hotel-booking

April 23, 2024

[2]:

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

[4]:

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

[5]:

```python
df
```

[5]:

```
               hotel  is_canceled  lead_time  arrival_date_year  \
0       Resort Hotel            0        342               2015
1       Resort Hotel            0        737               2015
2       Resort Hotel            0          7               2015
3       Resort Hotel            0         13               2015
4       Resort Hotel            0         14               2015
…              …          …        …                …
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  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
…                   …                       …
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                        0
1                               1                        0
2                               1                        0
3                               1                        0
4                               1                        0
…                           …                      …
119385                         30                        2
119386                         31                        2
119387                         31                        2
119388                         31                        2
119389                         29                        2

        stays_in_week_nights  adults  …  customer_type    adr  \
0                          0       2  …      Transient    0.00
1                          0       2  …      Transient    0.00
2                          1       1  …      Transient   75.00
3                          1       1  …      Transient   75.00
4                          2       2  …      Transient   98.00
…                      …     …  …            …     …
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  \
0                                0                         0
1                                0                         0
2                                0                         0
3                                0                         0
4                                0                         1
…                            …                       …
119385                           0                         0
119386                           0                         2
119387                           0                         4
119388                           0                         0
119389                           0                         2

       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
…                   …                     …                 …
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
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
…                               …           …                …
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

[119390 rows x 36 columns]
```

[29]:

```
df.head()
```

[29]:

```
          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]
```

[30]:

```
df.tail()
```

[30]:

```
             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]
```

0.1 Checking for the Null values in columns

[31]:

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

[31]:

```
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
```

0.2 Converting int to date time for reservation date

[29]:

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

[30]:

```
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  datetime64[ns]
 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: datetime64[ns](1), float64(4), int64(16), object(15)
memory usage: 32.8+ MB
```

[34]:

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

[34]:

```
                 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             name  \
count        119390        119390             119390           119390
unique            3             4                  3            81503
top      No Deposit     Transient          Check-Out  Michael Johnson
freq         104641         89613              75166               48

                      email  phone-number       credit_card
count                119390        119390            119390
unique               115889        119390              9000
top     Michael.C@gmail.com  669-792-1661  ************4923
freq                      6             1                28
```

0.2.1 checking the unique value for the above column

[35]:

```python
for col in df.describe(include='object').columns:
    print (col)
    print(df[col].unique())
    print('-'*100)
```

```
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']
--------------------------------------------------------------------------------
--------------------
```

```
[36]: df.columns
```

```
[36]: 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')
```

0.3 Removing 'agent','company' columns beacuse it has more than 1 lakh missing value and Removing all the null value from 'babies', 'country' column

[37]:

```
df.drop(['agent','company'],axis=1,inplace= True)
#axis =1 for column removal and inplace for permanent removal
```

[38]:

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

[39]:

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

[39]:

```
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
```

[40]:

```
df.describe()
```

[40]:

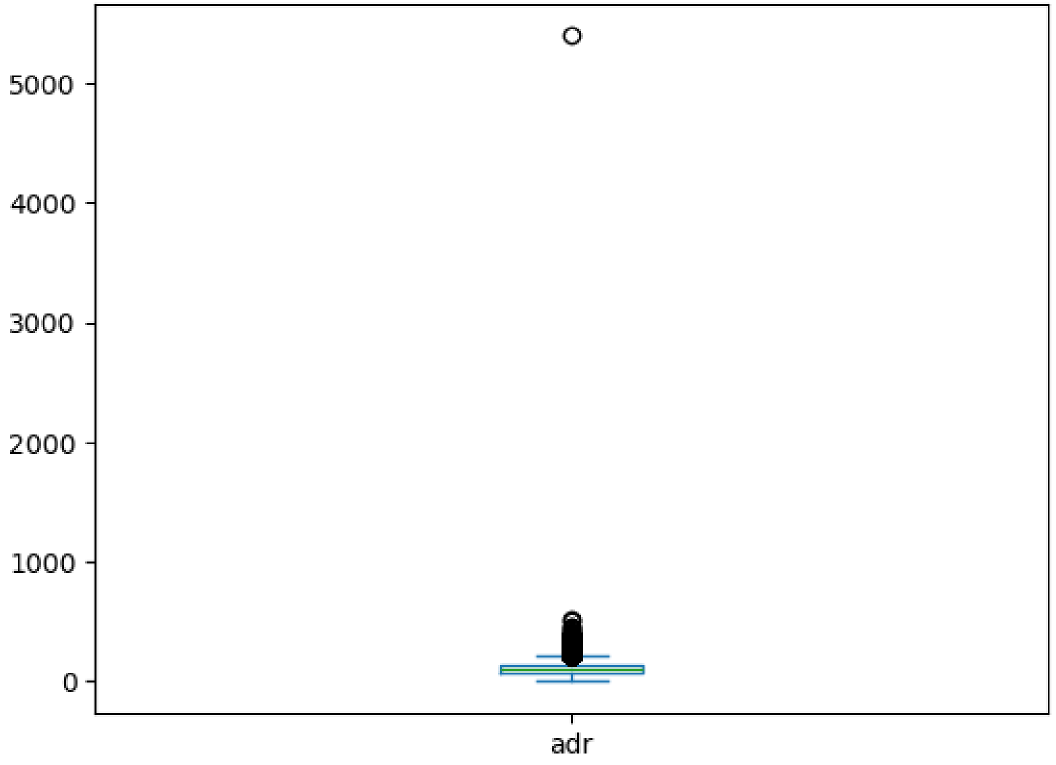
```
         is_canceled      lead_time  arrival_date_year  \
count  118898.000000  118898.000000      118898.000000
mean        0.371352     104.311435        2016.157656
min         0.000000       0.000000        2015.000000
25%         0.000000      18.000000        2016.000000
50%         0.000000      69.000000        2016.000000
75%         1.000000     161.000000        2017.000000
max         1.000000     737.000000        2017.000000
std         0.483168     106.903309           0.707459

       arrival_date_week_number  arrival_date_day_of_month  \
count             118898.000000              118898.000000
mean                  27.166555                  15.800880
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
std                   13.589971                   8.780324

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            118898.000000         118898.000000  118898.000000
mean                  0.928897              2.502145       1.858391
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
std                   0.996216              1.900168       0.578576

            children         babies  is_repeated_guest  \
count  118898.000000  118898.000000      118898.000000
mean        0.104207       0.007948           0.032011
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
std         0.399172       0.097380           0.176029

       previous_cancellations  previous_bookings_not_canceled  \
count           118898.000000                   118898.000000
mean                 0.087142                        0.131634
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
std                  0.845869                        1.484672

       booking_changes  days_in_waiting_list            adr  \
count    118898.000000         118898.000000  118898.000000
mean          0.221181              2.330754     102.003243
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
std           0.652785             17.630452      50.485862

       required_car_parking_spaces  total_of_special_requests  \
count                118898.000000              118898.000000
mean                      0.061885                   0.571683
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
std                       0.244172                   0.792678

             reservation_status_date
count                         118898
mean   2016-07-30 07:37:53.336809984
min              2014-10-17 00:00:00
25%              2016-02-02 00:00:00
50%              2016-08-08 00:00:00
75%              2017-02-09 00:00:00
max              2017-09-14 00:00:00
std                              NaN
```

0.4 checking the outlier in adr column

[41]:
```
df['adr'].plot(kind = 'box')
```

[41]:
```
<Axes: >
```



[42]:
```
df = df[df['adr']<5000]
```

[43]:
```
df['adr'].describe()
```

[43]:
```
count    118897.000000
mean        101.958683
std          48.091199
min          -6.380000
25%          70.000000
50%          95.000000
75%         126.000000
max         510.000000
Name: adr, dtype: float64
```

1 EDA

[44]:
```
df.columns
```

[44]:
```
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', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

1.1 Cancellation Count

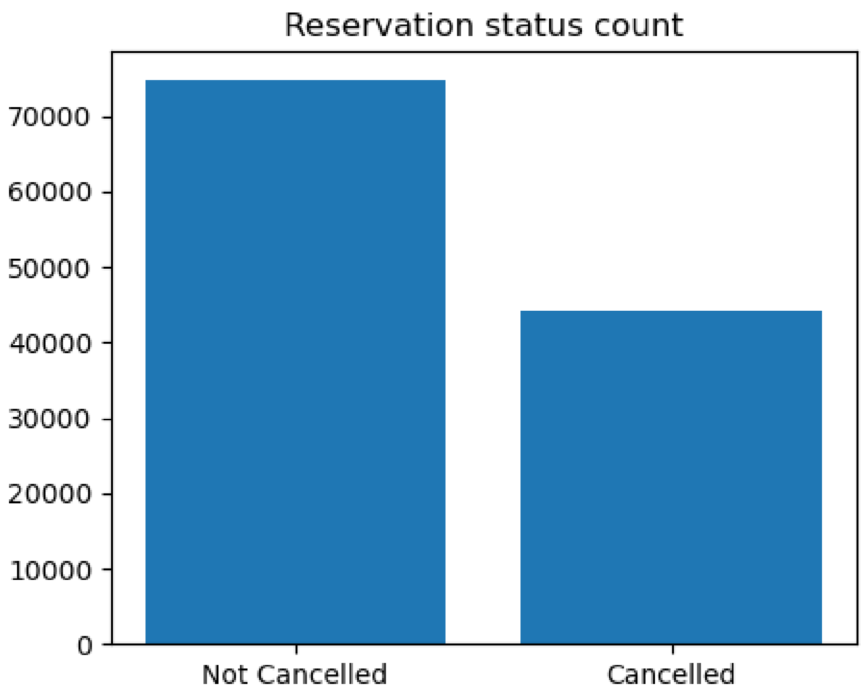
[45]:
```
cancel_percent = df['is_canceled'].value_counts(normalize = True)
# (normalize = True for the data present in percent  for all the rows in column␣
 ↪like 62% didnt cancel the booking
```

[46]:
```
cancel_percent
```

[46]:
```
is_canceled
0    0.628653
1    0.371347
Name: proportion, dtype: float64
```

[47]:
```
cancel_percent = df['is_canceled'].value_counts(normalize = True)
print(cancel_percent)

plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['Not Cancelled','Cancelled'],df['is_canceled'].value_counts())
plt.show()
```

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



[48]:
```
df.columns
```

[48]:
```
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', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

1.2 Converting the datatype of 'is_canceled' from int to string for visualization

[49]:
```
df['is_canceled'] = df['is_canceled'].astype(str)
```

[50]:
```
df.info()
```

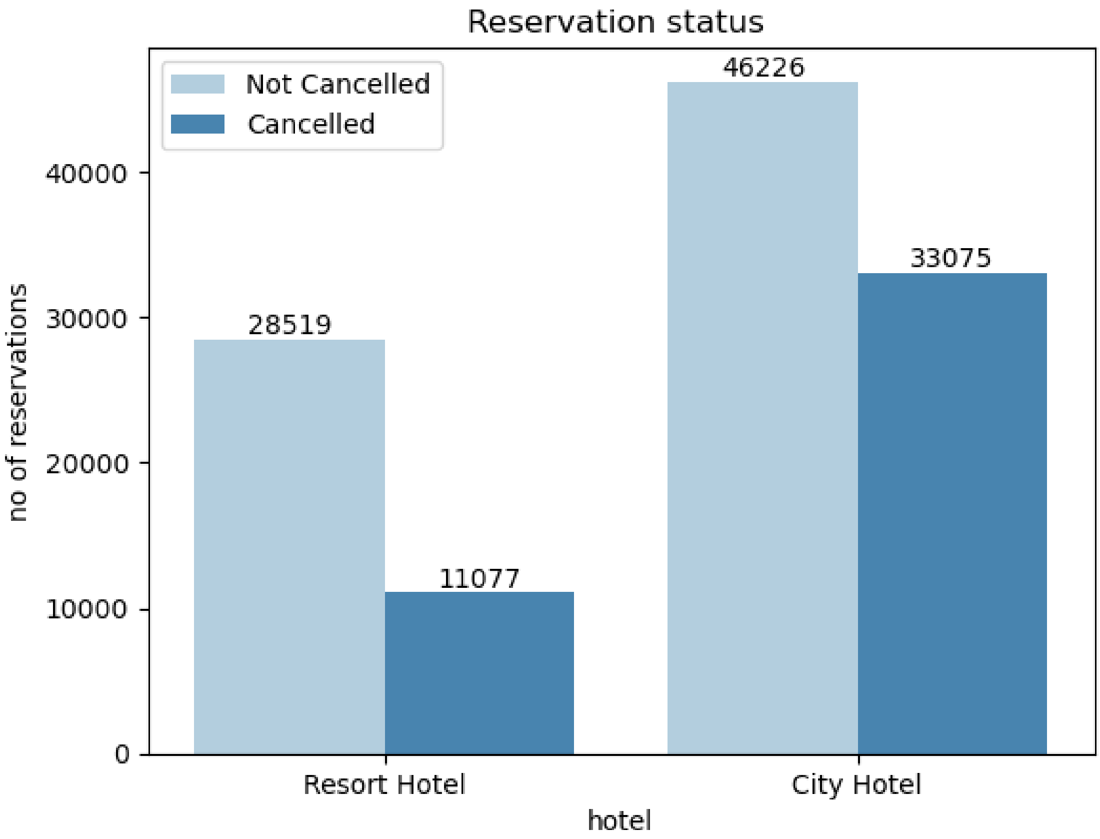
```
<class 'pandas.core.frame.DataFrame'>
Index: 118897 entries, 0 to 119389
```

```
Data columns (total 34 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  object
 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  name                            118897 non-null  object
 31  email                           118897 non-null  object
 32  phone-number                    118897 non-null  object
 33  credit_card                     118897 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(15), object(16)
memory usage: 31.7+ MB
```

[ ]:

2 Reservation status for each hotels

[51]:
```
#plt.figure(figsize=(8,4))
#ax= sns.countplot( x='hotel',hue = 'is_canceled',data= df, palette = 'Blues')
#legend_labels,_ = ax.get_legend_handles_labels()
#ax.legend(bbox_to_anchor(1,1))    # bbox_to_anchor(1,1 is not defined
#plt.title('Reservation status ',size =20)
#plt.xlabel('hotel')
#plt.ylabel('no of reservations')
#plt.legend(['Not Cancelled','Cancelled'])
#plt.show()


ax = sns.countplot (  data=df, x= 'hotel', hue= 'is_canceled',palette = 'Blues')
for bars in ax.containers : ax.bar_label(bars)
plt.title('Reservation status ')
plt.xlabel('hotel')
plt.ylabel('no of reservations')
plt.legend(['Not Cancelled','Cancelled'])
plt.show()
```



2.1 Percentage of cancellation for both hotels

[52]:
```
df['is_canceled'] = df['is_canceled'].astype(float)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 118897 entries, 0 to 119389
Data columns (total 34 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  float64
 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  name                            118897 non-null  object
 31  email                           118897 non-null  object
 32  phone-number                    118897 non-null  object
 33  credit_card                     118897 non-null  object
dtypes: datetime64[ns](1), float64(3), int64(15), object(15)
memory usage: 31.7+ MB
```

[53]:

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

[53]:

```
is_canceled
0.0    0.72025
1.0    0.27975
Name: proportion, dtype: float64
```

[54]:

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

[54]:

```
is_canceled
0.0    0.582918
1.0    0.417082
Name: proportion, dtype: float64
```

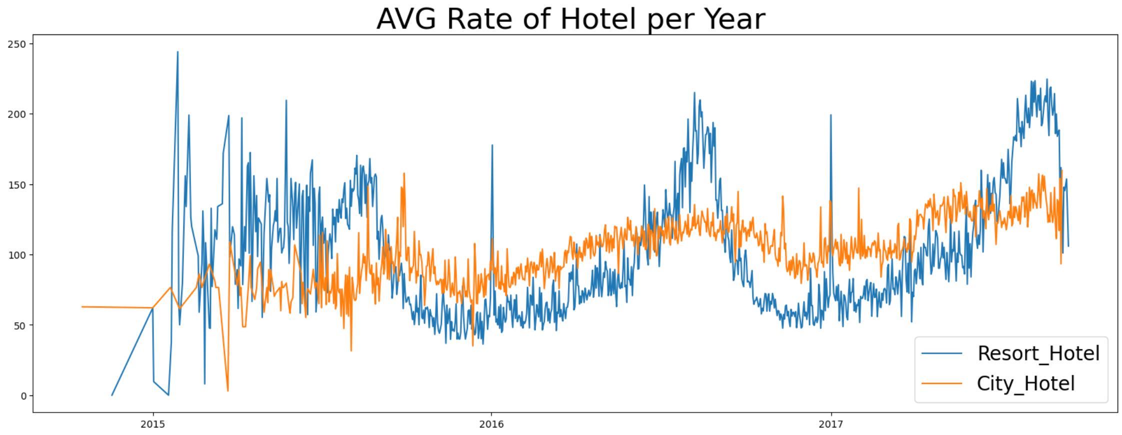
3 AVG Rate of Hotel per Year

[55]:

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

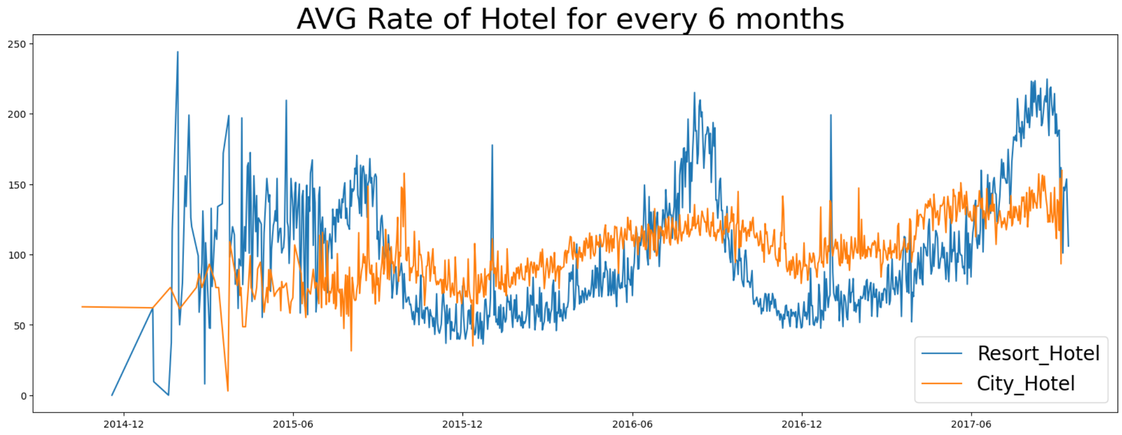
[56]:

```python
import matplotlib.dates as mdates
import pandas as pd

# Convert the datetime index to a datetime format
resort_hotel.index = pd.to_datetime(resort_hotel.index)
city_hotel.index = pd.to_datetime(city_hotel.index)

plt.figure(figsize=(20,7))

plt.title('AVG Rate of Hotel per Year ', 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)

# Set the x-axis tick labels to display years
years = mdates.YearLocator(base=1)   # Locate the years on the x-axis, base=1␣
 ↪for yearly ticks
years_fmt = mdates.DateFormatter('%Y')  # Format the years
plt.gca().xaxis.set_major_locator(years)
plt.gca().xaxis.set_major_formatter(years_fmt)

plt.show()
```



4 AVG Rate of Hotel for every 6 months

[57]:
```
import matplotlib.dates as mdates
import pandas as pd

# Convert the datetime index to a datetime format
resort_hotel.index = pd.to_datetime(resort_hotel.index)
city_hotel.index = pd.to_datetime(city_hotel.index)

plt.figure(figsize=(20,7))

plt.title('AVG Rate of Hotel for every 6 months ', 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)

# Set the x-axis tick labels to display every 6 months
months = mdates.MonthLocator(interval=6)   # Locate every 6 months on the x-axis
months_fmt = mdates.DateFormatter('%Y-%m')  # Format the months as 'YYYY-MM'
plt.gca().xaxis.set_major_locator(months)
plt.gca().xaxis.set_major_formatter(months_fmt)

plt.show()
```



```
[35]: df['is_canceled'] = df['is_canceled'].astype(str)
```

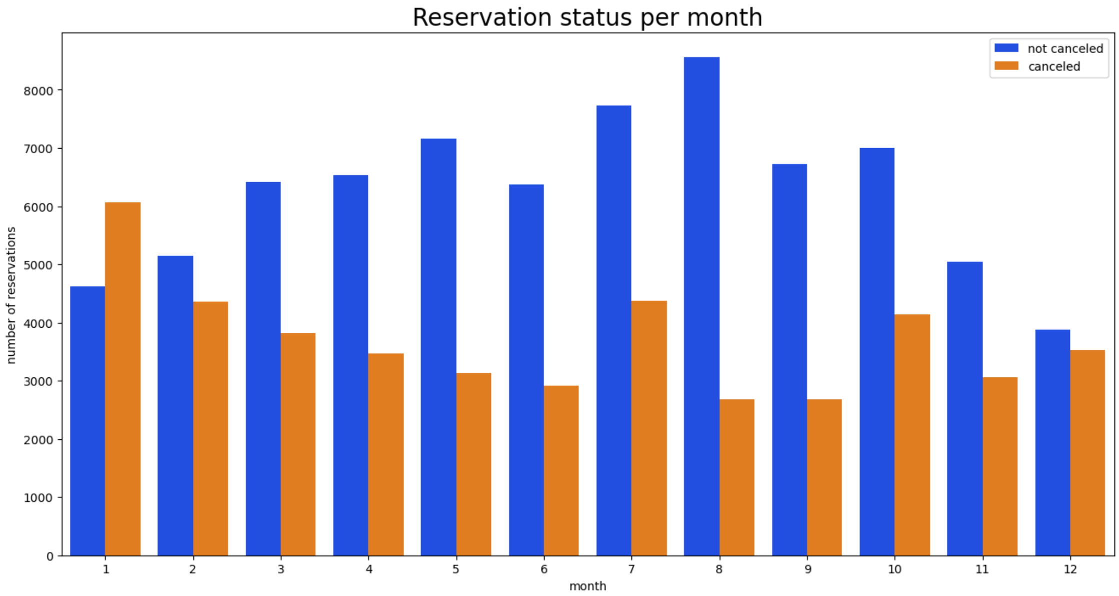
```
[36]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 37 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  object
 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]
 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
 36  month                           119390 non-null  int32
dtypes: datetime64[ns](1), float64(4), int32(1), int64(15), object(16)
memory usage: 33.2+ MB
```

5 Reservation status per month

[38]:

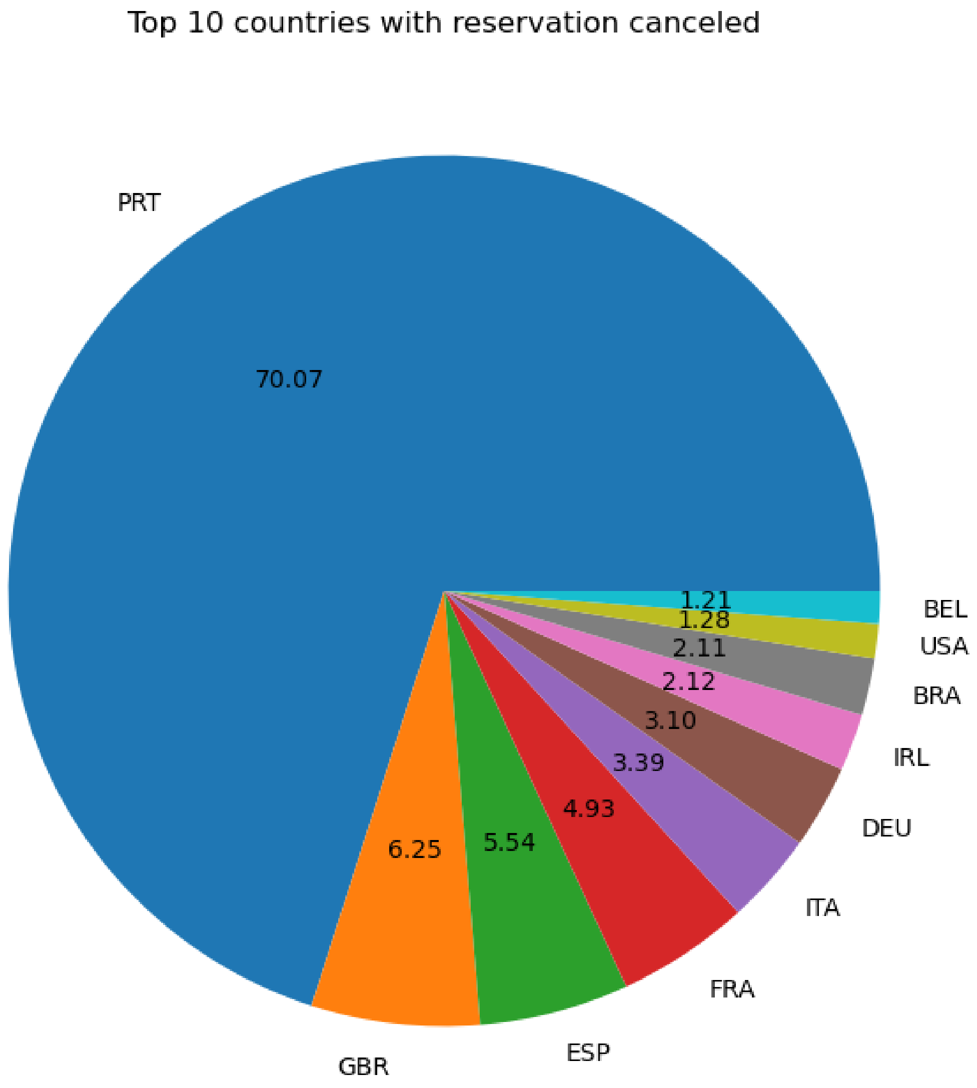
```python
# Create a new column 'month' in the DataFrame 'df' by extracting the month from the 'reservation_status_date' column
df['month'] = df['reservation_status_date'].dt.month
# Create a new figure with a specified size
plt.figure(figsize =(16,8))
# Create a count plot using Seaborn
# 'x' parameter specifies the column to use for the x-axis (in this case, 'month')
# 'hue' parameter specifies the column to use for creating different color-coded subsets (in this case, 'is_canceled')
# 'data' parameter specifies the DataFrame to use
# 'palette' parameter specifies the color palette to use
ax= sns.countplot(x= 'month', hue ='is_canceled', data =df, palette ='bright')
# Get the legend handles and labels from the count plot
legend_labels,_= ax. get_legend_handles_labels()
# Adjust the position of the legend
ax.legend(bbox_to_anchor= (1,1))
# Set the title of the plot
plt.title('Reservation status per month', size=20)
# Set the label for the x-axis
plt.xlabel('month')
# Set the label for the y-axis
plt.ylabel('number of reservations')
# Customize the legend labels
plt.legend(['not canceled', 'canceled'])
# Display the plot
plt.show()
```



6 Top 10 countries with reservation cancelled

[72]:

```
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()
```



```
[6]: df.columns
```

```
[6]: 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')
```

6.1 Checking out the total count of online and ofline booking

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

```
[8]: market_segment
     Online TA        56477
     Offline TA/TO    24219
     Groups           19811
     Direct           12606
     Corporate         5295
     Complementary      743
     Aviation           237
     Undefined            2
     Name: count, dtype: int64
```

```
[10]: df['market_segment'].value_counts(normalize=True)
      # percentage
```

```
[10]: market_segment
      Online TA        0.473046
      Offline TA/TO    0.202856
      Groups           0.165935
      Direct           0.105587
      Corporate        0.044350
      Complementary    0.006223
      Aviation         0.001985
      Undefined        0.000017
      Name: proportion, dtype: float64
```

7 Checking the total count of online and offline cancellation

```
[13]: import pandas as pd

      # Load data from a CSV file
      cancelled_data = pd.read_csv('hotel_booking.csv')

      # Now you can use cancelled_data
      cancelled_data['market_segment'].value_counts(normalize=True)
```

```
[13]: market_segment
      Online TA        0.473046
      Offline TA/TO    0.202856
```

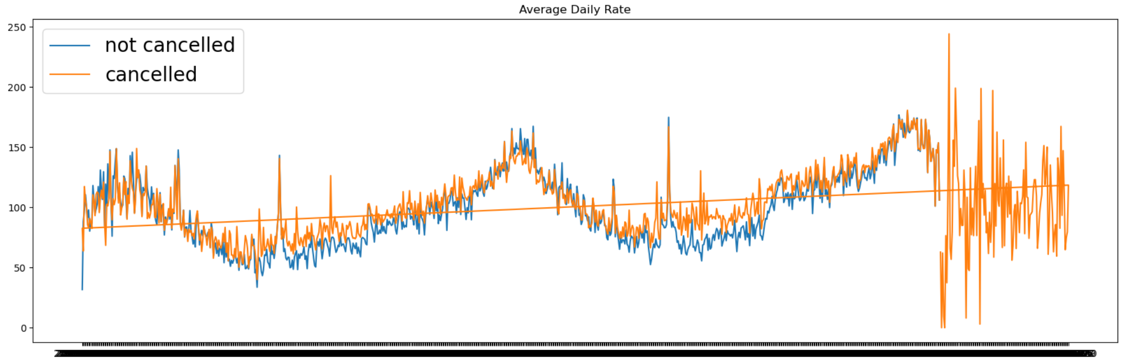
```
Groups           0.165935
Direct           0.105587
Corporate        0.044350
Complementary    0.006223
Aviation         0.001985
Undefined        0.000017
Name: proportion, dtype: float64
```

7.1 Checking the total count of online and offline cancellation over the year by graph

[25]:

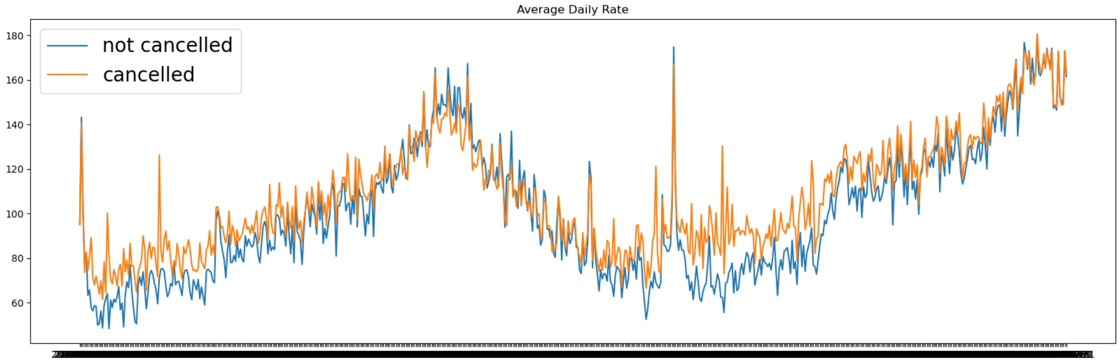
```
# Group the cancelled_data DataFrame by 'reservation_status_date' and calculate␣
  ↪the mean of 'adr' for each group
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[['adr']].
  ↪mean()
# Reset the index of cancelled_df_adr to turn the date into a column
cancelled_df_adr.reset_index(inplace=True)
# Sort the values in cancelled_df_adr by 'reservation_status_date' in ascending␣
  ↪order
cancelled_df_adr.sort_values('reservation_status_date', inplace=True)
# Create a new DataFrame 'not_cancelled_data' by filtering the original 'df'␣
  ↪DataFrame
# to include only rows where 'is_canceled' is 0 (not cancelled)
not_cancelled_data = df[df['is_canceled'] == 0]
# Similar to cancelled_df_adr, group not_cancelled_data by␣
  ↪'reservation_status_date'
# and calculate the mean of 'adr' for each group
not_cancelled_df_adr = not_cancelled_data.
  ↪groupby('reservation_status_date')[['adr']].mean()
# Reset the index of not_cancelled_df_adr to turn the date into a column
not_cancelled_df_adr.reset_index(inplace=True)
# Sort the values in not_cancelled_df_adr by 'reservation_status_date' in␣
  ↪ascending order
not_cancelled_df_adr.sort_values('reservation_status_date', inplace=True)
# Create a new figure with a specified size
plt.figure(figsize=(20, 6))
# Set the title of the plot
plt.title('Average Daily Rate')
# Plot the 'adr' column of not_cancelled_df_adr against␣
  ↪'reservation_status_date'
# and add a label to the line
plt.plot(not_cancelled_df_adr['reservation_status_date'],␣
  ↪not_cancelled_df_adr['adr'], label='not cancelled')
# Plot the 'adr' column of cancelled_df_adr against 'reservation_status_date'
# and add a label to the line
```

```
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label='cancelled')
plt.legend(fontsize =20)
plt.show()
```



[22]:
```
# Filter the cancelled_df_adr DataFrame to include only rows
# where the 'reservation_status_date' is greater than '2016' (i.e., after␣
 ↪December 31, 2016)
# and less than '2017-09' (i.e., before September 2017)
cancelled_df_adr =␣
 ↪cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') &␣
 ↪(cancelled_df_adr['reservation_status_date']<'2017-09')]
# Filter the not_cancelled_df_adr DataFrame to include only rows
# where the 'reservation_status_date' is greater than '2016' (i.e., after␣
 ↪December 31, 2016)
# and less than '2017-09' (i.e., before September 2017)
not_cancelled_df_adr =␣
 ↪not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016')␣
 ↪& (not_cancelled_df_adr['reservation_status_date']<'2017-09')]
```

[23]:
```
plt.figure(figsize=(20, 6))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'],␣
 ↪not_cancelled_df_adr['adr'], label='not cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label='cancelled')
plt.legend(fontsize =20)
plt.show()
```


Average Daily Rate
not cancelled
cancelled
180
160
140
120
100
80
60

8 END