	stays_in_weekend_nights	stays_in_week_nights	adults	…	customer_type	\
0	0	0	2	…	Transient	
1	0	0	2	…	Transient	
2	0	1	1	…	Transient	

```
3                       0                   1       1  …      Transient
4                       0                   2       2  …      Transient

    adr required_car_parking_spaces total_of_special_requests  \
0   0.0                           0                          0
1   0.0                           0                          0
2  75.0                           0                          0
3  75.0                           0                          0
4  98.0                           0                          1

  reservation_status reservation_status_date              name  \
0          Check-Out              2015-07-01    Ernest Barnes
1          Check-Out              2015-07-01     Andrea Baker
2          Check-Out              2015-07-02   Rebecca Parker
3          Check-Out              2015-07-02    Laura Murray
4          Check-Out              2015-07-03     Linda Hines

                         email  phone-number       credit_card
0  Ernest.Barnes31@outlook.com  669-792-1661  ************4322
1       Andrea_Baker94@aol.com  858-637-6955  ************9157
2   Rebecca_Parker@comcast.net  652-885-2745  ************3734
3            Laura_M@gmail.com  364-656-8427  ************5677
4           LHines@verizon.com  713-226-5883  ************5498

[5 rows x 36 columns]
```

[56]:
```
df.tail()
```

[56]:
```
             hotel  is_canceled  lead_time  arrival_date_year  \
119385  City Hotel            0         23               2017
119386  City Hotel            0        102               2017
119387  City Hotel            0         34               2017
119388  City Hotel            0        109               2017
119389  City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
119385                         30                        2
119386                         31                        2
119387                         31                        2
119388                         31                        2
```

```
119389                          29                          2

        stays_in_week_nights  adults  …  customer_type     adr  \
119385                     5       2  …      Transient   96.14
119386                     5       3  …      Transient  225.43
119387                     5       2  …      Transient  157.71
119388                     5       2  …      Transient  104.40
119389                     7       2  …      Transient  151.20

       required_car_parking_spaces total_of_special_requests  \
119385                           0                          0
119386                           0                          2
119387                           0                          4
119388                           0                          0
119389                           0                          2

       reservation_status reservation_status_date               name  \
119385          Check-Out              2017-09-06     Claudia Johnson
119386          Check-Out              2017-09-07      Wesley Aguilar
119387          Check-Out              2017-09-07        Mary Morales
119388          Check-Out              2017-09-07  Caroline Conley MD
119389          Check-Out              2017-09-07      Ariana Michael

                         email  phone-number       credit_card
119385       Claudia.J@yahoo.com  403-092-5582  ************8647
119386      WAguilar@xfinity.com  238-763-0612  ************4333
119387  Mary_Morales@hotmail.com  395-518-4100  ************1821
119388   MD_Caroline@comcast.net  531-528-1017  ************7860
119389      Ariana_M@xfinity.com  422-804-6403  ************4482

[5 rows x 36 columns]
```

[57]:
```
df.shape
```

[57]:
```
(119390, 36)
```

[58]:
```
df.columns
```

[58]:
```
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', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

[59]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
```

```
 35  credit_card                     119390 non-null  object
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

[60]:
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

[61]:
```
df.describe(include = 'object')
```

[61]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status	name	email	phone-number	credit_card
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3	81503	115889	119390	9000
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out	Michael Johnson	Michael.C@gmail.com	669-792-1661	************4923
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166	48	6	1	28

[62]:
```
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']
--------------------------------------------------
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']
--------------------------------------------------
```

[63]:
```
df.isnull().sum()
```

[63]:
```
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
```

[64]:
```
df.drop(['company','agent'], axis = 1, inplace=True)
df.dropna(inplace = True)
```

[65]:
```
df.isnull().sum()
```

[65]:
```
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
```

[66]:
```
df.describe()
```

[66]:
```
          is_canceled      lead_time  arrival_date_year  \
count  118898.000000  118898.000000      118898.000000
mean        0.371352     104.311435        2016.157656
std         0.483168     106.903309           0.707459
```

```
min         0.000000       0.000000        2015.000000
25%         0.000000      18.000000        2016.000000
50%         0.000000      69.000000        2016.000000
75%         1.000000     161.000000        2017.000000
max         1.000000     737.000000        2017.000000

       arrival_date_week_number  arrival_date_day_of_month  \
count             118898.000000              118898.000000
mean                  27.166555                  15.800880
std                   13.589971                   8.780324
min                    1.000000                   1.000000
25%                   16.000000                   8.000000
50%                   28.000000                  16.000000
75%                   38.000000                  23.000000
max                   53.000000                  31.000000

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            118898.000000         118898.000000  118898.000000
mean                  0.928897              2.502145       1.858391
std                   0.996216              1.900168       0.578576
min                   0.000000              0.000000       0.000000
25%                   0.000000              1.000000       2.000000
50%                   1.000000              2.000000       2.000000
75%                   2.000000              3.000000       2.000000
max                  16.000000             41.000000      55.000000

            children         babies  is_repeated_guest  \
count  118898.000000  118898.000000      118898.000000
mean        0.104207       0.007948           0.032011
std         0.399172       0.097380           0.176029
min         0.000000       0.000000           0.000000
25%         0.000000       0.000000           0.000000
50%         0.000000       0.000000           0.000000
75%         0.000000       0.000000           0.000000
max        10.000000      10.000000           1.000000

       previous_cancellations  previous_bookings_not_canceled  \
count           118898.000000                   118898.000000
mean                 0.087142                        0.131634
std                  0.845869                        1.484672
min                  0.000000                        0.000000
25%                  0.000000                        0.000000
50%                  0.000000                        0.000000
75%                  0.000000                        0.000000
max                 26.000000                       72.000000

       booking_changes  days_in_waiting_list            adr  \
```

```
count    118898.000000         118898.000000  118898.000000
mean          0.221181              2.330754     102.003243
std           0.652785             17.630452      50.485862
min           0.000000              0.000000      -6.380000
25%           0.000000              0.000000      70.000000
50%           0.000000              0.000000      95.000000
75%           0.000000              0.000000     126.000000
max          21.000000            391.000000    5400.000000

       required_car_parking_spaces  total_of_special_requests
count                118898.000000              118898.000000
mean                      0.061885                   0.571683
std                       0.244172                   0.792678
min                       0.000000                   0.000000
25%                       0.000000                   0.000000
50%                       0.000000                   0.000000
75%                       0.000000                   1.000000
max                       8.000000                   5.000000
```

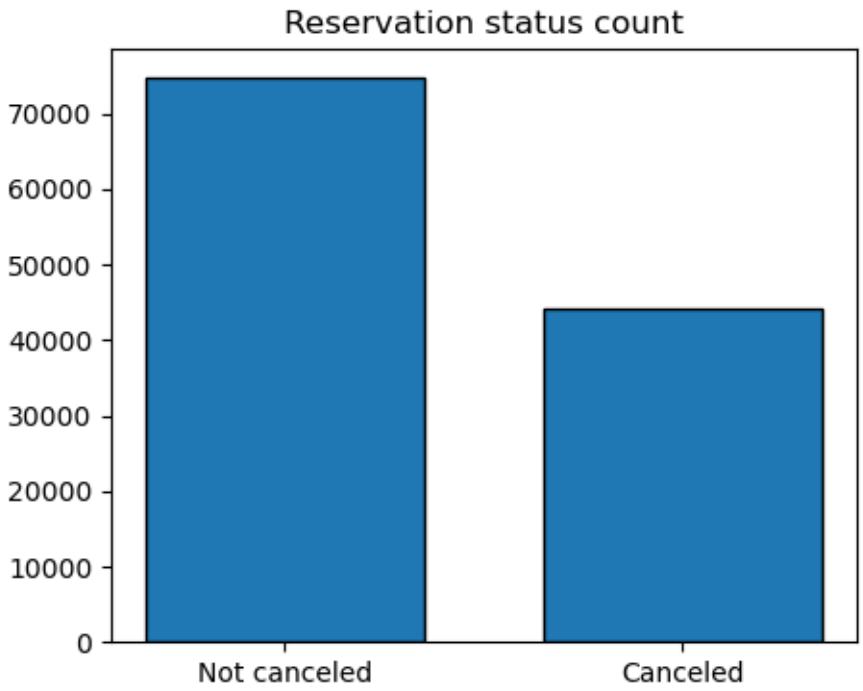
[67]:
```
df = df[df['adr']<5000]
```

4 Data Analysis and Visualizations

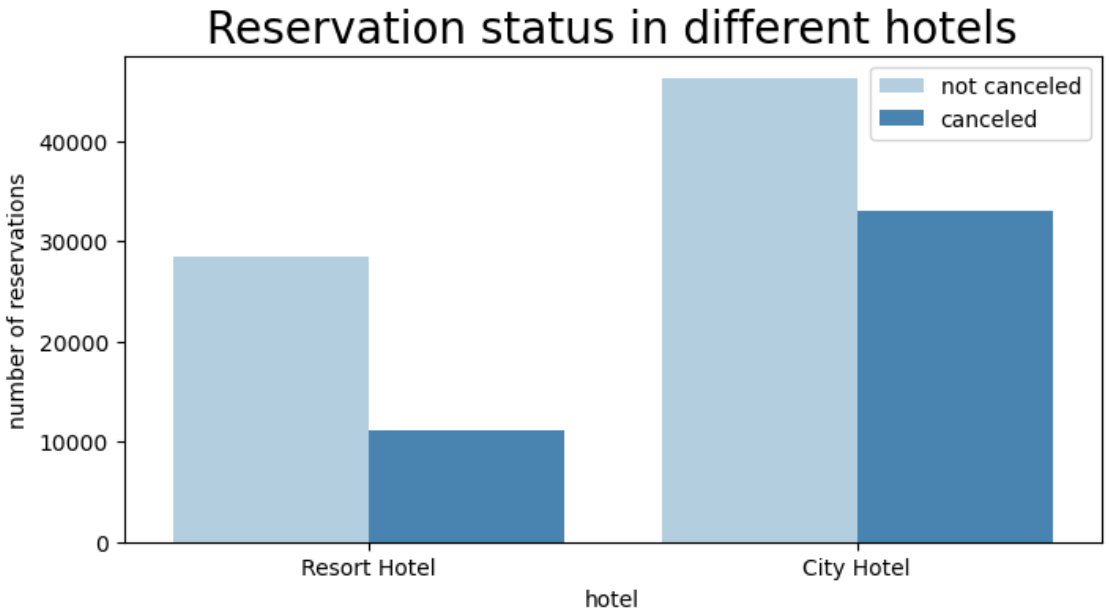
[68]:
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)

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

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



[69]:
```
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()
```



[70]:
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

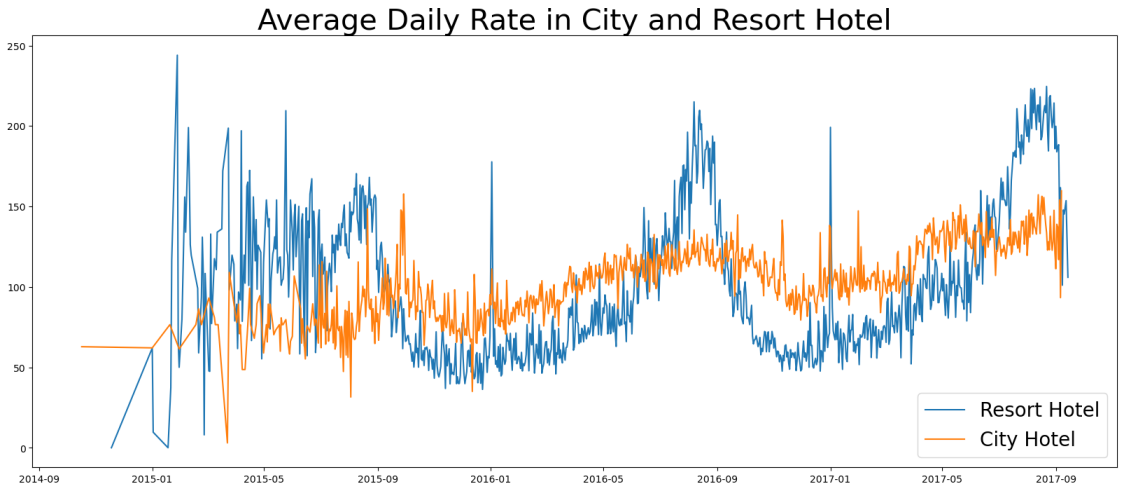
[70]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

[71]:
```
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

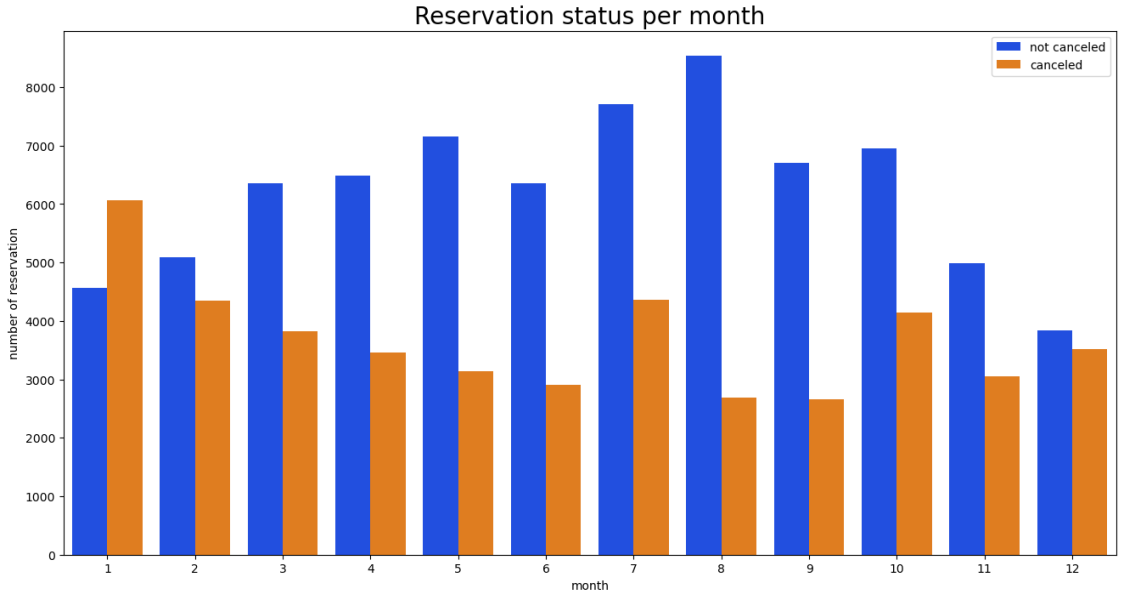
[71]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

[72]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

[73]:
```
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()
```

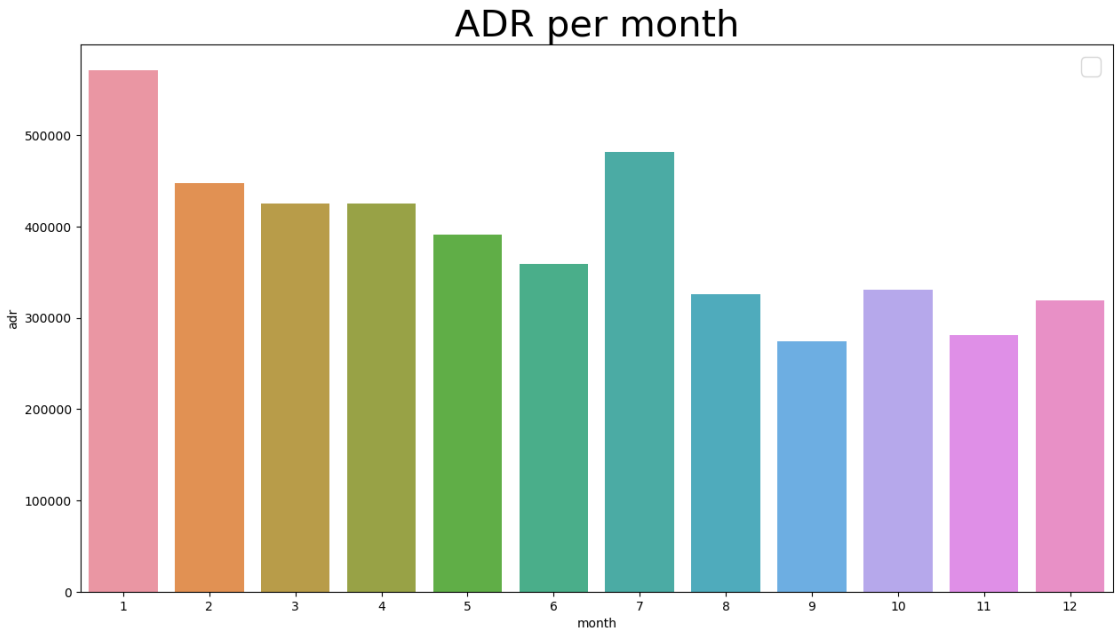


[74]:
```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x = 'month', hue = 'is_canceled', data = df, palette =␣
 ↪'bright')
legend_labels,_ = ax1. get_legend_handles_labels()
ax1.legend(bbox_to_anchor= (1,1))
plt.title('Reservation status per month', size = 20)
plt.xlabel('month')
plt.ylabel('number of reservation')
plt.legend(['not canceled', 'canceled'])
plt.show()
```

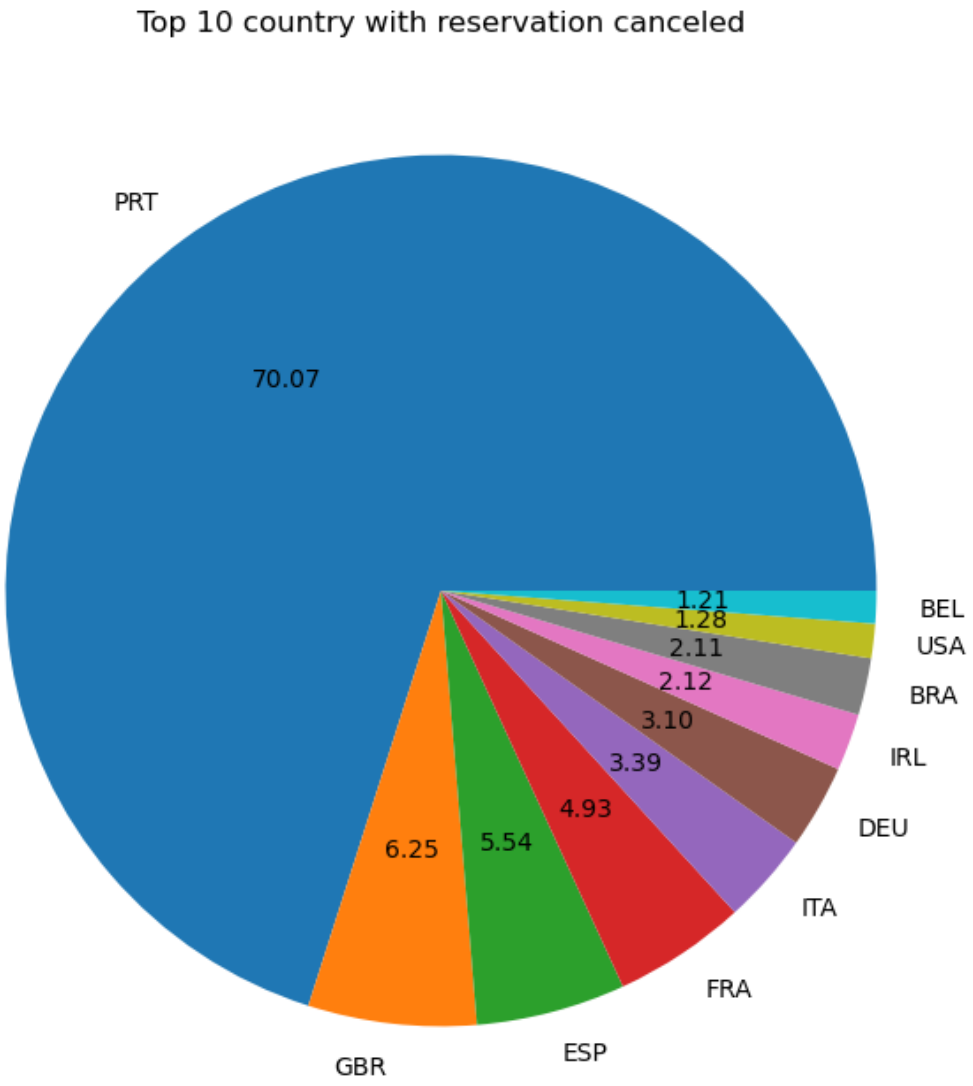


[75]:
```
plt.figure(figsize = (15,8))
plt.title('ADR per month', fontsize = 30)
sns.barplot('month', 'adr', data = df[df['is_canceled'] == 1].
 ↪groupby('month')[['adr']].sum().reset_index())
plt.legend(fontsize = 20)
plt.show()
```

```
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.
```



[77]:
```
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 country with reservation canceled')
plt.pie(top_10_country, autopct = '%.2f', labels = top_10_country.index)
plt.show()
```



[78]:
```
df['market_segment'].value_counts()
```

[78]:
```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```