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

```
d=pd.read_csv("hotel.csv")
```

```
d.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_d
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 × 32 columns

```
d.shape
```

```
(119390, 32)
```

```
d.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')
```

```
d.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
```

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

```
d.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
```

```
for col in d.describe(include='object').columns:
  print(col)
  print(d[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']
--------------------------------------------------
```

```
d.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
dtype: int64
```

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

```
d.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000
mean	0.371352	104.311435	2016.157656	27.166555	15.800880
std	0.483168	106.903309	0.707459	13.589971	8.780324
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

```
#removing the adr which is >5000 and saving only<5000 to handle more effectiently
#because if >5000 is present in data then the client did n't able to understand the graphs
#
d=d[d['adr']<5000] #adr= average daily rate
```

```
d.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month
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

```
c=d['is_canceled'].value_counts(normalize=True)  #c=cancelled_percentage
print(c)

plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['Not cancelled','Cancelled'],d['is_canceled'].value_counts(),edgecolor='k',width=0.7)
plt.show()
```

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

```
#plotting the Reservation status in different hotels
plt.figure(figsize=(8,4))
ax1=sns.countplot(x='hotel', hue='is_canceled',data=d,palette='Blues')
legen_=ax1.get_legend_handles_labels()

plt.title('Reservation status in different hotels')
plt.xlabel('Hotel')
plt.ylabel('No.of reservations')
plt.legend('Not canceled','canceled')
plt.show()
```


```
#finding the resort hotel cancellation %
resort=d[d['hotel']=='Resort Hotel']
resort['is_canceled'].value_counts(normalize=True)
```

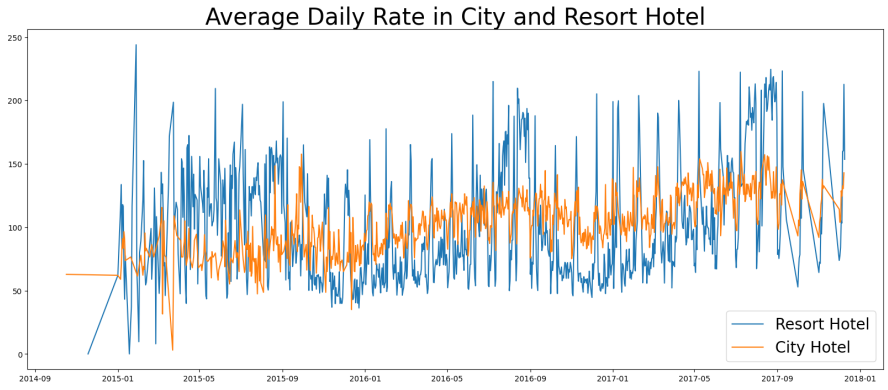
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

```
# in hotel section there are resort and city hotel are present
#finding the city hotel cancellation %
city=d[d['hotel']=='City Hotel']
city['is_canceled'].value_counts(normalize=True)
```

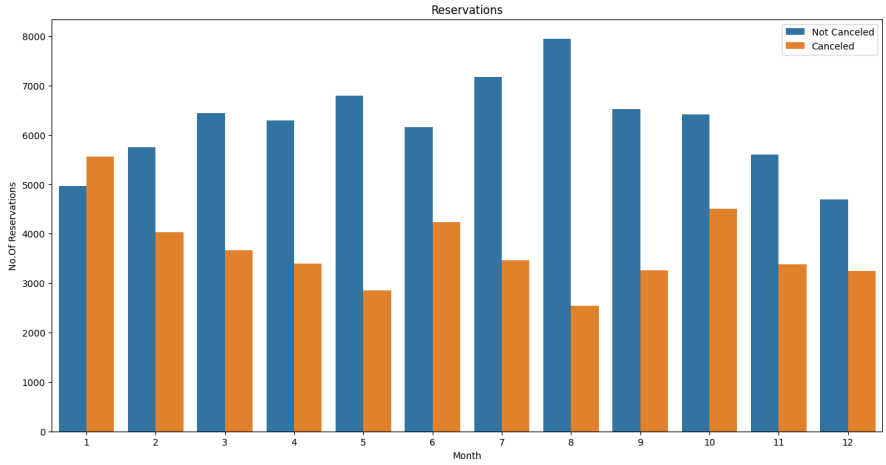
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

```
city=city.groupby('reservation_status_date')[['adr']].mean() # grouping the adr and finding mean adr to understand avg adr
resort=resort.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.index,resort['adr'],label='Resort Hotel')
plt.plot(city.index,city['adr'],label='City Hotel')
plt.legend(fontsize=20)
plt.show()
```

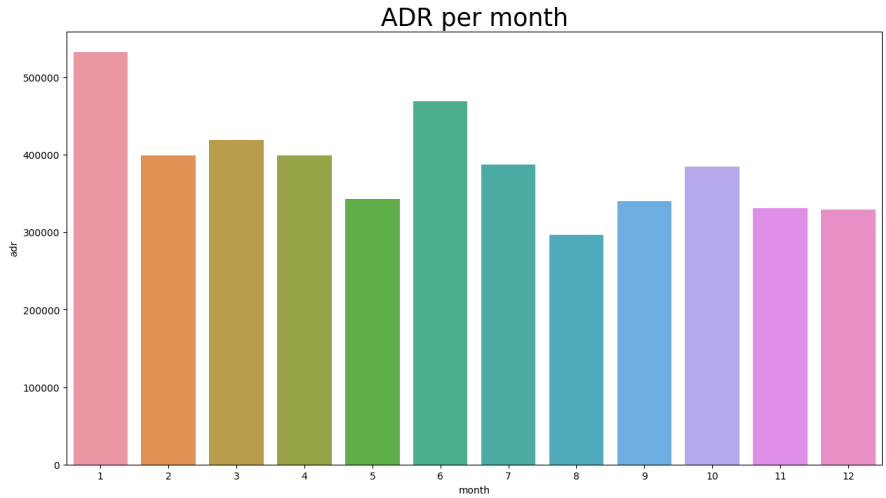


```
d['month']=d['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
ax1=sns.countplot(x='month',hue='is_canceled',data=d )
plt.title("Reservations")
plt.xlabel('Month')
plt.ylabel('No.Of Reservations')
plt.legend(['Not Canceled','Canceled'])
plt.show()
```




```
plt.figure(figsize=(15,8))
plt.title('ADR per month', fontsize=25)
sns.barplot(x='month', y='adr', data=d[d['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_index()) #reset index indicates the modified or altered data is replace
```

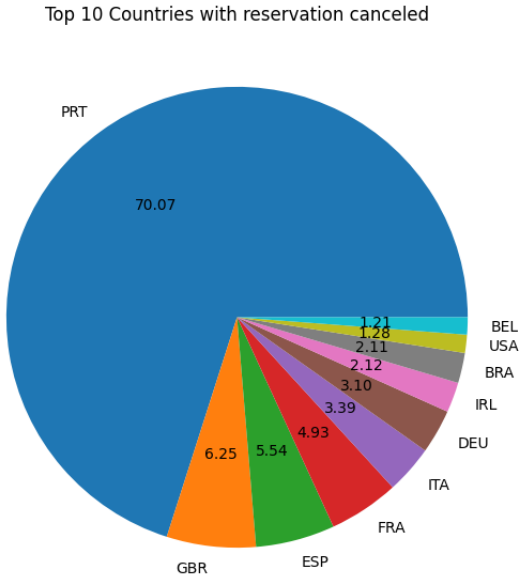
```
plt.show()
```




```
'''from the above two graphs 1.reservations and 2.adr/ month:concluded that if the adr is higher then the cancellation rates is higher
. Check with 1st month in adr is far high so, the cancellations rates are also high in 1st month of reservation graph'''
```

```
'from the above two graphs 1.reservations and 2.adr/ month:concluded that if the adr is higher then the can
cellation rates is higher\n. Check with 1st month in adr is far high so, the cancellations rates are also h
igh in 1st month of reservation graph'
```

```
canc=d[d['is_canceled']==1] #canc=cancellation data
top10=canc['country'].value_counts()[:10] #:10 indicates top 10 countries
plt.figure(figsize=(7,7))
plt.title('Top 10 Countries with reservation canceled')
plt.pie(top10,autopct='%.2f',labels=top10.index)
plt.show()
```




```
d['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
```

```
canc['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
```

```
'''by this attempt we can see that most reservations is coming from online and most
ancellation is happening from online
'''
```

```
'by this attempt we can see that most reservations is coming from online and most \nancellation is happening from online\n'
```

```
cancelled_d_adr=canc.groupby('reservation_status_date')[['adr']].mean() # getting the data from cancellation data that incliudes average daily rate in reservation_status_d
cancelled_d_adr.reset_index(inplace=True)  #reset index indicates the modified or altered data is replaced by default values after exceuting the pervious operations
cancelled_d_adr.sort_values('reservation_status_date',inplace=True)

not_cancelled=d[d['is_canceled']==0]
not_cancelled_adr=not_cancelled.groupby('reservation_status_date')[['adr']].mean()
not_cancelled_adr.reset_index(inplace=True)
not_cancelled_adr.sort_values('reservation_status_date',inplace=True)

plt.figure(figsize=(20,7))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_adr['reservation_status_date'],not_cancelled_adr['adr'],label='not cancelled')
plt.plot(cancelled_d_adr['reservation_status_date'],cancelled_d_adr['adr'],label='cancelled')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x7836b9b38100>
```


```
cancelled_d_adr=cancelled_d_adr[(cancelled_d_adr['reservation_status_date']>'2016')&(cancelled_d_adr['reservation_status_date']<'2017-09')]
not_cancelled_adr= not_cancelled_adr[(not_cancelled_adr['reservation_status_date']>'2016')& (not_cancelled_adr['reservation_status_date']<'2017-09')]
''' removing the data from which is previous of 2016 and after 2017-09 due to lack of contineous data in 2014, 2015 and 2018'''
```

```
' removing the data from which is previous of 2016 and after 2017-09 due to lack of contineous data in 2014, 2015 and 2018'
```



```
#plotting the same graph after removing <2016 and >2017-09
plt.figure(figsize=(20,6))
plt.title('Average Daily Rate',fontsize=20)
plt.plot(not_cancelled_adr['reservation_status_date'],not_cancelled_adr['adr'],label='not cancelled')
plt.plot(cancelled_d_adr['reservation_status_date'],cancelled_d_adr['adr'],label='cancelled')
plt.legend(fontsize=20)
plt.show()
```