DATA ANALYSIS ON HOTEL BOOKING

Author-lavanya GT

IMPORTING LIBRARIES

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

LOADING DATA SET

```
df = pd.read_csv('/content/drive/MyDrive/hotel_booking.csv')
```

EDA & DATA CLEANING

```
df.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_dat
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	

5 rows × 35 columns

```
df.tail()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
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

5 rows × 36 columns

```
df.shape
```

```
(119390, 36)
```

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

```
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
```

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

```
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
```

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

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

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

```
hotel
<bound method Series.unique of 0         Resort Hotel
1         Resort Hotel
2         Resort Hotel
3         Resort Hotel
4         Resort Hotel
              ...
119385      City Hotel
119386      City Hotel
119387      City Hotel
119388      City Hotel
119389      City Hotel
Name: hotel, Length: 119390, dtype: object>
__________________________________________________
arrival_date_month
<bound method Series.unique of 0           July
1           July
2           July
3           July
4           July
           ...
119385    August
119386    August
119387    August
119388    August
119389    August
Name: arrival_date_month, Length: 119390, dtype: object>
__________________________________________________
meal
<bound method Series.unique of 0         BB
1         BB
```

```
2         BB
3         BB
4         BB
          ..
119385    BB
119386    BB
119387    BB
119388    BB
119389    HB
Name: meal, Length: 119390, dtype: object>
_________________________________________________
country
<bound method Series.unique of 0         PRT
1         PRT
2         GBR
3         GBR
4         GBR
         ...
119385    BEL
119386    FRA
119387    DEU
119388    GBR
119389    DEU
Name: country, Length: 119390, dtype: object>
_________________________________________________
market_segment
<bound method Series.unique of 0               Direct
```

```
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                            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
```

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

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

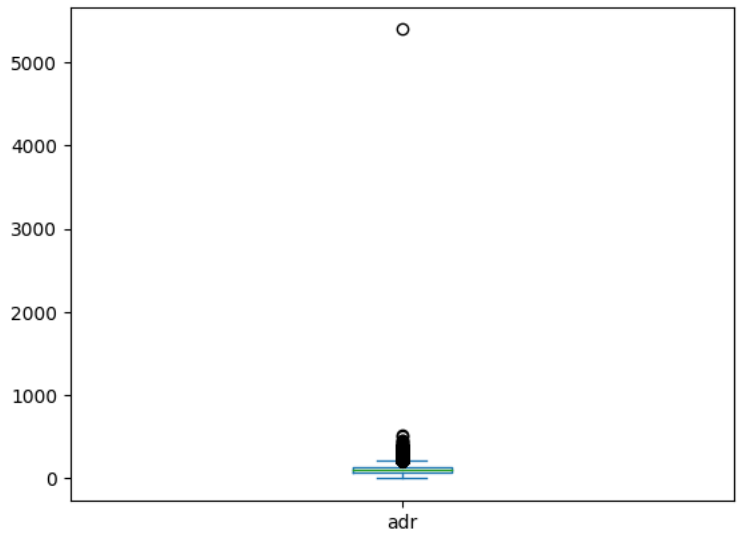
```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        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
```

```
df.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_m
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.00(
mean	0.371352	104.311435	2016.157656	27.166555	15.80(
std	0.483168	106.903309	0.707459	13.589971	8.78(
min	0.000000	0.000000	2015.000000	1.000000	1.00(
25%	0.000000	18.000000	2016.000000	16.000000	8.00(
50%	0.000000	69.000000	2016.000000	28.000000	16.00(
75%	1.000000	161.000000	2017.000000	38.000000	23.00(
max	1.000000	737.000000	2017.000000	53.000000	31.00(

```
df['adr'].plot(kind = 'box')
plt.savefig('my_plot.png')
```




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

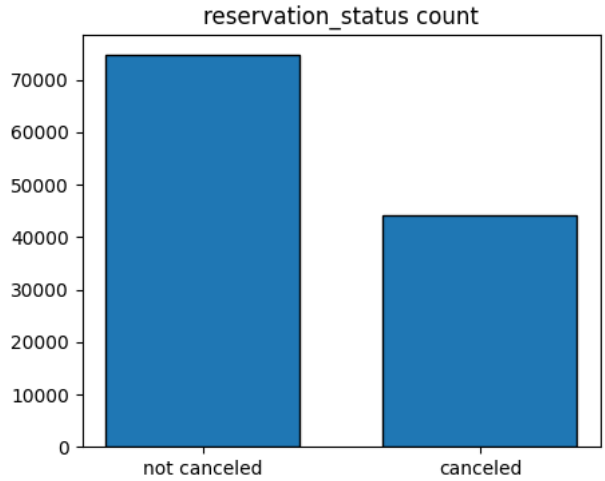
```
df.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_m
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.00(
mean	0.371347	104.312018	2016.157657	27.166674	15.80(
std	0.483167	106.903570	0.707462	13.589966	8.78(
min	0.000000	0.000000	2015.000000	1.000000	1.00(
25%	0.000000	18.000000	2016.000000	16.000000	8.00(

DATA ANALYSIS AND VISUALISATION

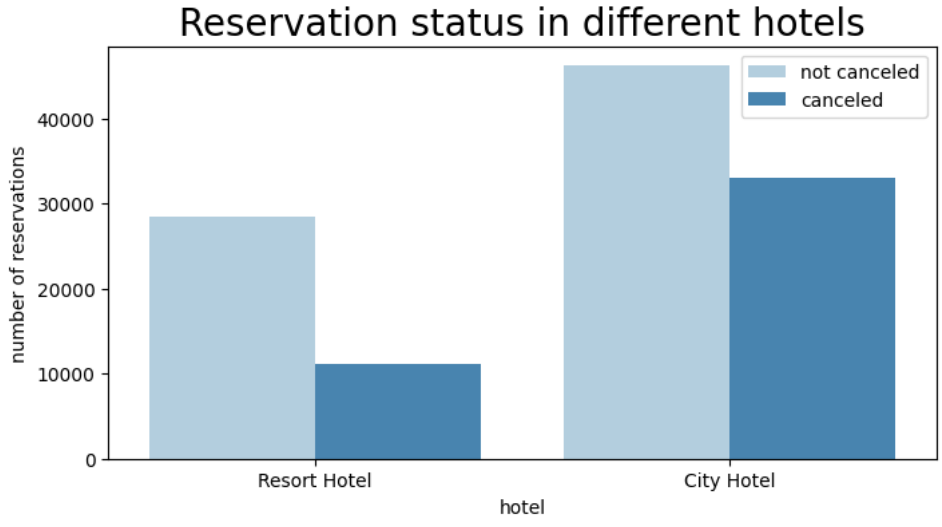
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)
plt.figure(figsize = (5,4))
plt.title('reservation_status count')
plt.bar(['not canceled','canceled'],df['is_canceled'].value_counts(),edgecolor = 'k',width = 0.7)
plt.show
plt.savefig('my_plot1.png')
```

```
0    0.628648
1    0.371352
Name: is_canceled, dtype: float64
```




```
plt.figure(figsize = (8,4))
ax1= sns.countplot(x = 'hotel', hue = 'is_canceled',data = df,palette = 'Blues')
legend_labels,_ =ax1.get_legend_handles_labels()
plt.legend(loc='upper right', bbox_to_anchor=(1.2, 1))

plt.title('Reservation status in different hotels',size = 20)
plt.xlabel('hotel')
plt.ylabel('number of reservations')
plt.legend(['not canceled','canceled'])
plt.show
plt.savefig('my_plot2.png')
```



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

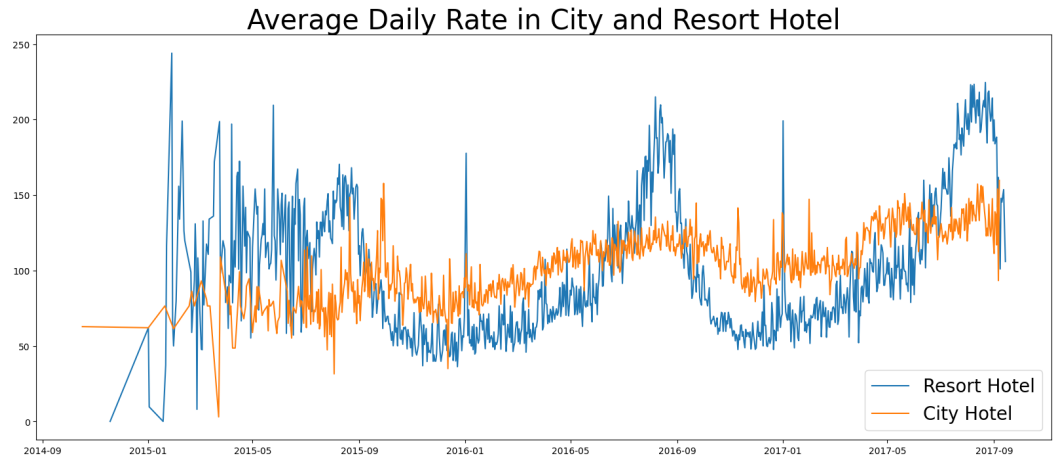
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

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

```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

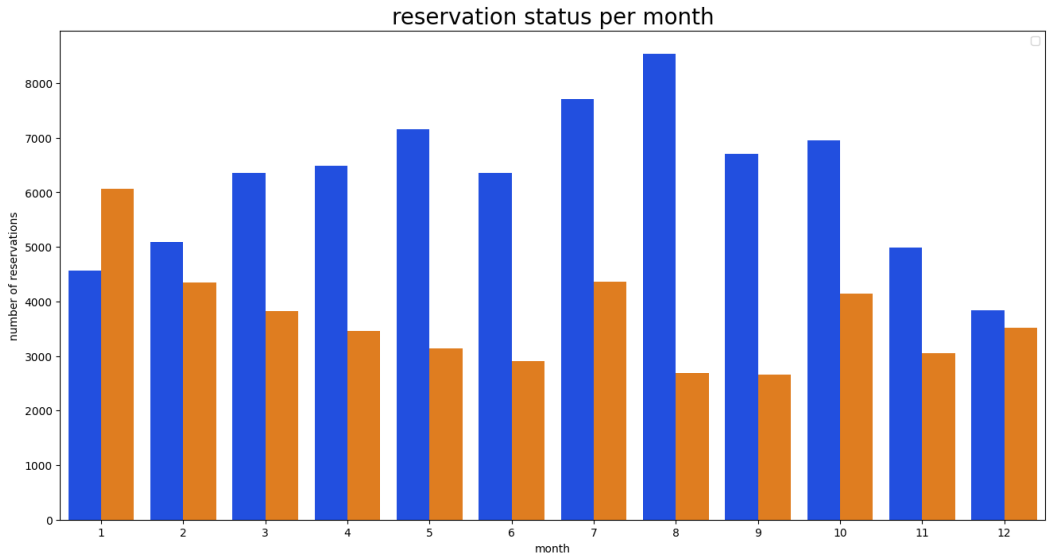
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

```
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()
plt.savefig('my_plot3.png')
```



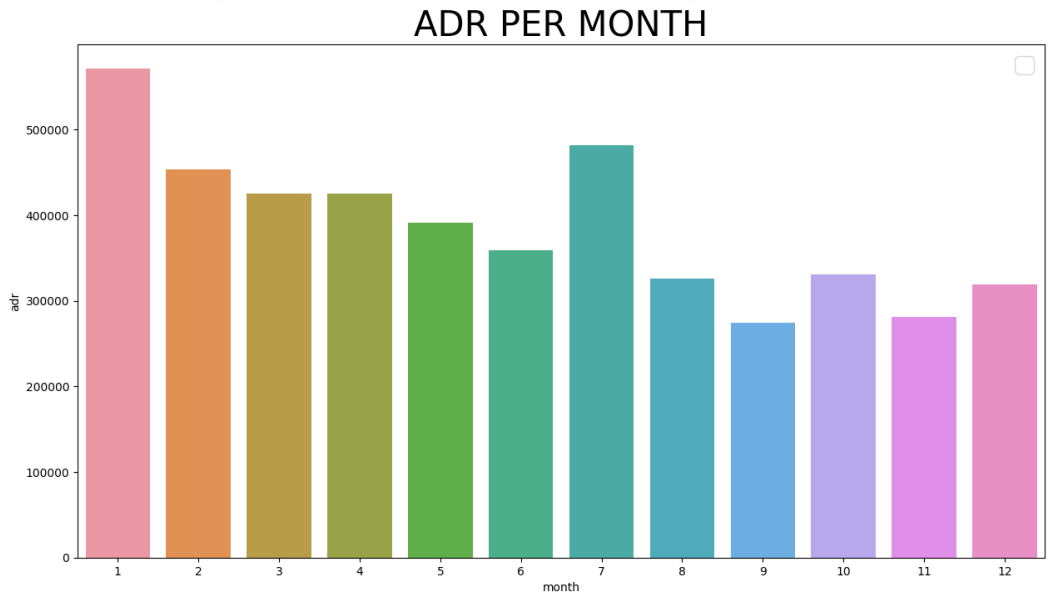

```
<Figure size 640x480 with 0 Axes>
```

```
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',size = 20)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.legend('[not canceled','canceled]')
plt.show
plt.savefig('my_plot3.png')
```



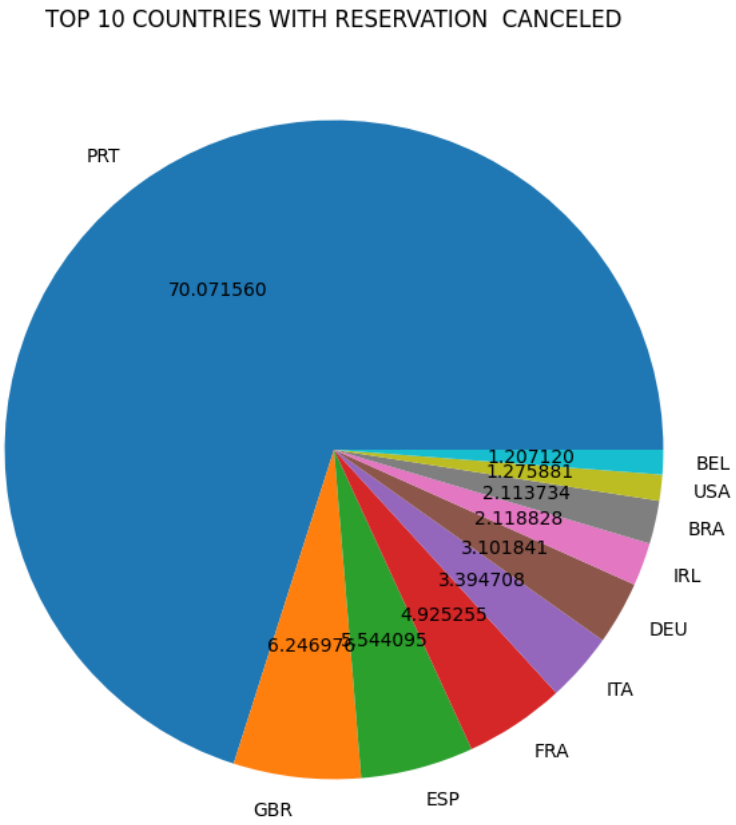

```
plt.figure(figsize = (15,8))
plt.title('ADR PER MONTH',fontsize = 30)
sns.barplot(x ='month',y ='adr',data = df[df['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_index())
plt.legend(fontsize = 20)
plt.show
plt.savefig('my_plot4.png')
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label
```




```
canceled_data = df[df['is_canceled']==1]
```

```
top_10_country = canceled_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
plt.savefig('my_plot5.png')
```




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

```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

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

```
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
```

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

```
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
```

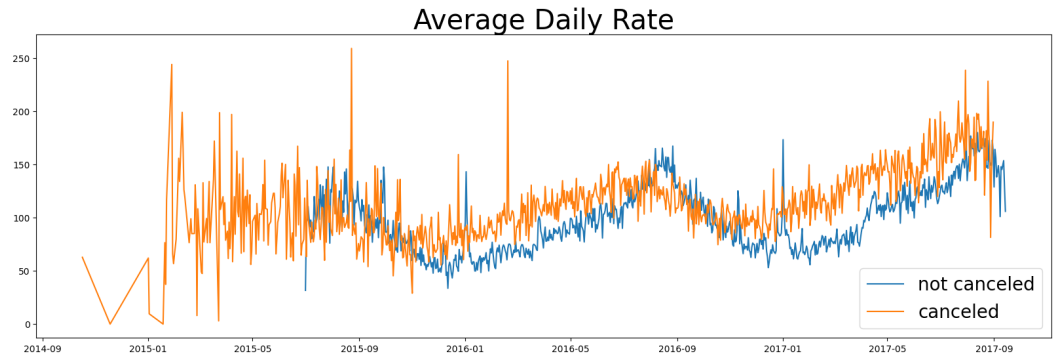
```
canceled_df_adr = canceled_data.groupby('reservation_status_date')[['adr']].mean()
canceled_df_adr.reset_index(inplace = True)
canceled_df_adr.sort_values('reservation_status_date',inplace = True)

not_canceled_data =df[df['is_canceled'] == 0]
not_canceled_df_adr = not_canceled_data.groupby('reservation_status_date')[['adr']].mean()
```

```
not_canceled_df_adr.reset_index(inplace = True)
not_canceled_df_adr.sort_values('reservation_status_date',inplace = True)


plt.figure(figsize = (20,6))
plt.title('Average Daily Rate',fontsize = 30)
plt.plot(not_canceled_df_adr['reservation_status_date'],not_canceled_df_adr['adr'],label='not canceled')
plt.plot(canceled_df_adr['reservation_status_date'],canceled_df_adr['adr'],label='canceled')

plt.legend(fontsize=20)
plt.savefig('my_plot6.png')
```




```
canceled_df_adr = canceled_df_adr[(canceled_df_adr['reservation_status_date']>'2016') & (canceled_df_adr['reservation_status_date']<'201
not_canceled_df_adr = not_canceled_df_adr[(not_canceled_df_adr['reservation_status_date']>'2016') & (not_canceled_df_adr['reservation_st
```

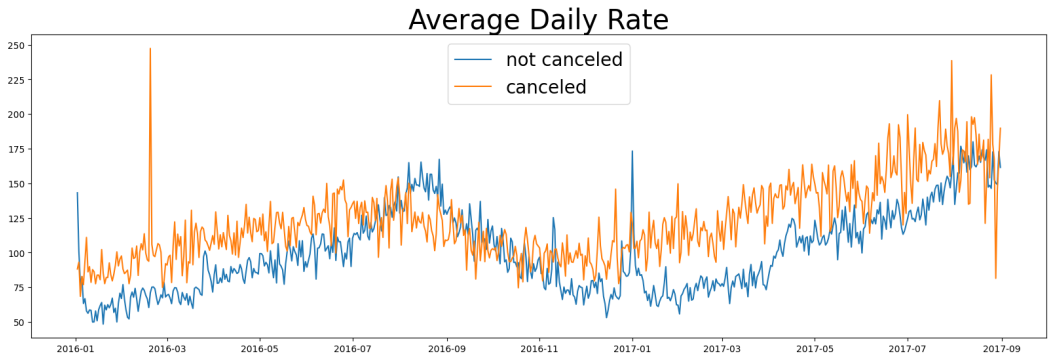
```
print(canceled_df_adr)
print(not_canceled_df_adr)
```

```
    reservation_status_date         adr
295              2016-01-02   88.147143
296              2016-01-03   93.053810
297              2016-01-04   68.428750
298              2016-01-05   82.809895
299              2016-01-06   76.845820
..                      ...         ...
897              2017-08-26  178.200000
898              2017-08-27  167.300000
899              2017-08-28   81.416667
900              2017-08-29  144.253333
901              2017-08-31  189.750000

[607 rows x 2 columns]
    reservation_status_date         adr
185              2016-01-02  143.154565
186              2016-01-03  100.776163
187              2016-01-04   76.913125
188              2016-01-05   82.525818
189              2016-01-06   63.240952
..                      ...         ...
788              2017-08-27  152.494744
789              2017-08-28  150.247197
790              2017-08-29  149.113494
791              2017-08-30  172.943662
792              2017-08-31  161.504861

[608 rows x 2 columns]
```

```
plt.figure(figsize = (20,6))
plt.title('Average Daily Rate',fontsize = 30)
plt.plot(not_canceled_df_adr['reservation_status_date'],not_canceled_df_adr['adr'],label='not canceled')
plt.plot(canceled_df_adr['reservation_status_date'],canceled_df_adr['adr'],label='canceled')
plt.legend(fontsize = 20)
plt.savefig('my_plot7.png')
```



✓ 0s completed at 6:47 PM