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

```
#dataset
df=pd.read_csv('/content/hotel_booking.csv')
```

```
#EDA AND DATA CLEANING
df.head(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_c
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 × 36 columns

```
df.tail(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arri
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
```

```
(119390, 36)
```

```
df.dtypes
```

```
hotel                                 object
is_canceled                            int64
```

```
lead_time                           int64
arrival_date_year                   int64
arrival_date_month                 object
arrival_date_week_number            int64
arrival_date_day_of_month           int64
stays_in_weekend_nights             int64
stays_in_week_nights                int64
adults                              int64
children                          float64
babies                              int64
meal                               object
country                            object
market_segment                     object
distribution_channel               object
is_repeated_guest                   int64
previous_cancellations              int64
previous_bookings_not_canceled      int64
reserved_room_type                 object
assigned_room_type                 object
booking_changes                     int64
deposit_type                       object
agent                             float64
company                           float64
days_in_waiting_list                int64
customer_type                      object
adr                               float64
required_car_parking_spaces         int64
total_of_special_requests           int64
reservation_status                 object
reservation_status_date            object
name                               object
email                              object
phone-number                       object
credit_card                        object
dtype: object
```

```
df.isna().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.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')
```

```
# 'is_canceled' is our main veriable 0 represnt booking not canceled 1 represent booking canceled
```

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

```
df.dtypes
```

```
    hotel                                     object
    is_canceled                                int64
    lead_time                                  int64
    arrival_date_year                          int64
    arrival_date_month                        object
    arrival_date_week_number                   int64
    arrival_date_day_of_month                  int64
    stays_in_weekend_nights                    int64
    stays_in_week_nights                       int64
    adults                                     int64
    children                                 float64
    babies                                     int64
    meal                                      object
    country                                   object
    market_segment                            object
    distribution_channel                      object
    is_repeated_guest                          int64
    previous_cancellations                     int64
    previous_bookings_not_canceled             int64
    reserved_room_type                        object
    assigned_room_type                        object
    booking_changes                            int64
    deposit_type                              object
    agent                                    float64
    company                                  float64
    days_in_waiting_list                       int64
    customer_type                             object
    adr                                      float64
    required_car_parking_spaces                int64
    total_of_special_requests                  int64
    reservation_status                        object
    reservation_status_date           datetime64[ns]
    name                                      object
    email                                     object
    phone-number                              object
    credit_card                               object
    dtype: object
```

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

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

```
# find unique in each columns
for i in df.describe(include='object').columns:
  print(i)
  print(df[i].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']
--------------------------------------------------
```

```
df.isna().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.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	a
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.drop(['agent','company'],axis=1,inplace=True)
```

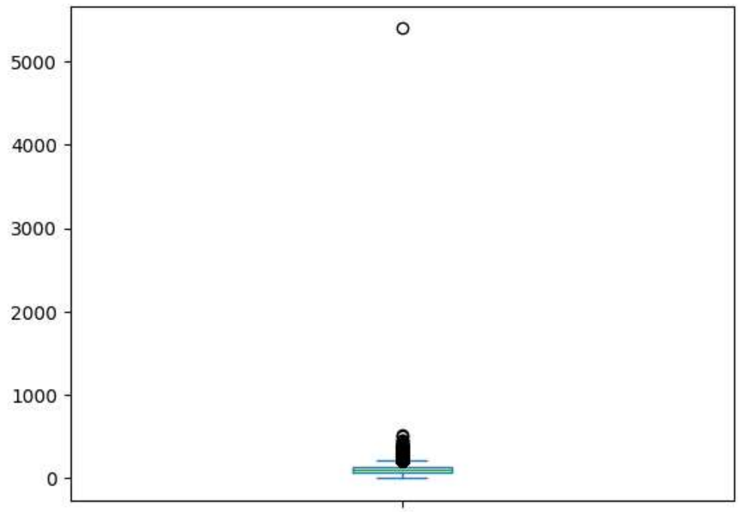
```
df.dropna(inplace=True)
```

```
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
```

```
#check Outliers
df['adr'].plot(kind='box') #adr-avg daily rate
```

```
<Axes: >
```



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

```
df.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	a
count	118897.000000	118897.000000	118897.000000	118897.000000	
mean	0.371347	104.312018	2016.157657	27.166674	
std	0.483167	106.903570	0.707462	13.589966	
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	

```
#Data Analysis and Visualizations
```

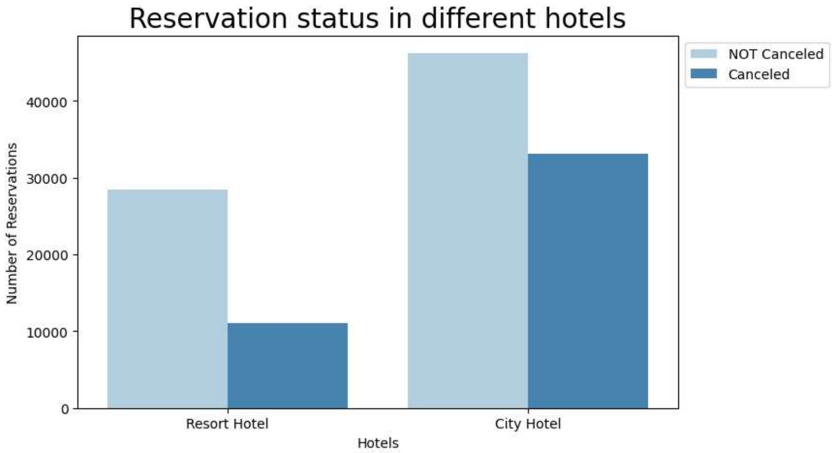
```
cancelled_per=df['is_canceled'].value_counts(normalize=True)
cancelled_per
```

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

```
plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['Not Canceled','Canceled'],df['is_canceled'].value_counts())
plt.show()
```

```
plt.figure(figsize=(8, 5))
ax1 = sns.countplot(x='hotel', hue='is_canceled', data=df, palette='Blues')

legend_labels, _ = ax1.get_legend_handles_labels()
ax1.legend(legend_labels, ['NOT Canceled', 'Canceled'], bbox_to_anchor=(1, 1))
plt.title("Reservation status in different hotels", size=20)
plt.xlabel('Hotels')
plt.ylabel('Number of Reservations')
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)
```

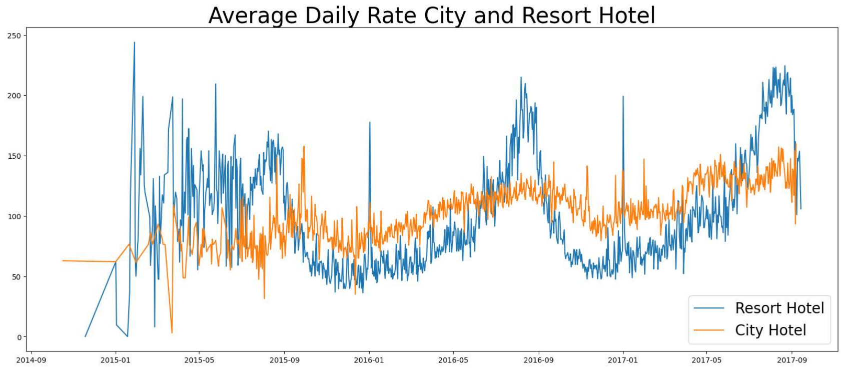
```
    0    0.582918
    1    0.417082
    Name: is_canceled, dtype: float64
```

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

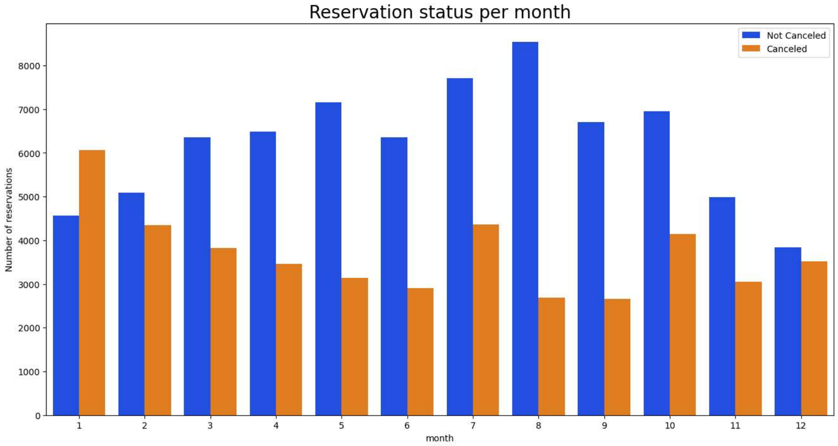
```
resort_hotel
```

	adr
reservation_status_date	
2014-11-18	0.000000
2015-01-01	61.966667
2015-01-02	9.633750
2015-01-18	0.000000
2015-01-21	37.301209
...	...

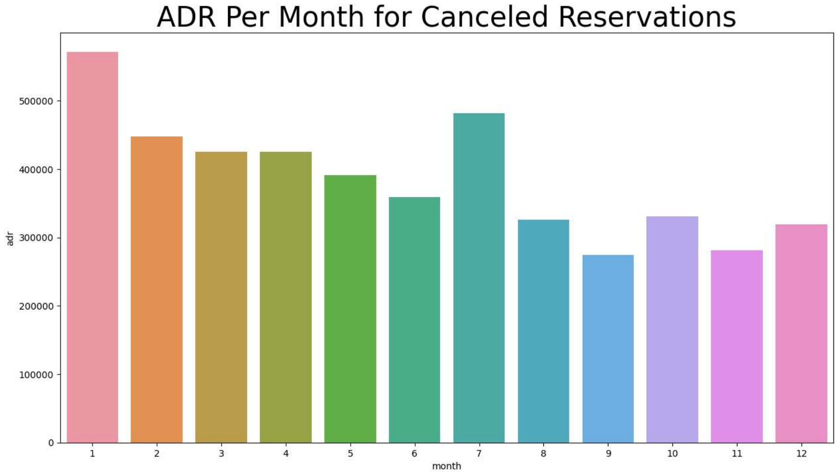
```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate City and Resort Hotel',fontsize=30)
plt.plot(resort_hotel.index,resort_hotel['adr'],label= 'Resort Hotel')
plt.plot(city_hotel.index,city_hotel['adr'],label= 'City Hotel')
plt.legend(fontsize=20)
plt.show()
```



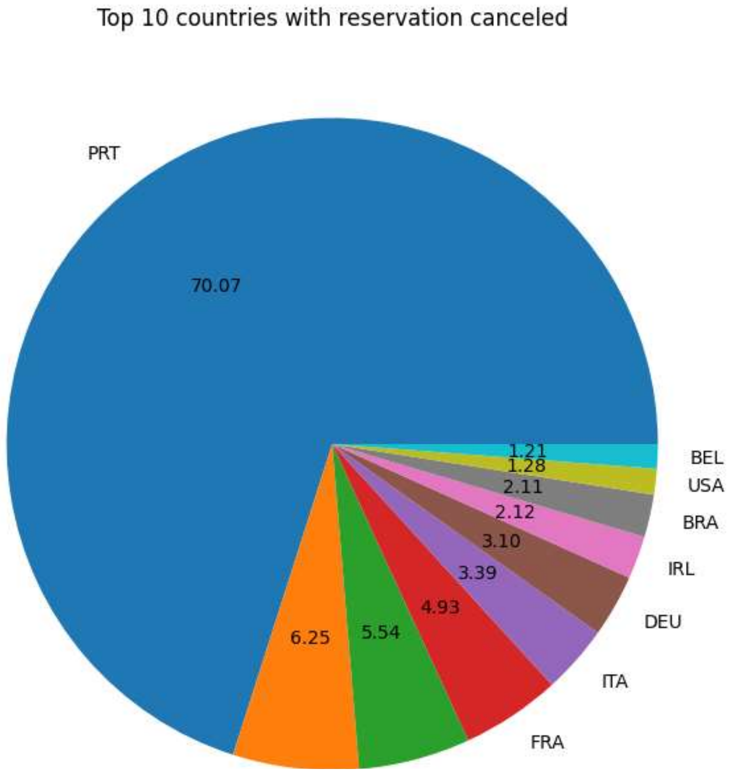
```
df['month']=df['reservation_status_date'].dt.month    #create month column
plt.figure(figsize=(16,8))
ax1=sns.countplot(x='month',hue='is_canceled',data=df,palette='bright')
legend_labels,_=ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per month',size=20)
plt.xlabel('month')
plt.ylabel('Number of reservations')
plt.legend(['Not Canceled' ,'Canceled'])
plt.show()
```



```
plt.figure(figsize =(15,8))
plt.title('ADR Per Month for Canceled Reservations', fontsize=30)
sns.barplot(x='month',y='adr',data=df[df['is_canceled']==1].groupby('month')[['adr']].sum().reset_index())
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)
```

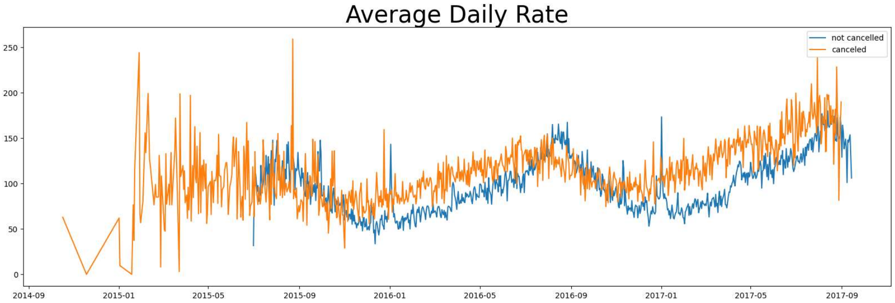
```
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: market_segment, dtype: float64
```

```
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)
```

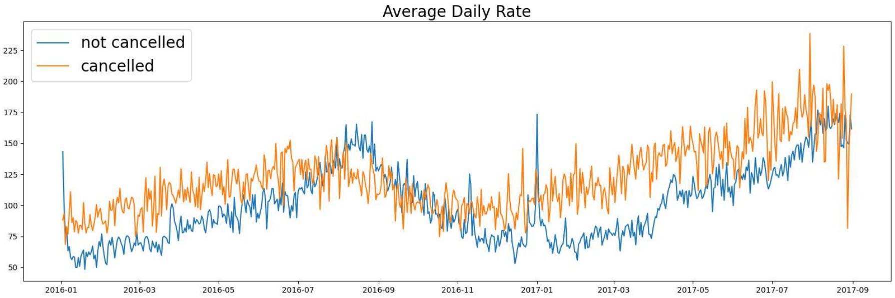
```
plt.figure(figsize=(20,6))
plt.title('Average Daily Rate',fontsize=30)
plt.plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr['adr'],label='not cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label='canceled')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x7cc029f37e20>
```



```
cancelled_df_adr=cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') & (cancelled_df_adr['reservation_status_date']<'2017-0
not_cancelled_df_adr=not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016') & (not_cancelled_df_adr['reservation_statu
```

```
plt.figure(figsize=(20,6))
plt.title('Average Daily Rate',fontsize=20)
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)
plt.show()
```


Average Daily Rate
not cancelled
cancelled