

```
In [10]: import pandas as pd
import os
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

```
In [53]: os.getcwd()
```

```
Out[53]: 'C:\\Users\\deepa\\Downloads\\hotel_booking.csv'
```

Loading the dataset

```
In [12]: os.chdir(r"C:\Users\deepa\Downloads\hotel_booking.csv")
```

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

```
In [14]: df
```

```
Out[14]:
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_
0	Resort Hotel	0	342	2015	July		27
1	Resort Hotel	0	737	2015	July		27
2	Resort Hotel	0	7	2015	July		27
3	Resort Hotel	0	13	2015	July		27
4	Resort Hotel	0	14	2015	July		27
...
119385	City Hotel	0	23	2017	August		35
119386	City Hotel	0	102	2017	August		35
119387	City Hotel	0	34	2017	August		35
119388	City Hotel	0	109	2017	August		35
119389	City Hotel	0	205	2017	August		35

119390 rows × 36 columns

Exploratory Data analysis and data cleaning

```
In [31]: df.head()
```

```
Out[31]:
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_
0	Resort Hotel	0	342	2015	July		27
1	Resort Hotel	0	737	2015	July		27
2	Resort Hotel	0	7	2015	July		27
3	Resort Hotel	0	13	2015	July		27
4	Resort Hotel	0	14	2015	July		27

5 rows × 37 columns

```
In [65]: df.columns
```

```
Out[65]: 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')
```

```
In [9]: df.shape
```

```
Out[9]: (119390, 36)
```

```
In [64]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 119390 entries, 0 to 119389
```

```
Data columns (total 36 columns):
```

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	object
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_week_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119386 non-null	float64
11	babies	119390 non-null	int64
12	meal	119390 non-null	object
13	country	118902 non-null	object
14	market_segment	119390 non-null	object
15	distribution_channel	119390 non-null	object
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	reserved_room_type	119390 non-null	object
20	assigned_room_type	119390 non-null	object
21	booking_changes	119390 non-null	int64
22	deposit_type	119390 non-null	object
23	agent	103050 non-null	float64
24	company	6797 non-null	float64
25	days_in_waiting_list	119390 non-null	int64
26	customer_type	119390 non-null	object
27	adr	119390 non-null	float64
28	required_car_parking_spaces	119390 non-null	int64
29	total_of_special_requests	119390 non-null	int64
30	reservation_status	119390 non-null	object
31	reservation_status_date	119390 non-null	object
32	name	119390 non-null	object
33	email	119390 non-null	object
34	phone-number	119390 non-null	object
35	credit_card	119390 non-null	object

```
dtypes: float64(4), int64(16), object(16)
```

```
memory usage: 32.8+ MB
```

```
In [66]: #there is reservation_status_date in Object we are convert in date formate
df["reservation_status_date"] = pd.to_datetime(df["reservation_status_date"])
```

```
In [71]: 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	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
36	month	119390 non-null	int64

```
dtypes: datetime64[ns](1), float64(4), int64(17), object(15)
```

```
memory usage: 33.7+ MB
```

```
In [33]: df.describe(include= "object")
```

```
Out[33]:
```

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type
count	119390	119390	119390	118902	119390	119390	119390
unique	2	12	5	177	8	5	10
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A
freq	79330	13877	92310	48590	56477	97870	85994

```
In [18]: for col in df.describe(include='object').columns:
          print(col)
          print(df[col].unique())
          print('-'*50)
```

```

hotel
['Resort Hotel' 'City Hotel']
-----
arrival_date_month
['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
-----
meal
['BB' 'FB' 'HB' 'SC' 'Undefined']
-----
country
['PRT' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' nan 'ROU' 'NOR' 'OMN' 'ARG' 'POL'
 'DEU' 'BEL' 'CHE' 'CN' 'GRC' 'ITA' 'NLD' 'DNK' 'RUS' 'SWE' 'AUS' 'EST'
 'CZE' 'BRA' 'FIN' 'MOZ' 'BWA' 'LUX' 'SVN' 'ALB' 'IND' 'CHN' 'MEX' 'MAR'
 'UKR' 'SMR' 'LVA' 'PRI' 'SRB' 'CHL' 'AUT' 'BLR' 'LTU' 'TUR' 'ZAF' 'AGO'
 'ISR' 'CYM' 'ZMB' 'CPV' 'ZWE' 'DZA' 'KOR' 'CRI' 'HUN' 'ARE' 'TUN' 'JAM'
 'HRV' 'HKG' 'IRN' 'GEO' 'AND' 'GIB' 'URY' 'JEY' 'CAF' 'CYP' 'COL' 'GGY'
 'KWT' 'NGA' 'MDV' 'VEN' 'SVK' 'FJI' 'KAZ' 'PAK' 'IDN' 'LBN' 'PHL' 'SEN'
 'SYC' 'AZE' 'BHR' 'NZL' 'THA' 'DOM' 'MKD' 'MYS' 'ARM' 'JPN' 'LKA' 'CUB'
 'CMR' 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SYR' 'SGP' 'BDI'
 'SAU' 'VNM' 'PLW' 'QAT' 'EGY' 'PER' 'MLT' 'MWI' 'ECU' 'MDG' 'ISL' 'UZB'
 'NPL' 'BHS' 'MAC' 'TGO' 'TWN' 'DJI' 'STP' 'KNA' 'ETH' 'IRQ' 'HND' 'RWA'
 'KHM' 'MCO' 'BGD' 'IMN' 'TJK' 'NIC' 'BEN' 'VGB' 'TZA' 'GAB' 'GHA' 'TMP'
 'GLP' 'KEN' 'LIE' 'GNB' 'MNE' 'UMI' 'MYT' 'FRO' 'MMR' 'PAN' 'BFA' 'LBY'
 'MLI' 'NAM' 'BOL' 'PRY' 'BRB' 'ABW' 'AIA' 'SLV' 'DMA' 'PYF' 'GUY' 'LCA'
 'ATA' 'GTM' 'ASM' 'MRT' 'NCL' 'KIR' 'SDN' 'ATF' 'SLE' 'LAO']
-----
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']
-----
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
-----
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
-----
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K']
-----
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
-----
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
-----
reservation_status
['Check-Out' 'Canceled' 'No-Show']
-----
name
['Ernest Barnes' 'Andrea Baker' 'Rebecca Parker' ... 'Wesley Aguilar'
 'Caroline Conley MD' 'Ariana Michael']
-----
email
['Ernest.Barnes31@outlook.com' 'Andrea_Baker94@aol.com'
 'Rebecca_Parker@comcast.net' ... 'Mary_Morales@hotmail.com'
 'MD_Caroline@comcast.net' 'Ariana_M@xfinity.com']
-----
phone-number
['669-792-1661' '858-637-6955' '652-885-2745' ... '395-518-4100'
 '531-528-1017' '422-804-6403']
-----
credit_card
['*****4322' '*****9157' '*****3734' ...

```

['*****9170' '*****6349' '*****7959']

```
In [70]: df.isnull().sum()
```

```
Out[70]: 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
month                                      0
dtype: int64
```

```
In [74]: df.drop(['company', 'agent', 'credit_card', 'phone-number'], axis=1, inplace=True)
df.columns
```

```
Out[74]: 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'],
              dtype='object')
```

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

```
Out[58]: Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
      'arrival_date_month', 'arrival_date_week_number',
      'arrival_date_day_of_month', 'stays_in_weekend_nights',
      'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
      'country', 'market_segment', 'distribution_channel',
      'is_repeated_guest', 'previous_cancellations',
      'previous_bookings_not_canceled', 'reserved_room_type',
      'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
      'company', 'days_in_waiting_list', 'customer_type', 'adr',
      'required_car_parking_spaces', 'total_of_special_requests',
      'reservation_status', 'reservation_status_date', 'name', 'email',
      'phone-number', 'credit_card'],
      dtype='object')
```

```
In [45]: df.dropna(inplace=True)
```

```
In [46]: df.isnull().sum()
```

```
Out[46]: 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
credit_card                              0
dtype: int64
```

```
In [50]: df.describe()
```

Out[50]:	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	s
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	
mean	0.371347	104.312018	2016.157657	27.166674	15.800802	
std	0.483167	106.903570	0.707462	13.589966	8.780321	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	

```
In [58]: df=df[df['adr']<5000]
```

```
In [117... ###Data analysis and Visualization
```

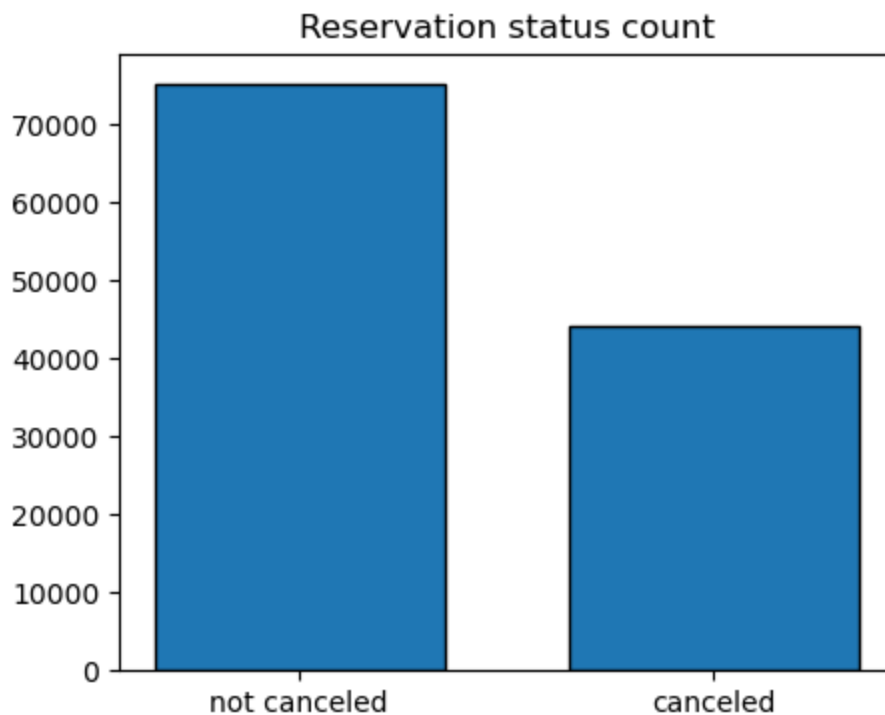
```
In [19]: canceled_perc=df['is_canceled'].value_counts(normalize=True)
canceled_perc
```

```
Out[19]: 0    0.629584
1    0.370416
Name: is_canceled, dtype: float64
```

```
In [17]: canceled_perc=df['is_canceled'].value_counts(normalize=True)

plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['not canceled','canceled'],df['is_canceled'].value_counts(), edgecolor='k' ,wid
```

```
Out[17]: <BarContainer object of 2 artists>
```



```
In [73]: plt.figure(figsize=(8,4))
ax1=sns.countplot(x= 'hotel',hue='is_canceled', data=df, palette='Blues')
legend_labels,_=ax1. get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status in different hostels')
```



```
plt.xlabel('Hotel')
plt.ylabel('Number of reservations')
plt.legend(['Not canceled', 'canceled'])
plt.show()
```



```
In [88]: #FILTER IN HOTEL COL TO RESORT HOTEL
resort_hotel=df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)#normalize is showing in percent
```

```
Out[88]: 0    0.722366
1    0.277634
Name: is_canceled, dtype: float64
```

```
In [89]: #FILTER IN HOTEL COL TO CITY HOTEL

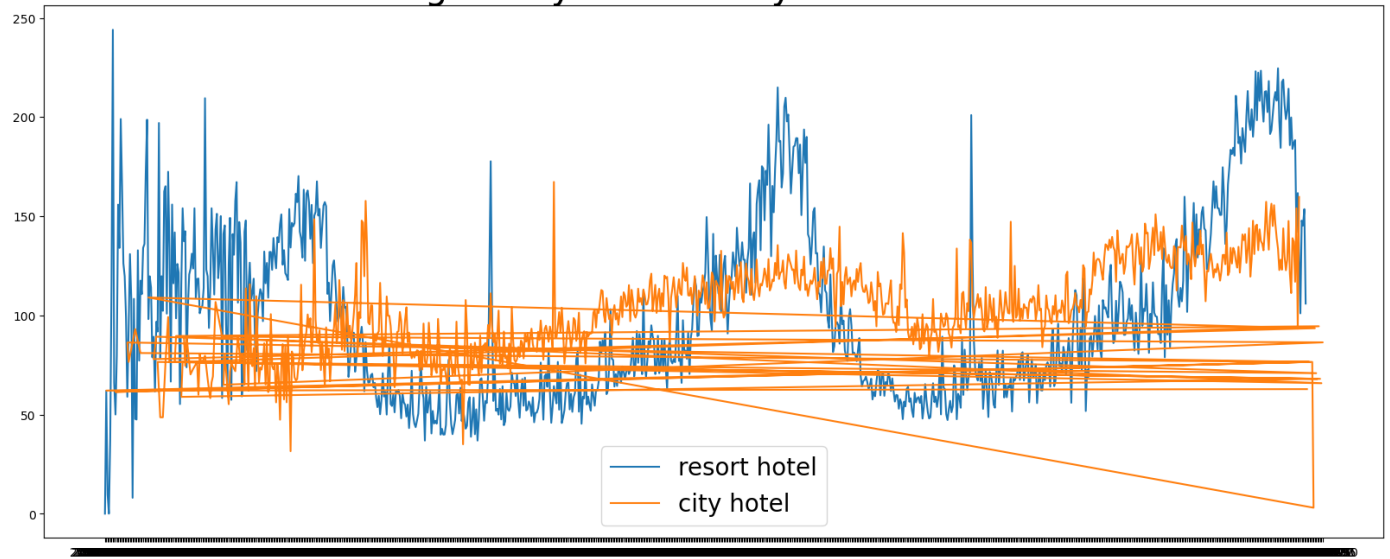
city_hotel= df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)
```

```
Out[89]: 0    0.58273
1    0.41727
Name: is_canceled, dtype: float64
```

```
In [122... resort_hotel=resort_hotel.groupby('c')[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

```
In [123... plt.figure(figsize=(20,8))
plt.title('Average Daily Rate in City and Resort Hotel', fontsize=30)
plt.plot(resort_hotel.index,resort_hotel['adr'],label='resort hotel')
plt.plot(city_hotel.index,city_hotel['adr'],label='city hotel')
plt.legend(fontsize=20)
plt.show()
```

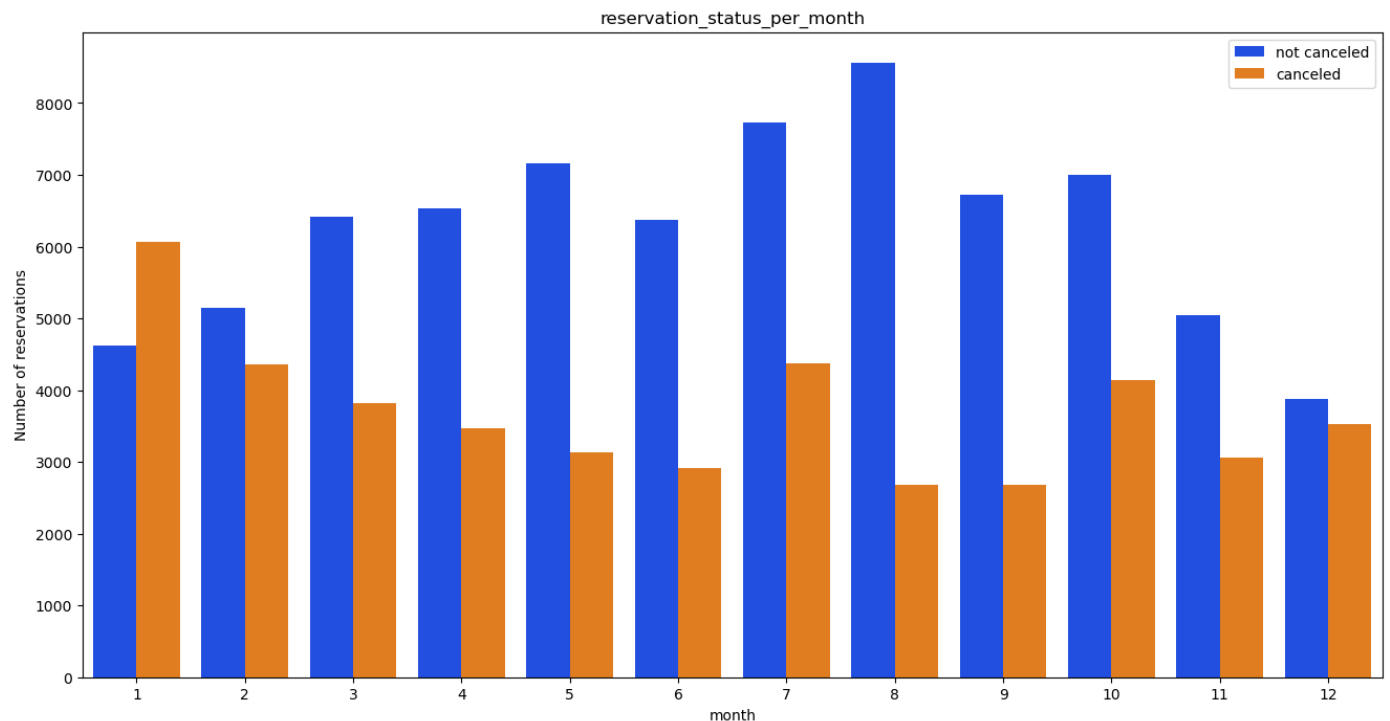
Average Daily Rate in City and Resort Hotel



```
In [46]: #df['month'] = df[''].dt.month

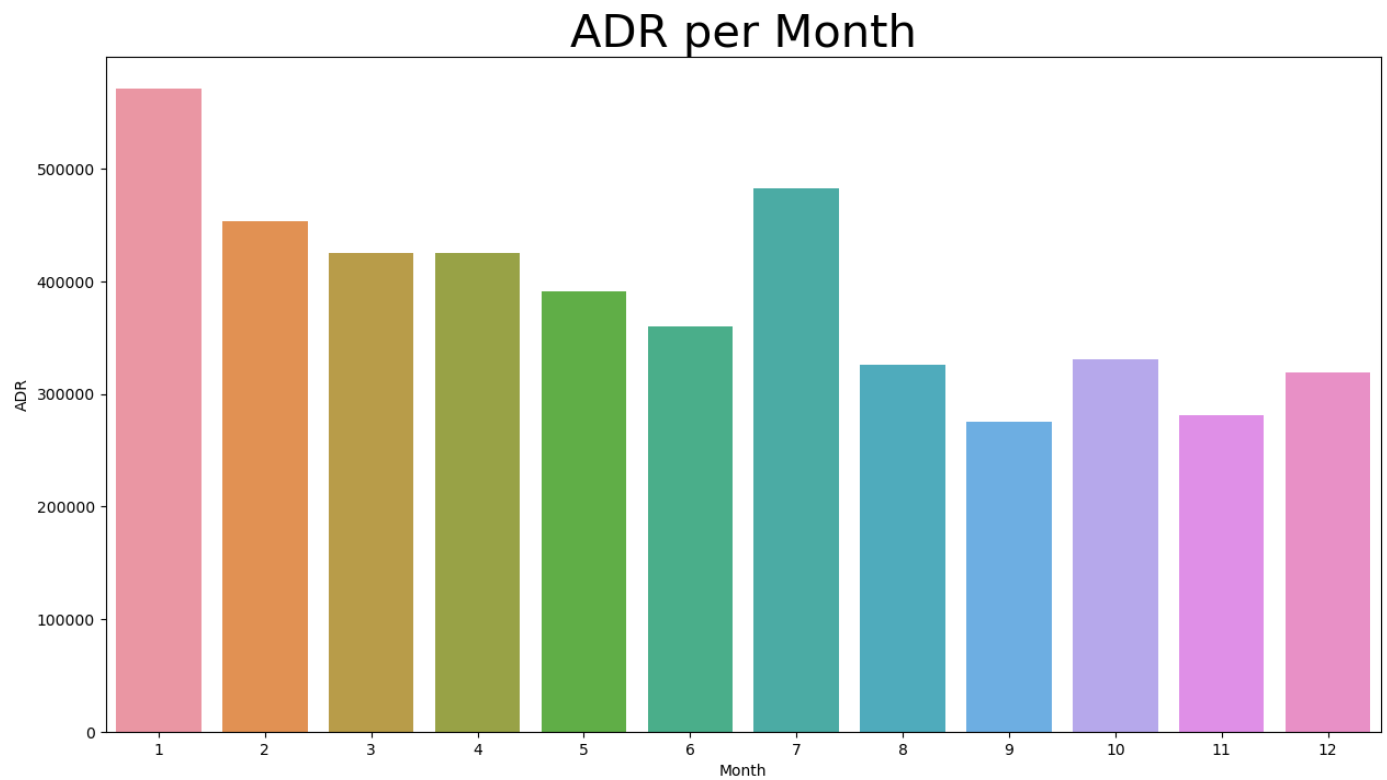
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

```
In [68]: df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
ax1= sns.countplot(x='month', hue='is_canceled', data=df, palette='bright')
#legend_labels,_= ax1.get_legend_handles_labels()
#ax1.legend(bbox_to_anchor=(1,1))
plt.title('reservation_status_per_month')
plt.xlabel('month')
plt.ylabel('Number of reservations')
plt.legend(['not canceled', 'canceled'])
plt.show()
```



```
In [86]: plt.figure(figsize=(15,8))
plt.title('ADR per Month', fontsize=30)
sns.barplot(x='month', y='adr', data=df[df['is_canceled']== 1].groupby('month')[['adr']])
plt.xlabel('Month')
plt.ylabel('ADR')
```

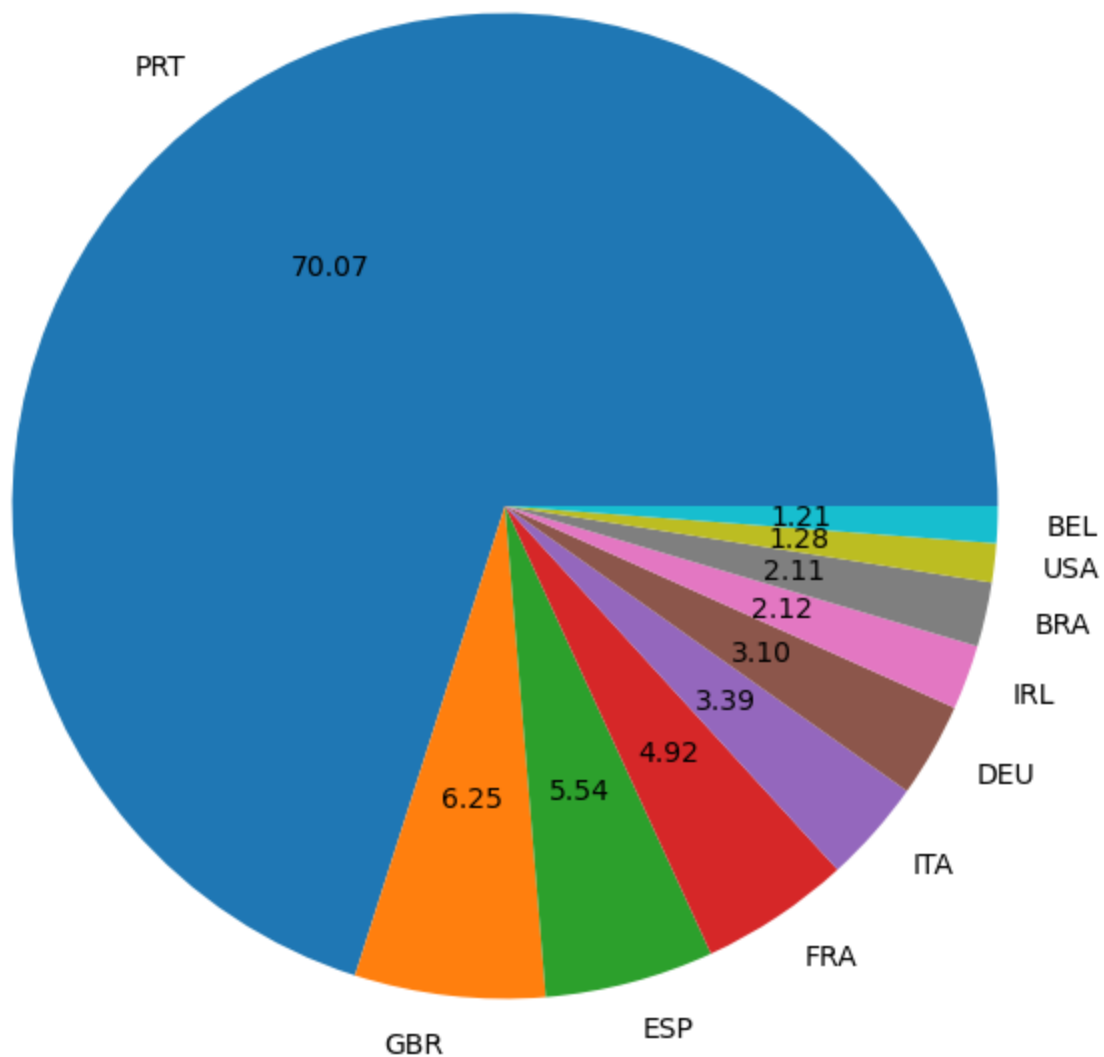
```
plt.show()
```



In [102...

```
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()
```

Top 10 Countries with reservation canceled



```
In [103]: df['market_segment'].value_counts()
```

```
Out[103]: Online TA      56477
Offline TA/T0    24219
Groups          19811
Direct          12606
Corporate        5295
Complementary    743
Aviation         237
Undefined         2
Name: market_segment, dtype: int64
```

```
In [105]: df['market_segment'].value_counts(normalize=True)
```

```
Out[105]: Online TA      0.473046
Offline TA/T0    0.202856
Groups          0.165935
Direct          0.105587
Corporate        0.044350
Complementary    0.006223
Aviation         0.001985
Undefined        0.000017
Name: market_segment, dtype: float64
```

```

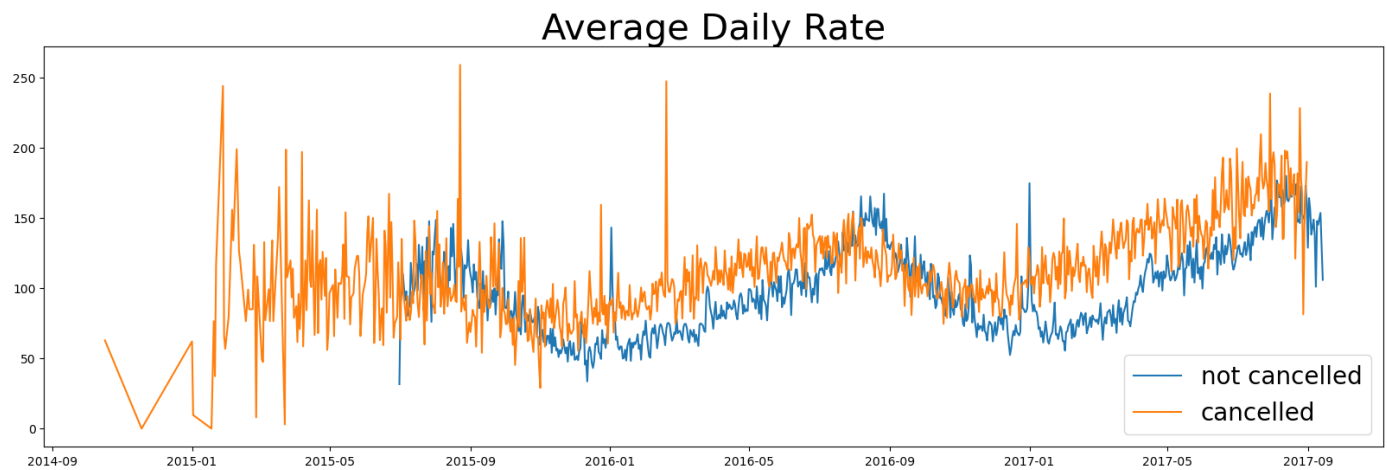
In [133... cancelled_df_adr = cancelled_data.groupby( 'reservation_status_date') [['adr']].mean()
cancelled_df_adr.reset_index(inplace = True)
cancelled_df_adr.sort_values('reservation_status_date', inplace = True)

not_cancelled_data = df[df['is_canceled'] == 0]
not_cancelled_df_adr = not_cancelled_data.groupby( 'reservation_status_date')[['adr']].m
not_cancelled_df_adr.reset_index(inplace = True)

not_cancelled_df_adr.sort_values('reservation_status_date', inplace = True)
plt.figure(figsize = (20,6))
plt.title('Average Daily Rate', fontsize= 30)
plt.plot(not_cancelled_df_adr[ 'reservation_status_date'], not_cancelled_df_adr['adr'],
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label = '
plt.legend(fontsize = 20)

```

Out[133]: <matplotlib.legend.Legend at 0x1d03f39da20>



```

In [134... cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr[ 'reservation_status_date']> '2016
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr[ 'reservation_status_d

```

In []: