Hotel_Booking_Project

January 16, 2024

1 Importing libraies

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

2 Loding the dataset

[2]:
```
df=pd.read_csv(r"C:\Users\omkar␣
↪mangrulkar\Desktop\DSA\AllDataSets\hotel_bookings.csv")
```

3 Exploratory Data Analysis and Data Clenaing

[3]:
```
df.head().T
```

[3]:

	0	1	2	\
hotel	Resort Hotel	Resort Hotel	Resort Hotel	
is_canceled	0	0	0	
lead_time	342	737	7	
arrival_date_year	2015	2015	2015	
arrival_date_month	July	July	July	
arrival_date_week_number	27	27	27	
arrival_date_day_of_month	1	1	1	
stays_in_weekend_nights	0	0	0	
stays_in_week_nights	0	0	1	
adults	2	2	1	
children	0.0	0.0	0.0	
babies	0	0	0	
meal	BB	BB	BB	
country	PRT	PRT	GBR	
market_segment	Direct	Direct	Direct	
distribution_channel	Direct	Direct	Direct	
is_repeated_guest	0	0	0	
previous_cancellations	0	0	0	

previous_bookings_not_canceled	0	0	0
reserved_room_type	C	C	A
assigned_room_type	C	C	C
booking_changes	3	4	0
deposit_type	No Deposit	No Deposit	No Deposit
agent	NaN	NaN	NaN
company	NaN	NaN	NaN
days_in_waiting_list	0	0	0
customer_type	Transient	Transient	Transient
adr	0.0	0.0	75.0
required_car_parking_spaces	0	0	0
total_of_special_requests	0	0	0
reservation_status	Check-Out	Check-Out	Check-Out
reservation_status_date	01-07-2015	01-07-2015	02-07-2015

	3	4
hotel	Resort Hotel	Resort Hotel
is_canceled	0	0
lead_time	13	14
arrival_date_year	2015	2015
arrival_date_month	July	July
arrival_date_week_number	27	27
arrival_date_day_of_month	1	1
stays_in_weekend_nights	0	0
stays_in_week_nights	1	2
adults	1	2
children	0.0	0.0
babies	0	0
meal	BB	BB
country	GBR	GBR
market_segment	Corporate	Online TA
distribution_channel	Corporate	TA/TO
is_repeated_guest	0	0
previous_cancellations	0	0
previous_bookings_not_canceled	0	0
reserved_room_type	A	A
assigned_room_type	A	A
booking_changes	0	0
deposit_type	No Deposit	No Deposit
agent	304.0	240.0
company	NaN	NaN
days_in_waiting_list	0	0
customer_type	Transient	Transient
adr	75.0	98.0
required_car_parking_spaces	0	0
total_of_special_requests	0	1
reservation_status	Check-Out	Check-Out

```
reservation_status_date           02-07-2015     03-07-2015
```

[4]:
```
df.shape
```

[4]:
```
(119390, 32)
```

[5]:
```
df.describe().T
```

[5]:

	count	mean	std	min	\
is_canceled	119390.0	0.370416	0.482918	0.00	
lead_time	119390.0	104.011416	106.863097	0.00	
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	
adults	119390.0	1.856403	0.579261	0.00	
children	119386.0	0.103890	0.398561	0.00	
babies	119390.0	0.007949	0.097436	0.00	
is_repeated_guest	119390.0	0.031912	0.175767	0.00	
previous_cancellations	119390.0	0.087118	0.844336	0.00	
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	
booking_changes	119390.0	0.221124	0.652306	0.00	
agent	103050.0	86.693382	110.774548	1.00	
company	6797.0	189.266735	131.655015	6.00	
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	
adr	119390.0	101.831122	50.535790	-6.38	
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	
total_of_special_requests	119390.0	0.571363	0.792798	0.00	

	25%	50%	75%	max
is_canceled	0.00	0.000	1.0	1.0
lead_time	18.00	69.000	160.0	737.0
arrival_date_year	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	16.00	28.000	38.0	53.0
arrival_date_day_of_month	8.00	16.000	23.0	31.0
stays_in_weekend_nights	0.00	1.000	2.0	19.0
stays_in_week_nights	1.00	2.000	3.0	50.0
adults	2.00	2.000	2.0	55.0
children	0.00	0.000	0.0	10.0
babies	0.00	0.000	0.0	10.0
is_repeated_guest	0.00	0.000	0.0	1.0
previous_cancellations	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	0.00	0.000	0.0	72.0
booking_changes	0.00	0.000	0.0	21.0
agent	9.00	14.000	229.0	535.0
company	62.00	179.000	270.0	543.0

```
days_in_waiting_list              0.00     0.000     0.0   391.0
adr                              69.29    94.575   126.0  5400.0
required_car_parking_spaces       0.00     0.000     0.0     8.0
total_of_special_requests         0.00     0.000     1.0     5.0
```

[6]:
```
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
```

[7]:
```
df.columns
```

[7]:
```
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')
```

[8]:
```
df['reservation_status_date']=pd.to_datetime(df['reservation_status_date'])
```

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

[10]:
```
df.describe(include=object).T
```

[10]:

	count	unique	top	freq
hotel	119390	2	City Hotel	79330
arrival_date_month	119390	12	August	13877
meal	119390	5	BB	92310
country	118902	177	PRT	48590
market_segment	119390	8	Online TA	56477
distribution_channel	119390	5	TA/TO	97870
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
customer_type	119390	4	Transient	89613
reservation_status	119390	3	Check-Out	75166

[11]:
```
df.hotel.value_counts()
```

[11]:
```
City Hotel      79330
Resort Hotel    40060
Name: hotel, dtype: int64
```

[12]:
```
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']
--------------------------------------------------
```

[13]:
```
df.isnull().sum()
```

[13]:
```
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
```

[14]:
```
df.drop(['agent','company'],axis=1,inplace=True)
df.dropna(inplace=True)
```

[15]:
```
df.isnull().sum()
```

[15]:
```
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
```

[16]:
```
df.describe()
```

[16]:

	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
```

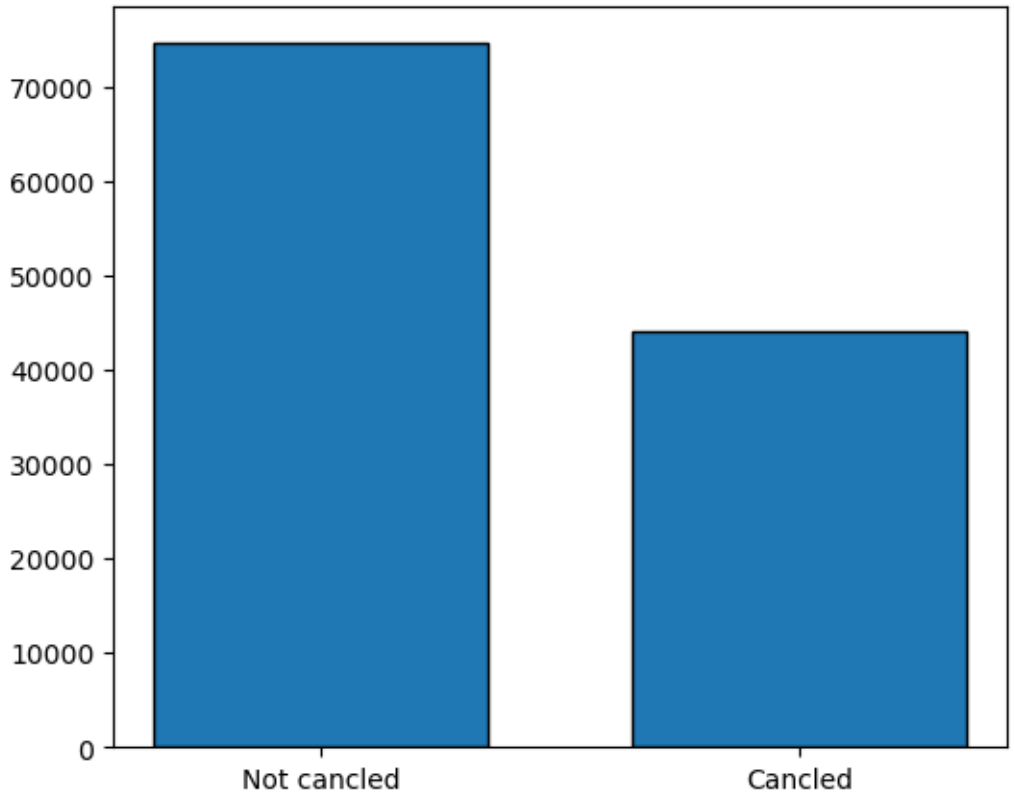
[17]:
```
df=df[df['adr']<5000]
```

4 Data Analysis and Visualizations

[18]:
```
cancled_perc=df['is_canceled'].value_counts(normalize=True)
print(cancled_perc)
plt.figure(figsize=(6,5))

plt.bar(['Not cancled','Cancled'],df['is_canceled'].
  ↪value_counts(),edgecolor='k',width=0.7)
plt.show()
```

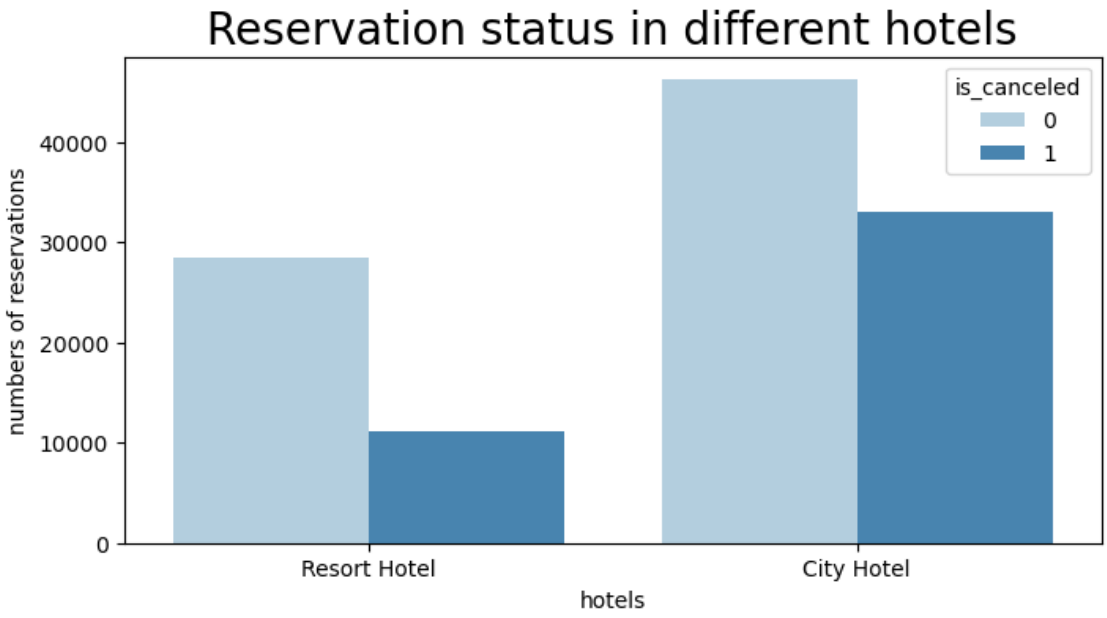
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```



[19]:
```
df['is_canceled'].value_counts()
```

[19]:
```
0    74745
1    44152
Name: is_canceled, dtype: int64
```

[20]:
```
plt.figure(figsize=(8,4))
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,palette='Blues')
plt.title('Reservation status in different hotels',size=20)
plt.xlabel('hotels')
plt.ylabel('numbers of reservations')
plt.show()
```



[21]:
```
# how many % of cancle and not cancle each hotels
resort_hotel=df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)
```

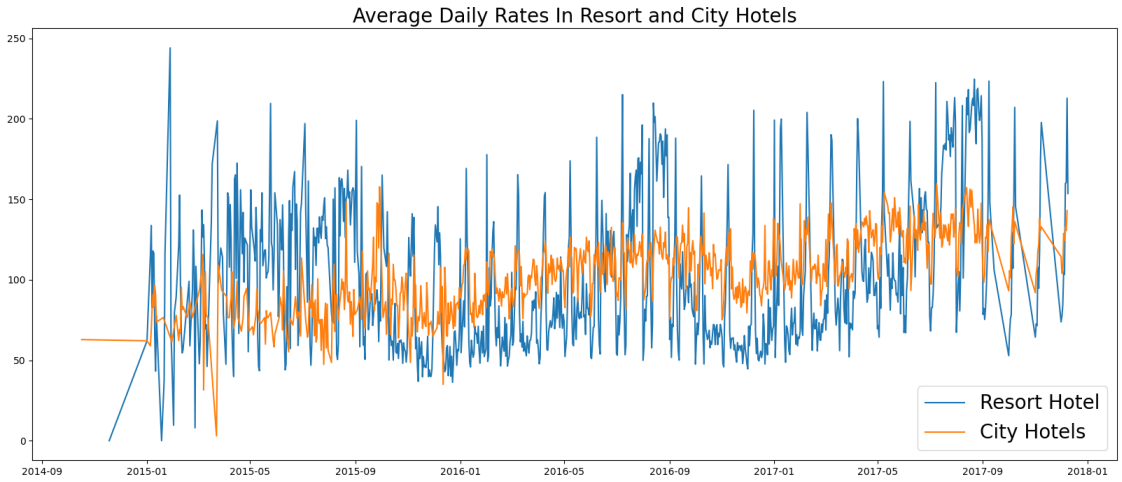
[21]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

[22]:
```
city_hotel=df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)
```

[22]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

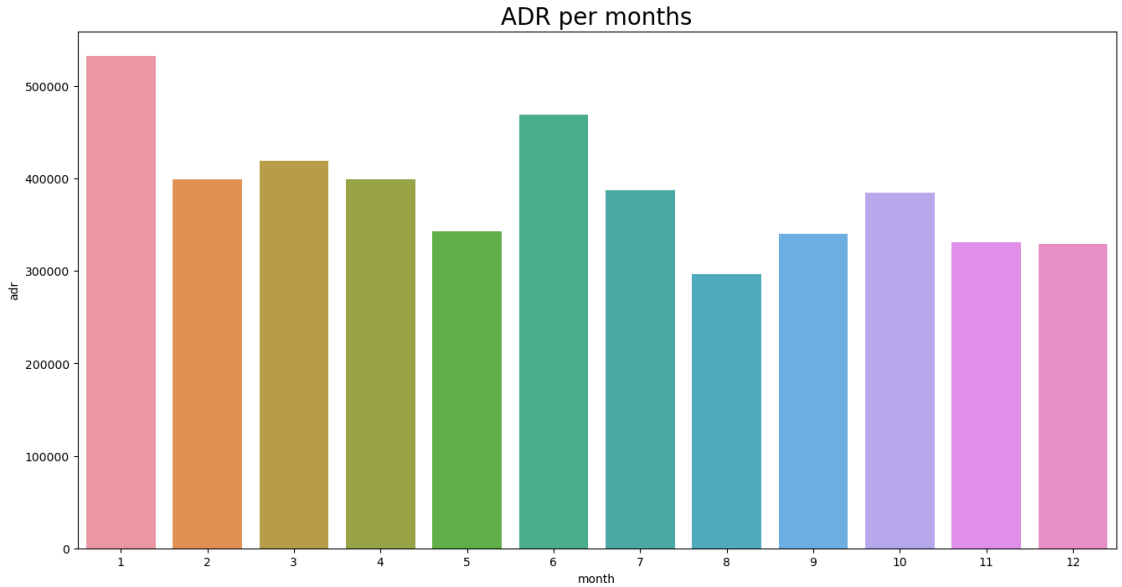
[23]:
```
resort_hotel=resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

[24]:
```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rates In Resort and City Hotels',fontsize=20)
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort Hotel')
plt.plot(city_hotel.index,city_hotel['adr'],label='City Hotels')
plt.legend(fontsize=20)
plt.show()
```

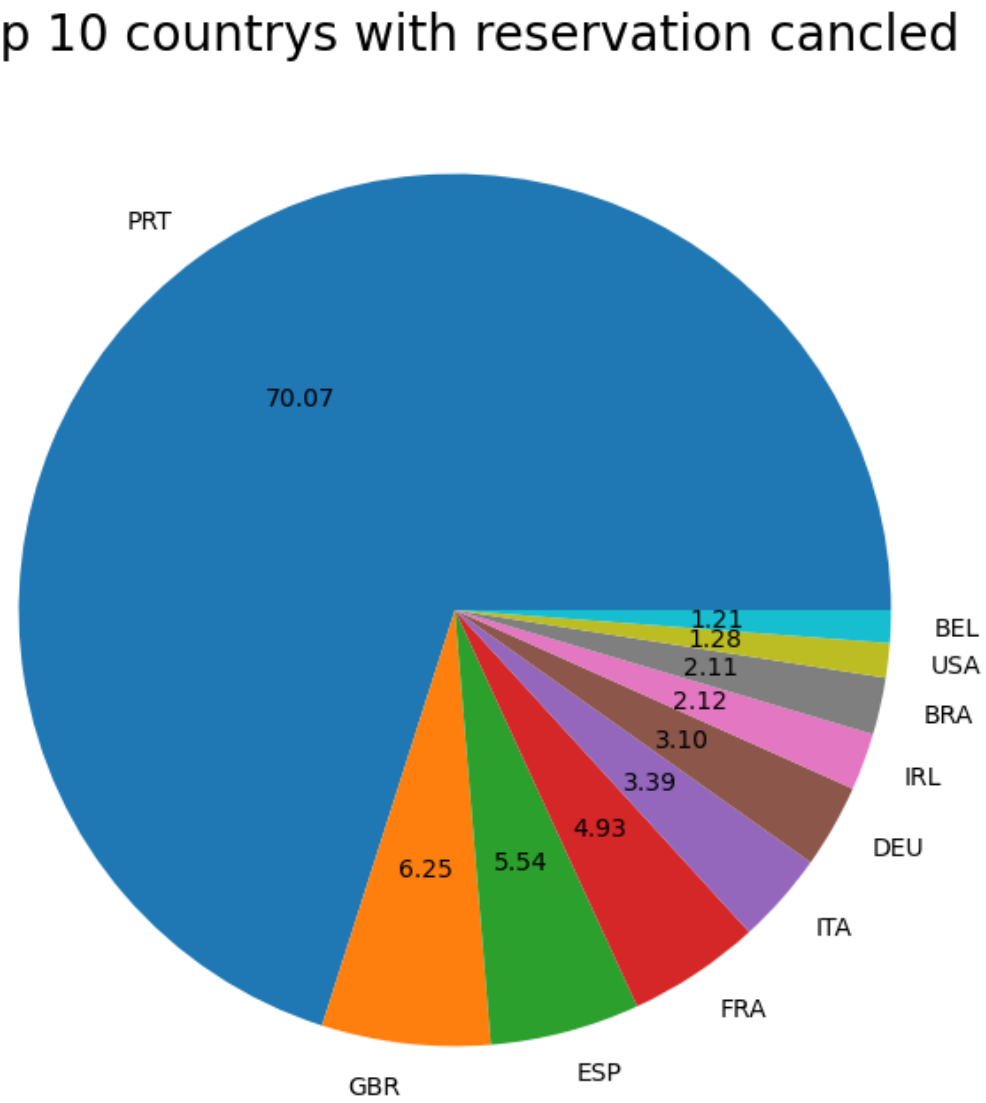


[25]:
```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
ax1=sns.countplot(x='month',hue='is_canceled',data=df,palette='prism')
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per months',size=20)
plt.xlabel('months')
plt.ylabel('numbers of reservations')
plt.legend(['Not Cancled','Cancled'])
plt.show()
```

[26]:
```
plt.figure(figsize=(16,8))
plt.title('ADR per months',size=20)
sns.barplot('month','adr',data=df[df['is_canceled']==1].
 ↪groupby('month')[['adr']].sum().reset_index())
plt.show()
```



[27]:
```
cancle_data=df[df['is_canceled']==1]
top_10_country=cancle_data['country'].value_counts()[:10]
plt.figure(figsize=(16,8))
plt.title('top 10 countrys with reservation cancled',size=20)
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.show()
```



[28]:
```
df['market_segment'].value_counts()
```

[28]:
```
Online TA        56402
Offline TA/TO    24159
```

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

[29]:
```
df['market_segment'].value_counts(normalize=True)
```

[29]:
```
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
```

[30]:
```
cancle_data['market_segment'].value_counts(normalize=True)
```

[30]:
```
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
```

[ ]:

[ ]: