hotel-data-analysis-booking-data

April 14, 2024

[ ]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib .pyplot as plt
%matplotlib inline
import csv
import pathlib as path
```

[ ]:
```
df=pd.read_csv('hotel_bookings 2.csv',encoding='unicode_escape')
```

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

[ ]:

	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	0.0	Transient	0.0	
1	NaN	NaN	0.0	Transient	0.0	
2	NaN	NaN	0.0	Transient	75.0	

```
3  304.0     NaN                  0.0     Transient  75.0
4  240.0     NaN                  0.0     Transient  98.0

   required_car_parking_spaces  total_of_special_requests  reservation_status  \
0                          0.0                        0.0           Check-Out
1                          0.0                        0.0           Check-Out
2                          0.0                        0.0           Check-Out
3                          0.0                        0.0           Check-Out
4                          0.0                        1.0           Check-Out

  reservation_status_date
0                1/7/2015
1                1/7/2015
2                2/7/2015
3                2/7/2015
4                3/7/2015

[5 rows x 32 columns]
```

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

[ ]:
```
            hotel  is_canceled  lead_time  arrival_date_year  \
51412  City Hotel            1        322               2016
51413  City Hotel            1        322               2016
51414  City Hotel            1        322               2016
51415  City Hotel            1         25               2016
51416  City Hotel            1        116               2016

      arrival_date_month  arrival_date_week_number  arrival_date_day_of_month  \
51412                May                        21                         19
51413                May                        21                         19
51414                May                        21                         19
51415                May                        21                         19
51416                May                        21                         19

       stays_in_weekend_nights  stays_in_week_nights  adults  …  \
51412                        1                     3       2  …
51413                        1                     3       2  …
51414                        1                     3       2  …
51415                        1                     3       2  …
51416                        1                     3       2  …

       deposit_type  agent company days_in_waiting_list customer_type    adr  \
51412    Non Refund   31.0     NaN                120.0     Transient   80.0
51413    Non Refund   31.0     NaN                120.0     Transient   80.0
51414    Non Refund   31.0     NaN                120.0     Transient   80.0
51415    No Deposit    7.0     NaN                  0.0     Transient  158.0
```

```
51416           NaN    NaN     NaN                   NaN           NaN    NaN

       required_car_parking_spaces  total_of_special_requests  \
51412                          0.0                        0.0
51413                          0.0                        0.0
51414                          0.0                        0.0
51415                          0.0                        0.0
51416                          NaN                        NaN

       reservation_status reservation_status_date
51412           Canceled               10/2/2016
51413           Canceled               10/2/2016
51414           Canceled               10/2/2016
51415           Canceled               14/5/2016
51416                NaN                     NaN

[5 rows x 32 columns]
```

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

[ ]:
```
(51417, 32)
```

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

[ ]:
```
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')
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51417 entries, 0 to 51416
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           51417 non-null  object
 1   is_canceled                     51417 non-null  int64
 2   lead_time                       51417 non-null  int64
```

```
 3   arrival_date_year               51417 non-null  int64
 4   arrival_date_month              51417 non-null  object
 5   arrival_date_week_number        51417 non-null  int64
 6   arrival_date_day_of_month       51417 non-null  int64
 7   stays_in_weekend_nights         51417 non-null  int64
 8   stays_in_week_nights            51417 non-null  int64
 9   adults                          51417 non-null  int64
 10  children                        51413 non-null  float64
 11  babies                          51417 non-null  int64
 12  meal                            51417 non-null  object
 13  country                         50939 non-null  object
 14  market_segment                  51417 non-null  object
 15  distribution_channel            51417 non-null  object
 16  is_repeated_guest               51416 non-null  float64
 17  previous_cancellations          51416 non-null  float64
 18  previous_bookings_not_canceled  51416 non-null  float64
 19  reserved_room_type              51416 non-null  object
 20  assigned_room_type              51416 non-null  object
 21  booking_changes                 51416 non-null  float64
 22  deposit_type                    51416 non-null  object
 23  agent                           42552 non-null  float64
 24  company                         3359 non-null   float64
 25  days_in_waiting_list            51416 non-null  float64
 26  customer_type                   51416 non-null  object
 27  adr                             51416 non-null  float64
 28  required_car_parking_spaces     51416 non-null  float64
 29  total_of_special_requests       51416 non-null  float64
 30  reservation_status              51416 non-null  object
 31  reservation_status_date         51416 non-null  object
dtypes: float64(11), int64(9), object(12)
memory usage: 12.6+ MB
```

#Changing the data type of reservation_status_date

[ ]:
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'],␣
  ↪format="%d/%m/%Y")
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51417 entries, 0 to 51416
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           51417 non-null  object
 1   is_canceled                     51417 non-null  int64
 2   lead_time                       51417 non-null  int64
 3   arrival_date_year               51417 non-null  int64
```

```
 4   arrival_date_month              51417 non-null  object
 5   arrival_date_week_number        51417 non-null  int64
 6   arrival_date_day_of_month       51417 non-null  int64
 7   stays_in_weekend_nights         51417 non-null  int64
 8   stays_in_week_nights            51417 non-null  int64
 9   adults                          51417 non-null  int64
 10  children                        51413 non-null  float64
 11  babies                          51417 non-null  int64
 12  meal                            51417 non-null  object
 13  country                         50939 non-null  object
 14  market_segment                  51417 non-null  object
 15  distribution_channel            51417 non-null  object
 16  is_repeated_guest               51416 non-null  float64
 17  previous_cancellations          51416 non-null  float64
 18  previous_bookings_not_canceled  51416 non-null  float64
 19  reserved_room_type              51416 non-null  object
 20  assigned_room_type              51416 non-null  object
 21  booking_changes                 51416 non-null  float64
 22  deposit_type                    51416 non-null  object
 23  agent                           42552 non-null  float64
 24  company                         3359 non-null   float64
 25  days_in_waiting_list            51416 non-null  float64
 26  customer_type                   51416 non-null  object
 27  adr                             51416 non-null  float64
 28  required_car_parking_spaces     51416 non-null  float64
 29  total_of_special_requests       51416 non-null  float64
 30  reservation_status              51416 non-null  object
 31  reservation_status_date         51416 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(11), int64(9), object(11)
memory usage: 12.6+ MB
```

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

[ ]:
```
                hotel arrival_date_month   meal country market_segment  \
count           51417              51417  51417   50939          51417
unique              2                 12      5     136              8
top     Resort Hotel             August     BB     PRT      Online TA
freq            40060               6472  38826   23566          22223

       distribution_channel reserved_room_type assigned_room_type  \
count                 51417              51416              51416
unique                    6                 10                 12
top                   TA/TO                  A                  A
freq                  39354              33323              25683

       deposit_type customer_type reservation_status
count         51416         51416              51416
```

```
unique            3             4                   3
top      No Deposit     Transient           Check-Out
freq          47383         36578               34210
```

#Displaying the unique data in the non numerical columns

[ ]:
```
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']
--------------------------------------------------
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']
--------------------------------------------------
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS' 'TA']
--------------------------------------------------
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B' nan]
--------------------------------------------------
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K' nan]
--------------------------------------------------
```

```
deposit_type
['No Deposit' 'Refundable' 'Non Refund' nan]
--------------------------------------------------
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group' nan]
--------------------------------------------------
reservation_status
['Check-Out' 'Canceled' 'No-Show' nan]
--------------------------------------------------
```

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

[ ]:
```
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                           478
market_segment                      0
distribution_channel                0
is_repeated_guest                   1
previous_cancellations              1
previous_bookings_not_canceled      1
reserved_room_type                  1
assigned_room_type                  1
booking_changes                     1
deposit_type                        1
agent                            8865
company                         48058
days_in_waiting_list                1
customer_type                       1
adr                                 1
required_car_parking_spaces         1
total_of_special_requests           1
reservation_status                  1
reservation_status_date             1
dtype: int64
```

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

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

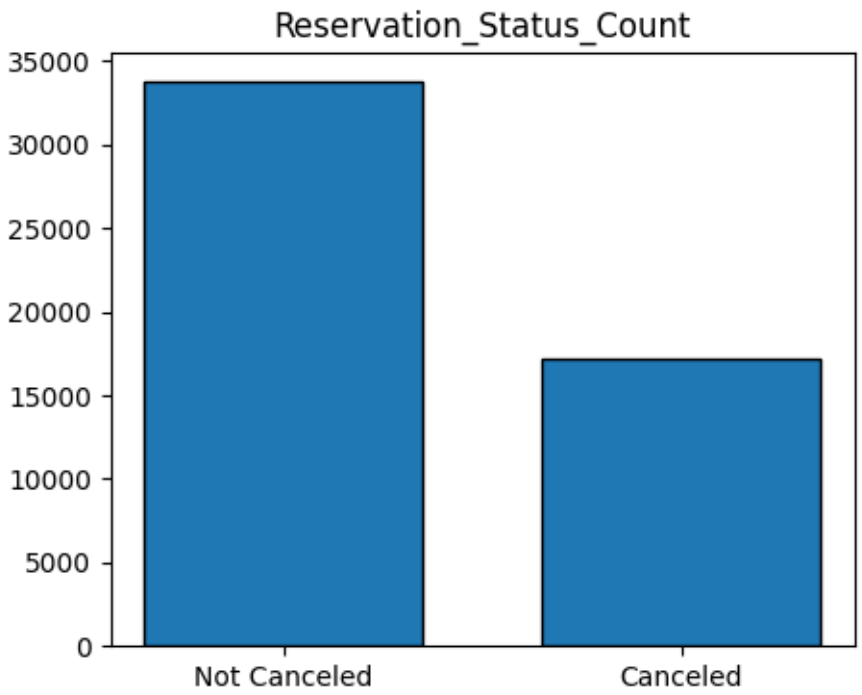
[ ]:
```
        is_canceled     lead_time  arrival_date_year  \
count  50934.000000  50934.000000       50934.000000
mean       0.336573     90.887659        2015.971748
min        0.000000      0.000000        2015.000000
25%        0.000000     13.000000        2015.000000
50%        0.000000     61.000000        2016.000000
75%        1.000000    142.000000        2017.000000
max        1.000000    737.000000        2017.000000
std        0.472542     93.732072           0.736109

       arrival_date_week_number  arrival_date_day_of_month  \
count              50934.000000               50934.000000
mean                  27.228315                  15.801842
min                    1.000000                   1.000000
25%                   16.000000                   8.000000
50%                   29.000000                  16.000000
75%                   38.000000                  23.000000
max                   53.000000                  31.000000
std                   13.908198                   8.797406

       stays_in_weekend_nights  stays_in_week_nights        adults  \
count             50934.000000          50934.000000  50934.000000
mean                  1.111831              2.933777      1.862116
min                   0.000000              0.000000      0.000000
25%                   0.000000              1.000000      2.000000
50%                   1.000000              2.000000      2.000000
75%                   2.000000              4.000000      2.000000
max                  16.000000             40.000000     55.000000
std                   1.107791              2.303412      0.651289

           children        babies  is_repeated_guest  previous_cancellations  \
count  50934.000000  50934.000000       50934.000000            50934.000000
mean       0.116033      0.011976           0.034829                0.079220
min        0.000000      0.000000           0.000000                0.000000
25%        0.000000      0.000000           0.000000                0.000000
50%        0.000000      0.000000           0.000000                0.000000
75%        0.000000      0.000000           0.000000                0.000000
max       10.000000     10.000000           1.000000               26.000000
std        0.425796      0.117959           0.183349                1.184453

       previous_bookings_not_canceled  booking_changes  days_in_waiting_list  \
count                    50934.000000     50934.000000          50934.000000
```

```
mean                          0.101111         0.260553              2.640908
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                          30.000000        20.000000            259.000000
std                           0.825892         0.696316             17.567284

                adr  required_car_parking_spaces  total_of_special_requests  \
count  50934.000000                 50934.000000               50934.000000
mean      94.304132                     0.109338                   0.554364
min       -6.380000                     0.000000                   0.000000
25%       56.570000                     0.000000                   0.000000
50%       79.500000                     0.000000                   0.000000
75%      117.000000                     0.000000                   1.000000
max     5400.000000                     8.000000                   5.000000
std       61.024978                     0.317306                   0.786958

            reservation_status_date
count                         50934
mean   2016-06-03 02:06:59.837436928
min              2014-11-18 00:00:00
25%              2015-11-15 00:00:00
50%              2016-04-21 00:00:00
75%              2016-12-20 00:00:00
max              2017-09-14 00:00:00
std                              NaN
```

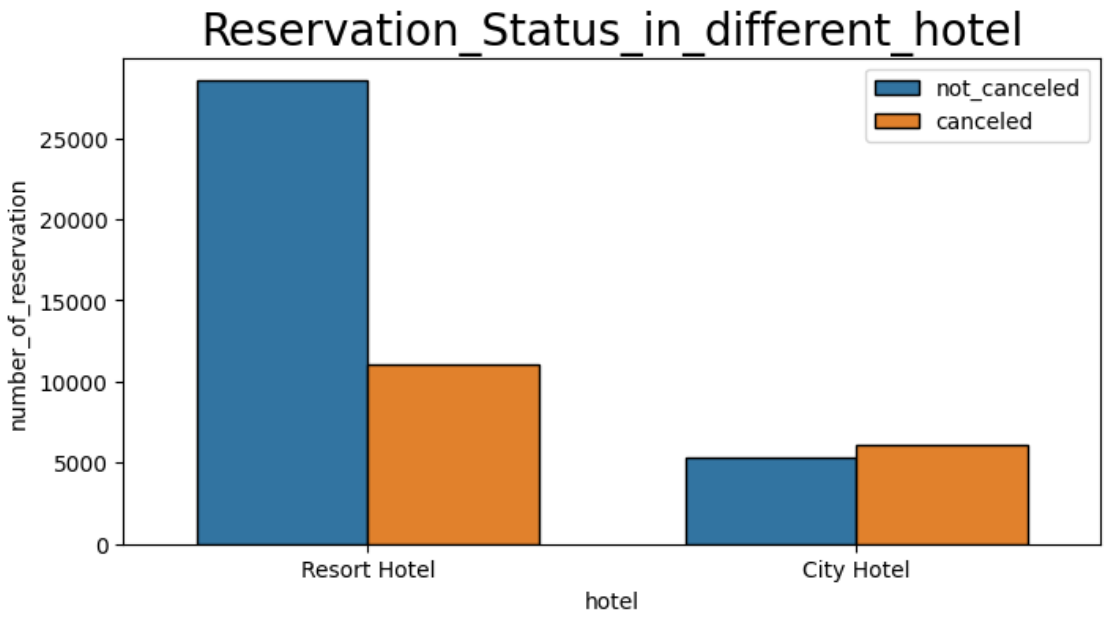
[ ]:
```
cancelled_prec=df['is_canceled'].value_counts(normalize=True)
cancelled_prec
```

[ ]:
```
is_canceled
0    0.663427
1    0.336573
Name: proportion, dtype: float64
```

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



[ ]:
```
plt.figure(figsize=(8,4))
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,edgecolor='k',width=0.7)
legend_labels=ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1, 1, 0, -0.2), loc='upper left')
plt.title('Reservation_Status_in_different_hotel',size=20)
plt.xlabel('hotel')
plt.ylabel('number_of_reservation')
plt.legend(['not_canceled','canceled'])
plt.show()
```



#Cancellation rate in resort

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

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

#Cancellation rate in hotel

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

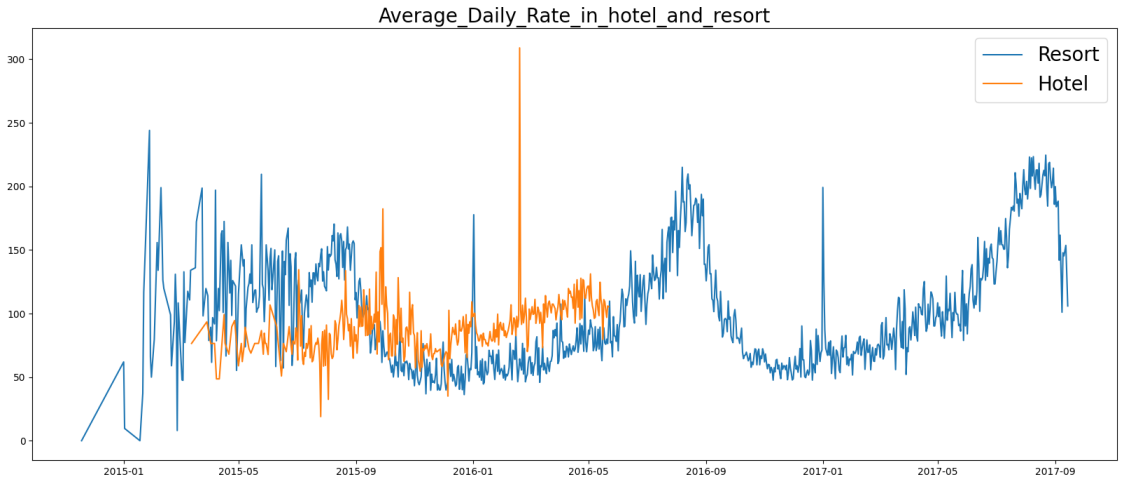
[ ]:
```
is_canceled
1    0.535015
0    0.464985
Name: proportion, dtype: float64
```

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

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

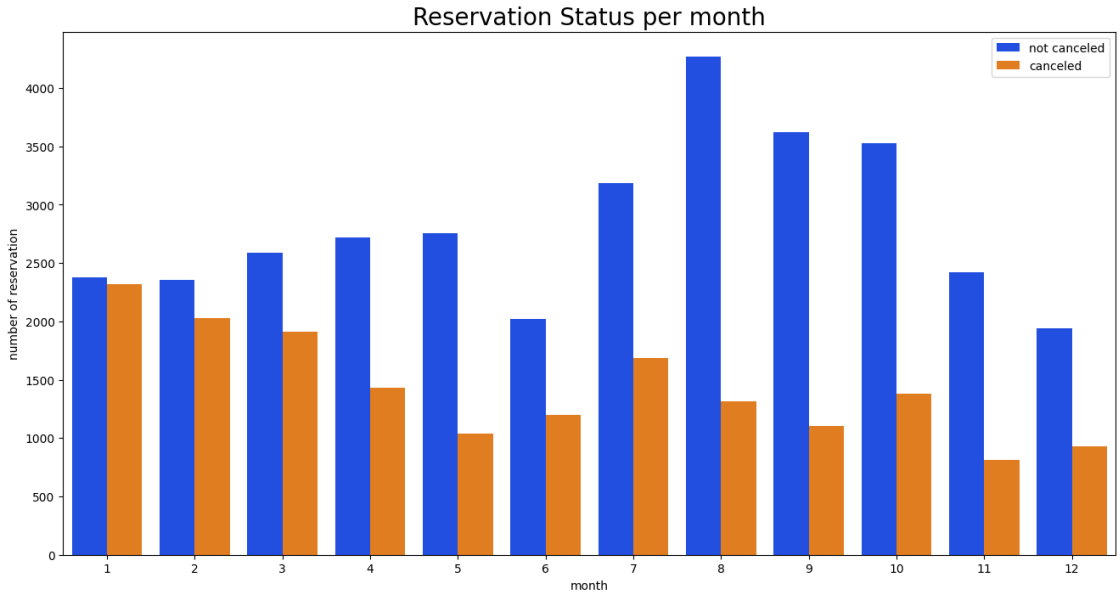
#ADR=AVERAGE DAILY RATE The average daily rate (ADR) measures the average rental revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using the ADR.

[ ]:
```
plt.figure(figsize=(20,8))
plt.title('Average_Daily_Rate_in_hotel_and_resort',size=20)
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort')
plt.plot(city_hotel.index,city_hotel['adr'],label='Hotel')
plt.legend(fontsize=20)
plt.show()
```



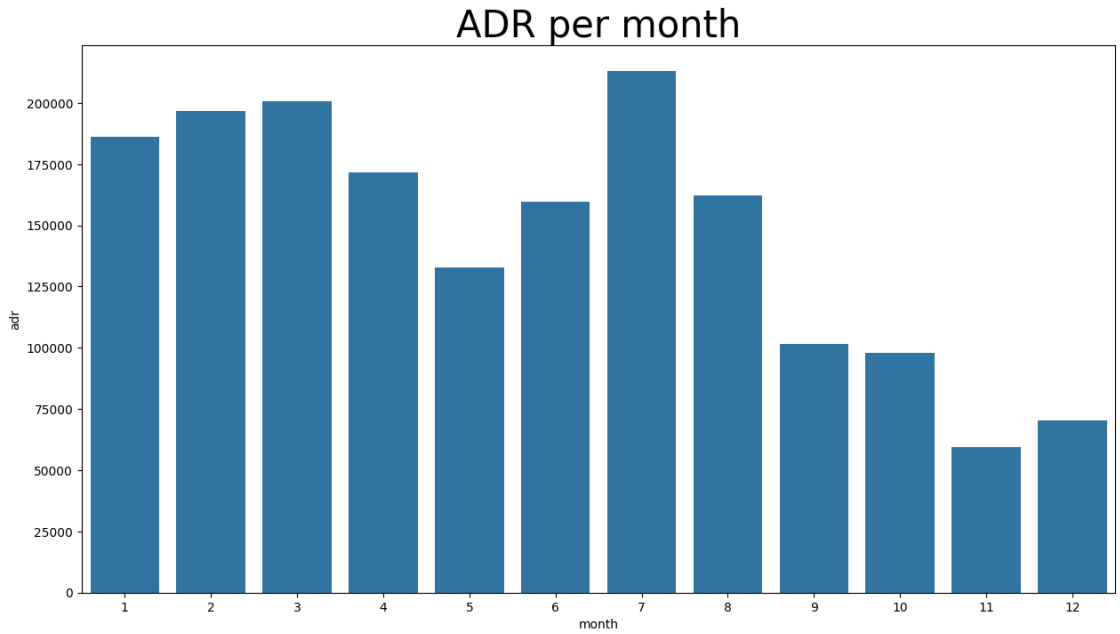
#Reservation Status per month

[ ]:
```
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')
plt.title("Reservation Status per month",size=20)
plt.xlabel('month')
plt.ylabel('number of reservation')
plt.legend(['not canceled','canceled'])
plt.show()
```



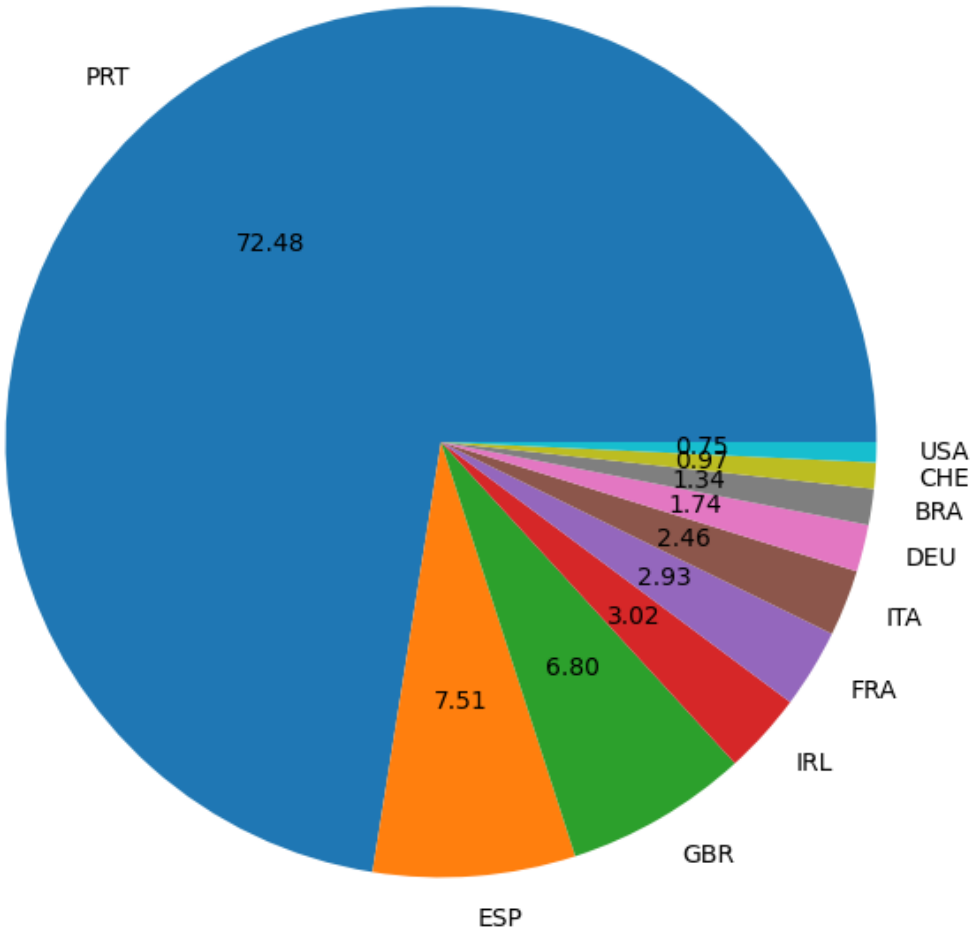
#ADR=AVERAGE DAILY RATE The average daily rate (ADR) measures the average rental revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using the ADR.

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



#TOP 10 COUNTRY WITH CANCELED BOOKING

[ ]:
```
top_10_country=df[df['is_canceled']==1]
top_10_country=top_10_country['country'].value_counts()[:10]
plt.figure(figsize=(8,8))
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.show()
```



#Market Segment Breakdown

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

[ ]:
```
market_segment
Online TA        22147
Offline TA/TO    10441
Groups            8777
Direct            6975
Corporate         2341
Complementary      245
Aviation             8
Name: count, dtype: int64
```

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

[ ]:
```
market_segment
Online TA        0.434818
Offline TA/TO    0.204991
Groups           0.172321
Direct           0.136942
Corporate        0.045961
Complementary    0.004810
Aviation         0.000157
Name: proportion, dtype: float64
```

[ ]:
```
cancelled_prec.index.dtype
```

[ ]:
```
dtype('int64')
```