data-analysis-hotel-booking

May 12, 2024

[2]:

```
# This Python 3 environment comes with many helpful analytics libraries␣
 ↪installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/
 ↪docker-python
# For example, here's several helpful packages to load

#import numpy as np # linear algebra
#import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list␣
 ↪all files under the input directory

#import os
#for dirname, _, filenames in os.walk('/kaggle/input'):
 #   for filename in filenames:
  #      print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that␣
 ↪gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved␣
 ↪outside of the current session
```

1 Importing Libraries

[3]:

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

```
/opt/conda/lib/python3.10/site-packages/scipy/__init__.py:146: UserWarning: A
NumPy version >=1.16.5 and <1.23.0 is required for this version of SciPy
(detected version 1.23.5
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

2 Loading the dataset

[4]:
```
df = pd.read_csv('/kaggle/input/hotel-booking-new/new_hotel_booking.csv')
```

3 Exploratory Data Analysis and Data Cleaning

[5]:
```
df.head()
```

[5]:
```
          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  ...  deposit_type  \
0                        0                     0       2  ...    No Deposit
1                        0                     0       2  ...    No Deposit
2                        0                     1       1  ...    No Deposit
3                        0                     1       1  ...    No Deposit
4                        0                     2       2  ...    No Deposit

   agent company days_in_waiting_list customer_type   adr  \
0    NaN     NaN                    0     Transient   0.0
1    NaN     NaN                    0     Transient   0.0
2    NaN     NaN                    0     Transient  75.0
3  304.0     NaN                    0     Transient  75.0
4  240.0     NaN                    0     Transient  98.0

   required_car_parking_spaces  total_of_special_requests reservation_status  \
0                            0                          0          Check-Out
1                            0                          0          Check-Out
2                            0                          0          Check-Out
3                            0                          0          Check-Out
4                            0                          1          Check-Out

  reservation_status_date
0               01-07-2015
1               01-07-2015
```

```
2              02-07-2015
3              02-07-2015
4              03-07-2015

[5 rows x 32 columns]
```

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

```
[6]:              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  …  deposit_type  agent company  \
119385                     5       2  …    No Deposit  394.0     NaN
119386                     5       3  …    No Deposit    9.0     NaN
119387                     5       2  …    No Deposit    9.0     NaN
119388                     5       2  …    No Deposit   89.0     NaN
119389                     7       2  …    No Deposit    9.0     NaN

       days_in_waiting_list customer_type     adr  \
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     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
119385           Check-Out              06-09-2017
119386           Check-Out              07-09-2017
119387           Check-Out              07-09-2017
119388           Check-Out              07-09-2017
119389           Check-Out              07-09-2017

[5 rows x 32 columns]
```

[7]:

```
df.shape
```

[7]:

```
(119390, 32)
```

[8]:

```
df.columns
```

[8]:

```
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')
```

[9]:

```
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
```

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

```
[11]: 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  datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

[12]:
```
df.describe(include= 'object')
```

[12]:
```
             hotel arrival_date_month    meal country market_segment  \
count       119390             119390  119390  118902         119390
unique           2                 12       5     177              8
top     City Hotel             August      BB     PRT      Online TA
freq         79330              13877   92310   48590          56477

       distribution_channel reserved_room_type assigned_room_type  \
count                119390             119390             119390
unique                    5                 10                 12
top                   TA/TO                  A                  A
freq                  97870              85994              74053

       deposit_type customer_type reservation_status
count        119390        119390             119390
unique            3             4                  3
top      No Deposit     Transient          Check-Out
freq         104641         89613              75166
```

[13]:
```
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']
--------------------------------------------------
```

[14]:

```
#displaying the number of missing values in each column

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

[14]:

```
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
```

[15]:

```
#dropping the columns with many null values as it is diffult to handle them.


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

[16]:

```
# removing the rows with NULL values

df.dropna(inplace=True)
```

[17]:
```
df.isnull().sum()
```

[17]:
```
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
```

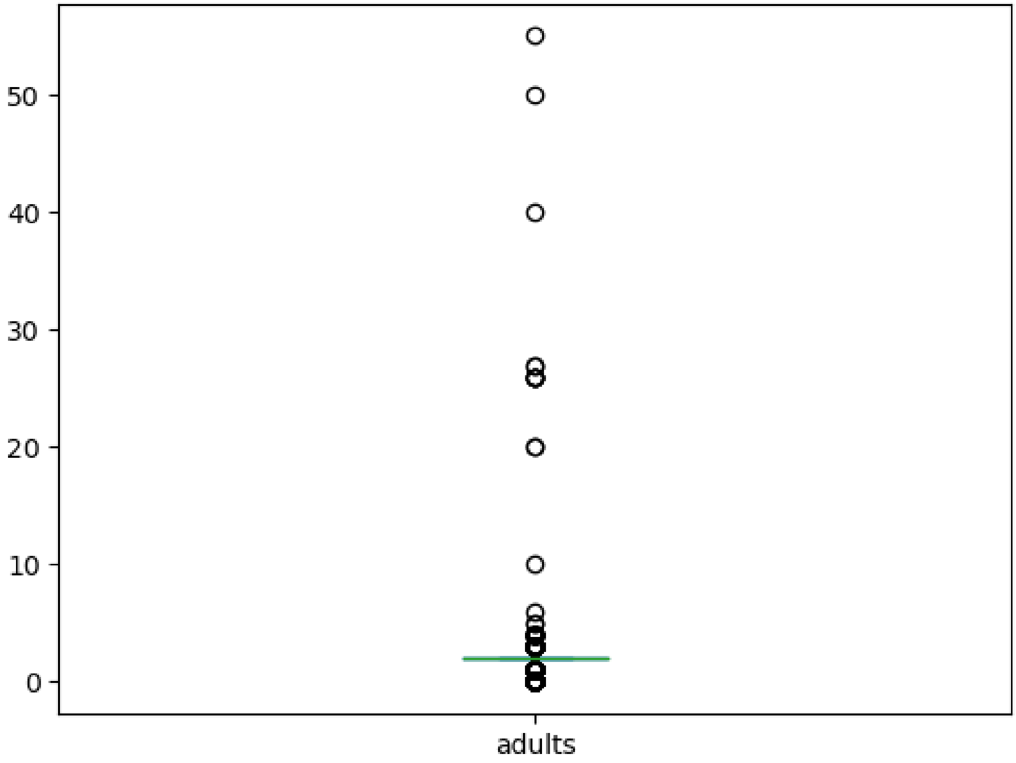
[18]:
```
df.describe()
```

[18]:
```
         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
```

[19]:
```
df['adults'].plot(kind='box')
```

[19]:
```
<Axes: >
```



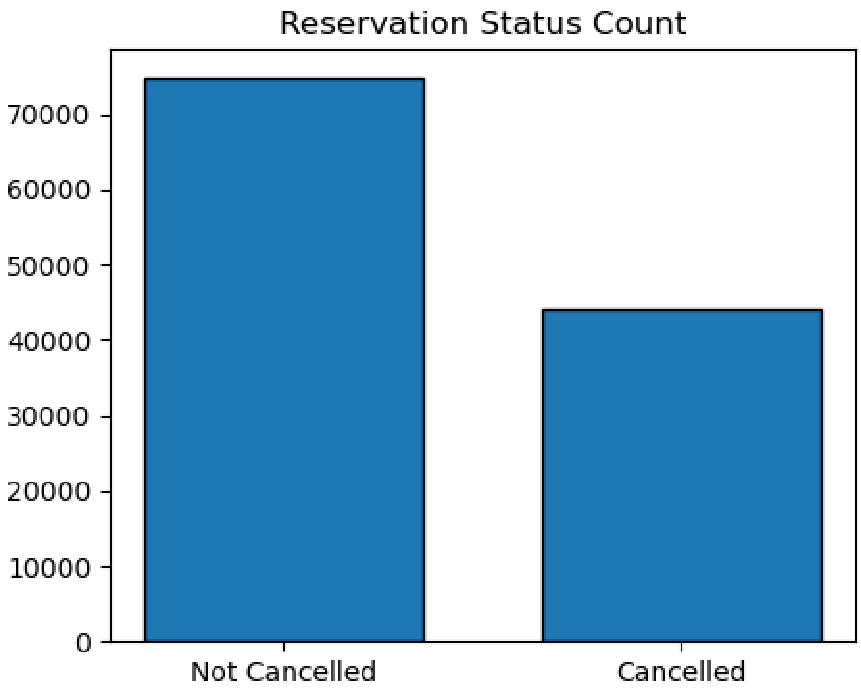
[20]:
```
df = df[df['adr']<5000]
```

4 Data Analysis and Visualizations

[21]:

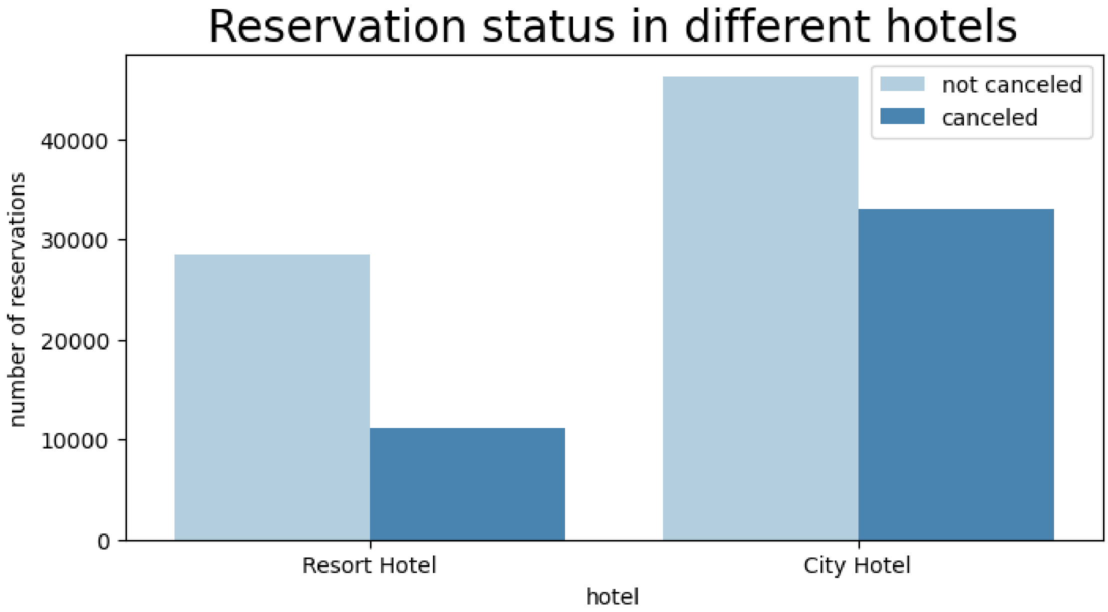
```
cancelled_percentage = df['is_canceled'].value_counts(normalize= True)
cancelled_percentage = cancelled_percentage*100
print(cancelled_percentage)

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

```
0    62.865337
1    37.134663
Name: is_canceled, dtype: float64
```



[22]:

```
plt.figure(figsize = (8,4))
ax1=sb.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()
```



[23]:

```
# Canceled and not canceled percentage of each hotel.

resort_hotel= df[df['hotel']== 'Resort Hotel']
cancelled_percentage = resort_hotel['is_canceled'].value_counts(normalize= True)
cancelled_percentage = cancelled_percentage*100
print(cancelled_percentage)
```

```
0    72.024952
1    27.975048
Name: is_canceled, dtype: float64
```

[24]:

```
city_hotel= df[df['hotel']== 'City Hotel']
cancelled_per = city_hotel['is_canceled'].value_counts(normalize= True)
cancelled_per = cancelled_per*100
print(cancelled_per)
```

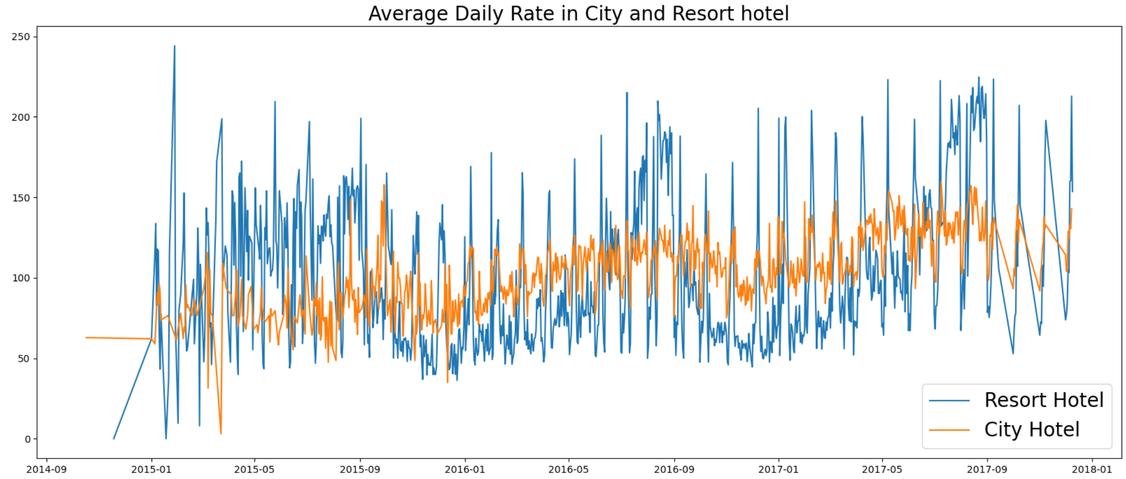
```
0    58.291825
1    41.708175
Name: is_canceled, dtype: float64
```

[25]:

```
# Grouping of booking status date using groupby

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

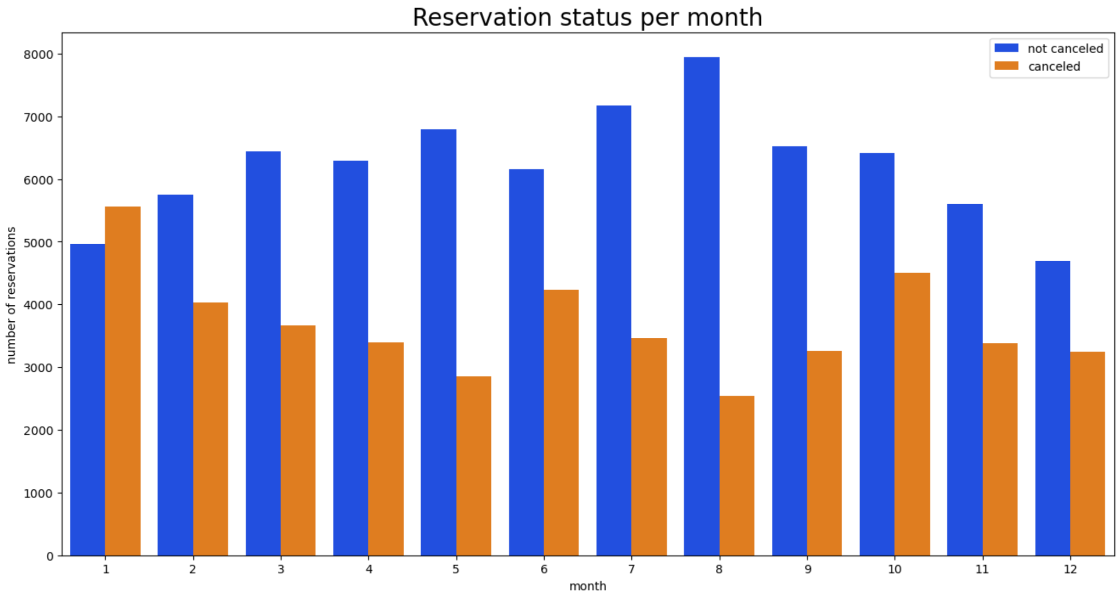
[26]:

```
# visualization of average daily rate

plt.figure(figsize= (20,8))
plt.title('Average Daily Rate in City and Resort hotel', fontsize= 20)
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()
```



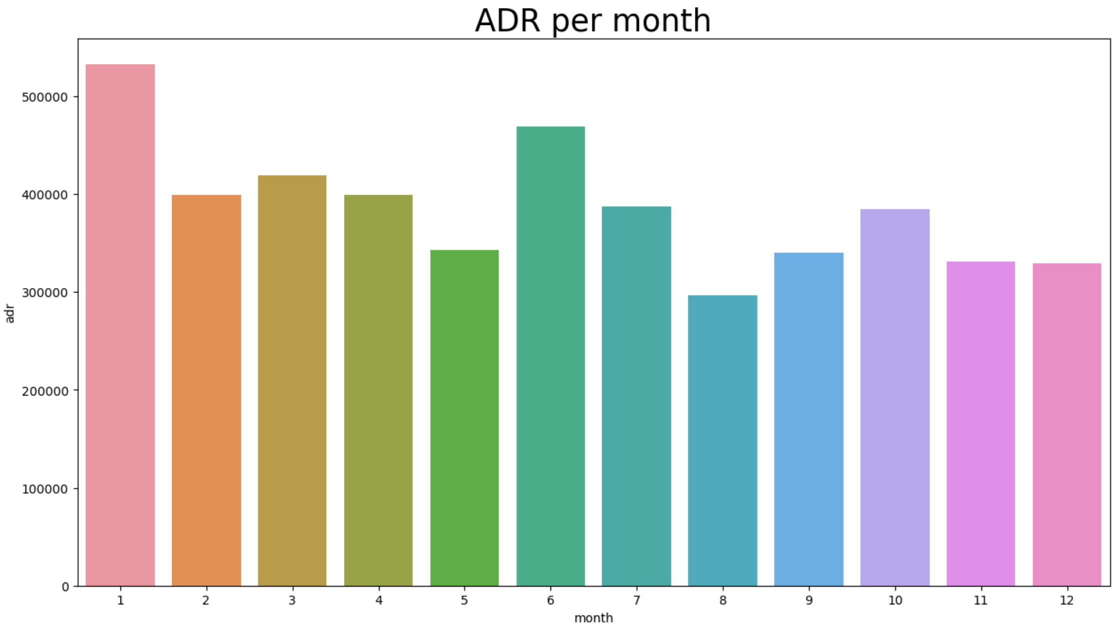
[27]:

```
# maximum and minimum nunber of cancellations in months

df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize =(16,8))
ax1= sb.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 reservations')
plt.legend(['not canceled','canceled'])
plt.show()
```

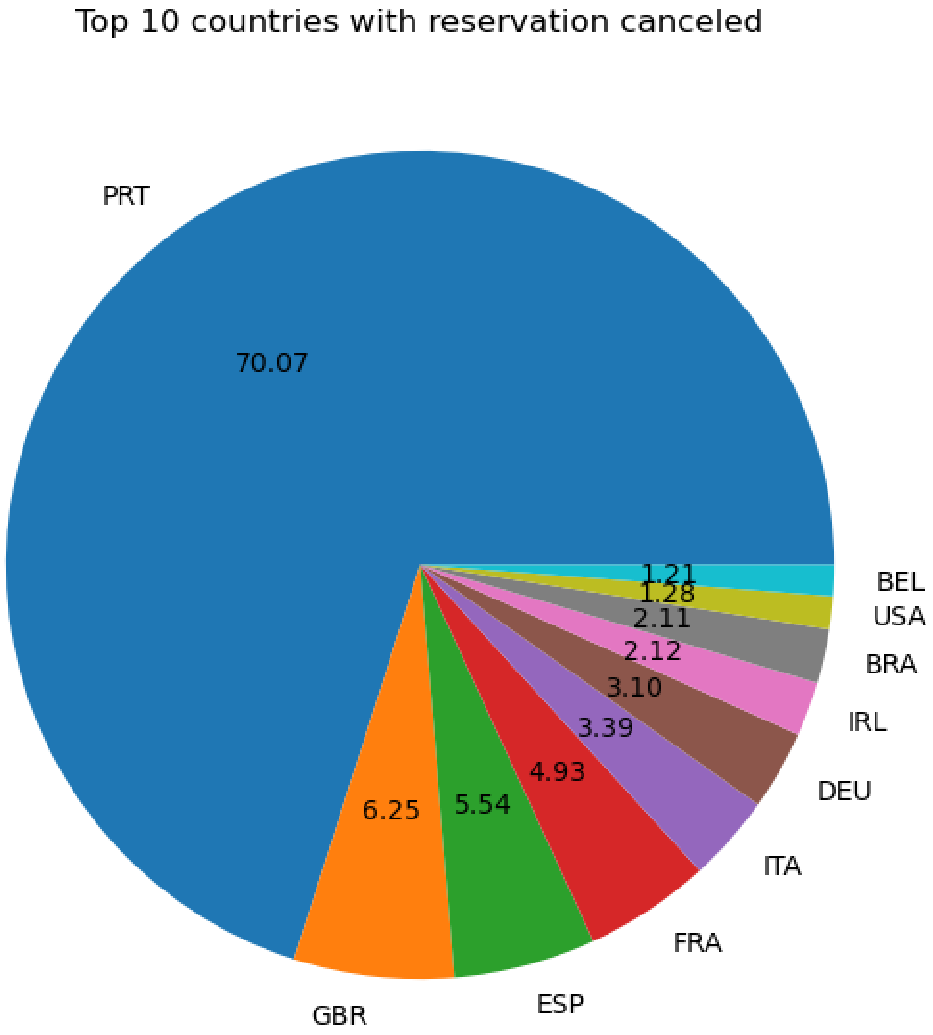


[28]:
```
# average daily rate for each month

plt.figure(figsize= (15,8))
plt.title('ADR per month', fontsize = 25)
sb.barplot(x = 'month', y = 'adr',data = df[df['is_canceled'] == 1].
  ↪groupby('month')[['adr']].sum().reset_index())
plt.show
```

[28]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



[29]:
```
cancelled_data = df[df['is_canceled'] == 1]
top_10_country = cancelled_data['country'].value_counts()[:10]
plt.figure(figsize = (7,7))
plt.title('Top 10 countries with reservation canceled')
plt.pie(top_10_country, autopct = '%.2f' , labels = top_10_country.index)
plt.show()
```



[30]:
```
df['market_segment'].value_counts()
```

[30]:
```
Online TA        56402
Offline TA/TO    24159
```

```
Groups            19806
Direct            12448
Corporate          5111
Complementary       734
Aviation            237
Name: market_segment, dtype: int64
```

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

```
[31]: 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
```

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

```
[32]: 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
```

```
[ ]:
```