```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
df = pd.read_csv('hotel_booking.csv')
print(df)
```

```
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]
```

```
df.head()           #staring 5 row
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_dat
0	Resort Hotel	0	342	2015	July	

```
df.tail()        #last five row
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival
119385	City Hotel	0	23	2017	August	
119386	City Hotel	0	102	2017	August	
119387	City Hotel	0	34	2017	August	
119388	City Hotel	0	109	2017	August	
119389	City Hotel	0	205	2017	August	

5 rows × 36 columns

```
df.shape        #total row and columns
```

```
(119390, 36)
```

```
df.columns           #checking column
```

```
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.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arri
count	119390.000000	119390.000000	119390.000000	119390.000000	
mean	0.370416	104.011416	2016.156554	27.165173	
std	0.482918	106.863097	0.707476	13.605138	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	160.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

```
df.info         #checking datatype of the columns
```

```
<bound method DataFrame.info of                 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
```

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

	hotel	arrival_date_month	meal	country	market_segment	distribution_cha
count	119390	119390	119390	118902	119390	11
unique	2	12	5	177	8	
top	City Hotel	August	BB	PRT	Online TA	T.
freq	79330	13877	92310	48590	56477	9

```
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']
--------------------------------------------------
reservation_status_date
['2015-07-01' '2015-07-02' '2015-07-03' '2015-05-06' '2015-04-22'
 '2015-06-23' '2015-07-05' '2015-07-06' '2015-07-07' '2015-07-08'
 '2015-05-11' '2015-07-15' '2015-07-16' '2015-05-29' '2015-05-19'
 '2015-06-19' '2015-05-23' '2015-05-18' '2015-07-09' '2015-06-02'
 '2015-07-13' '2015-07-04' '2015-06-29' '2015-06-16' '2015-06-18'
 '2015-06-12' '2015-06-09' '2015-05-26' '2015-07-11' '2015-07-12'
 '2015-07-17' '2015-04-15' '2015-05-13' '2015-07-10' '2015-05-20'
 '2015-05-12' '2015-07-14' '2015-06-17' '2015-05-01' '2015-03-30'
```

```
df.isnull().sum()  #return total missing values with name
```

```
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()
```

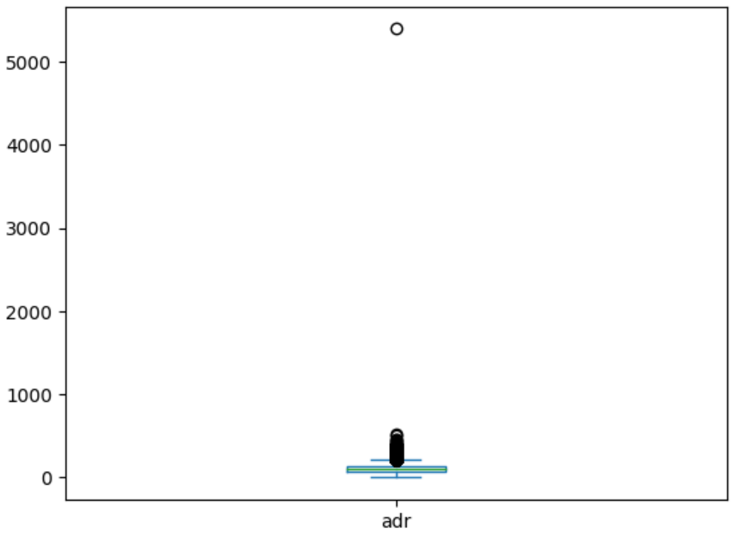
```
hotel                             0
is_canceled                       0
lead_time                         0
arrival_date_year                 0
arrival_date_month                0
arrival_date_week_number          0
arrival_date_day_of_month         0
stays_in_weekend_nights           0
stays_in_week_nights              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	arri
count	118898.000000	118898.000000	118898.000000	118898.000000	
mean	0.371352	104.311435	2016.157656	27.166555	
std	0.483168	106.903309	0.707459	13.589971	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	161.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

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

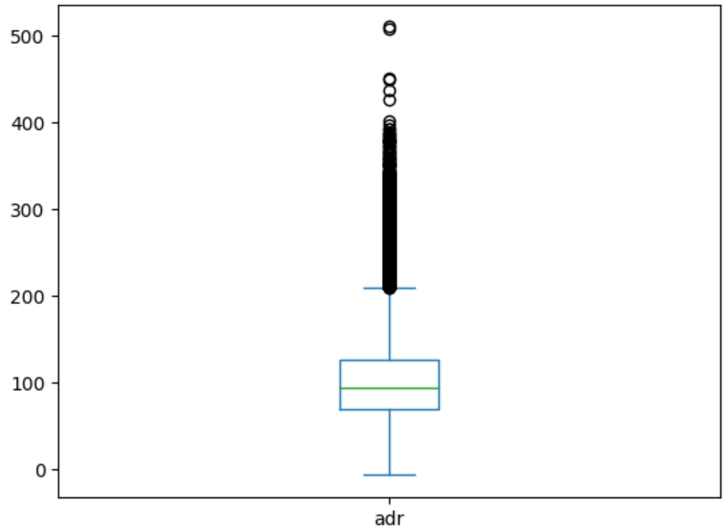
```
<Axes: >
```



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

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

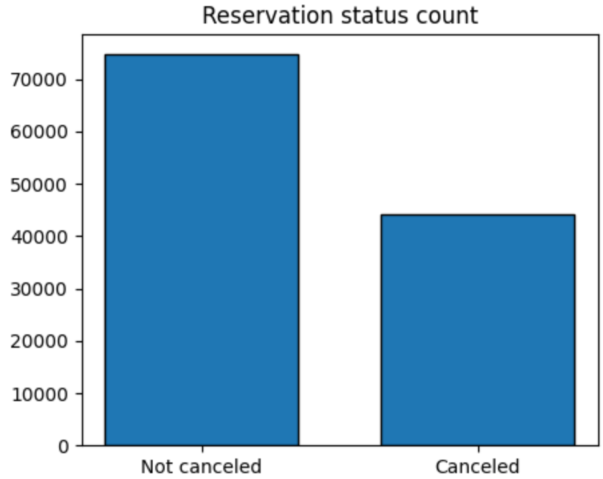
```
<Axes: >
```



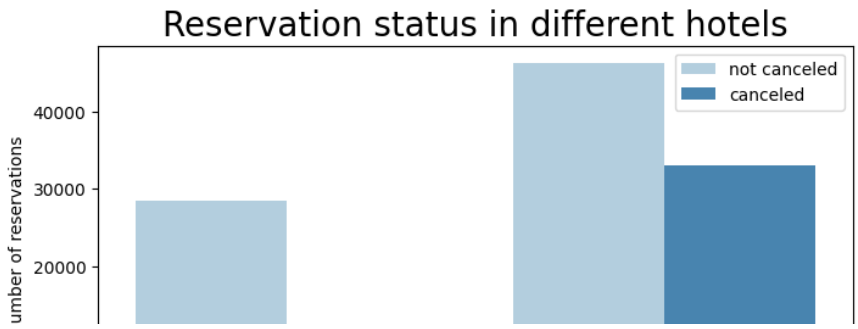
DATA ANALYSIS AND VISUALIZATIONS

```
import matplotlib.pyplot as plt
cancel_perc = df['is_canceled'].value_counts(normalize=True)
print(cancel_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()
```

```
0    0.628653
1    0.371347
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()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status in different hotels',size=20)
plt.xlabel('hotel')
plt.ylabel('number of reservations')
plt.legend(['not canceled','canceled'])
plt.show()
```

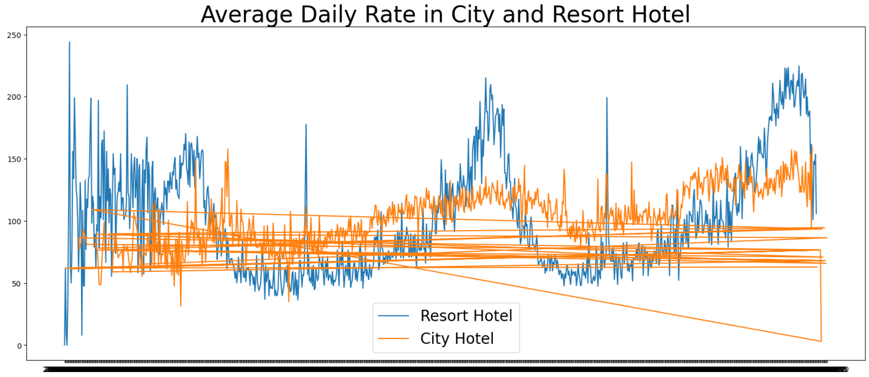


```
resort_hotel=df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)

     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
```

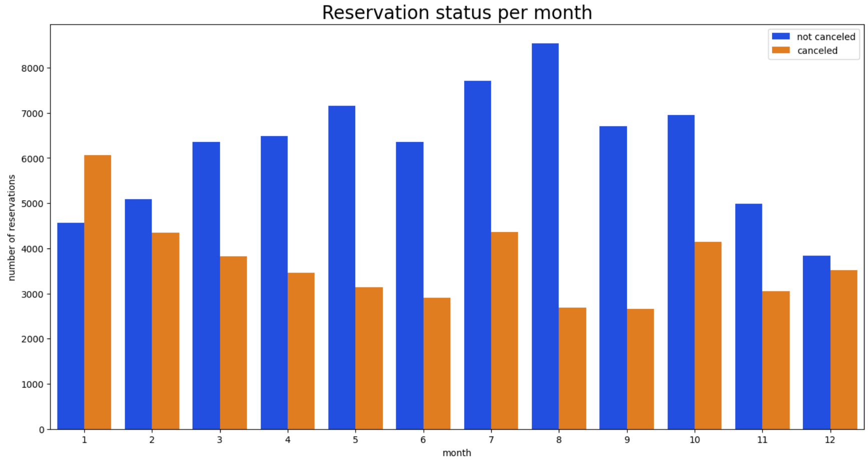
```
resort_hotel=resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

```
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()
```

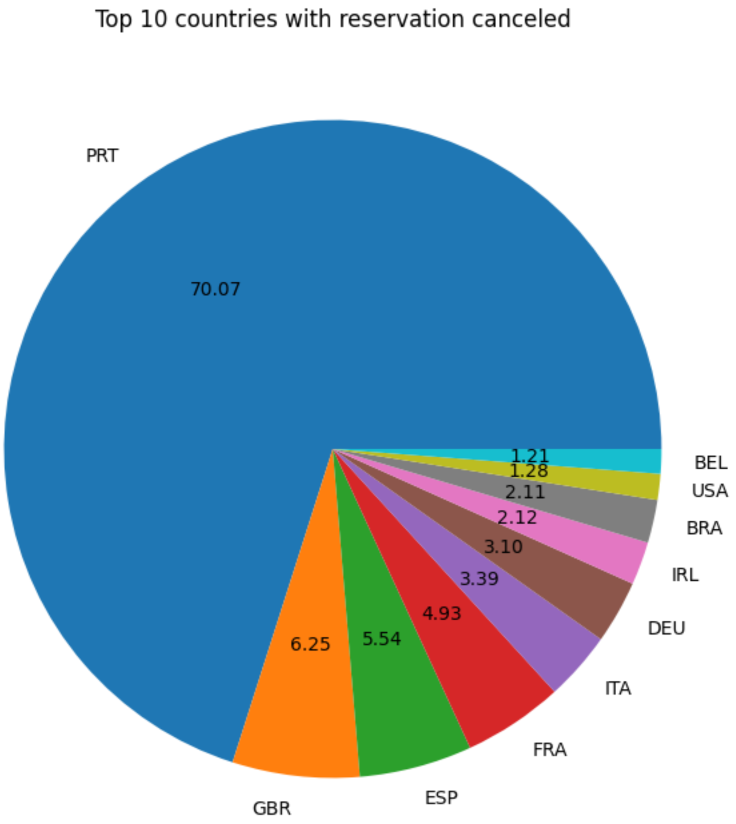


```
df['reservation_status_date']=pd.to_datetime(df['reservation_status_date'])
df['month'] = df['reservation_status_date'].dt.month
```

```
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()
```



```
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()
```



```
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)
```

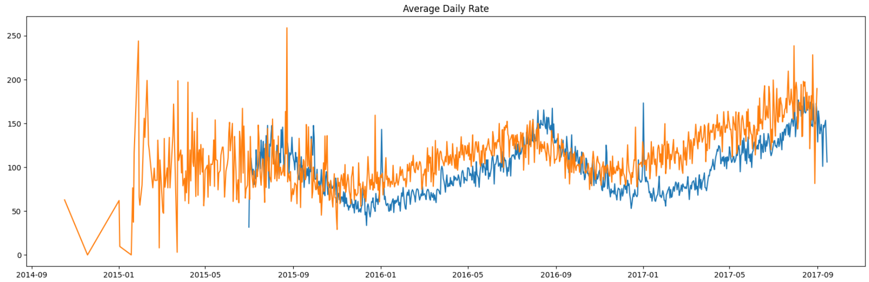
```
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: market_segment, dtype: float64
```

```
cancelled_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
```

```
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']].mean()
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')
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
```

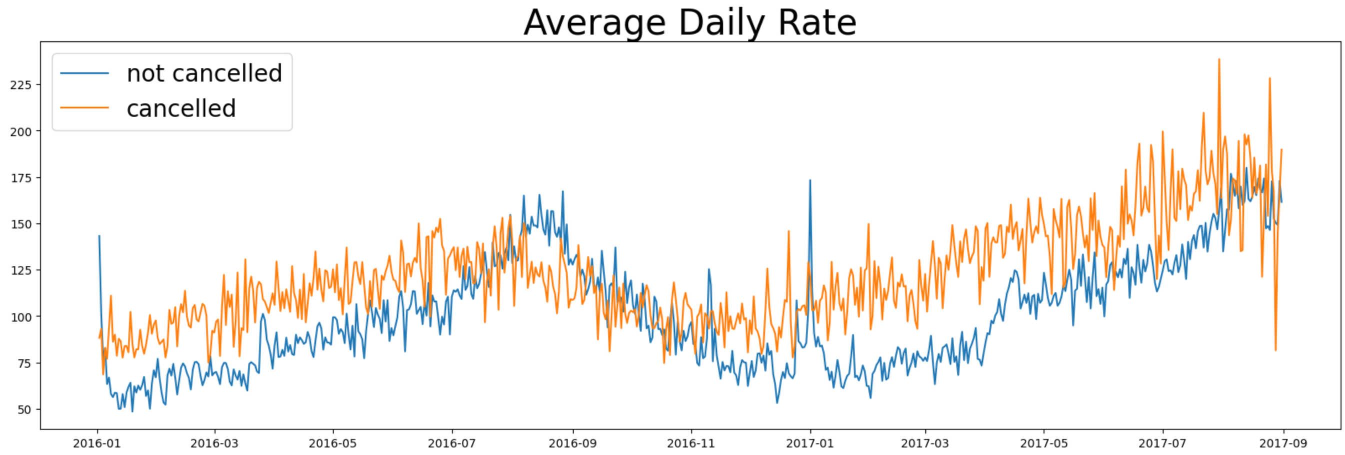
```
<function matplotlib.pyplot.legend(*args, **kwargs)>
```



```
cancelled_df_adr=cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016')&(cancelled_df_adr['reservation_status_date']<'2017
not_cancelled_df_adr=not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016')&(not_cancelled_df_adr['reservation_sta
```

```
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'],label='not cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label='cancelled')
plt.legend(fontsize=20)
```

```
<matplotlib.legend.Legend at 0x79ad18d7f760>
```


Average Daily Rate
not cancelled
cancelled
225
200
175
150
125
100
75
50
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09