data cleaning

In [8]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from tabulate import tabulate
import warnings
import pandoc
import nbconvert
warnings.filterwarnings('ignore')
```

In [9]:
```
#pip install nbconvert
```

In [10]:
```
df=pd.read_csv(r'C:\Users\ritika shukla\Downloads\archive (3)\hotel_booking.csv')
```

In [11]:
```
df.columns
```

Out[11]:
```
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 [12]:
```
df.drop(['name','email','phone-number','credit_card','company','children','babies'], axis='columns', inplace=Tr
```

In [13]:
```
df.head(100)
df.tail()
```

Out[13]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
119385	City Hotel	0	23	2017	August	35	30
119386	City Hotel	0	102	2017	August	35	31
119387	City Hotel	0	34	2017	August	35	31
119388	City Hotel	0	109	2017	August	35	31
119389	City Hotel	0	205	2017	August	35	29

5 rows × 29 columns

In [14]:
```
df.head()
```

Out[14]:

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

5 rows × 29 columns

In [15]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 29 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  meal                            119390 non-null  object
 11  country                         118902 non-null  object
 12  market_segment                  119390 non-null  object
 13  distribution_channel            119390 non-null  object
 14  is_repeated_guest               119390 non-null  int64
 15  previous_cancellations          119390 non-null  int64
 16  previous_bookings_not_canceled  119390 non-null  int64
 17  reserved_room_type              119390 non-null  object
 18  assigned_room_type              119390 non-null  object
 19  booking_changes                 119390 non-null  int64
 20  deposit_type                    119390 non-null  object
 21  agent                           103050 non-null  float64
 22  days_in_waiting_list            119390 non-null  int64
 23  customer_type                   119390 non-null  object
 24  adr                             119390 non-null  float64
 25  required_car_parking_spaces     119390 non-null  int64
 26  total_of_special_requests       119390 non-null  int64
 27  reservation_status              119390 non-null  object
 28  reservation_status_date         119390 non-null  object
dtypes: float64(2), int64(15), object(12)
memory usage: 26.4+ MB
```

In [16]:
```
df['reservation_status_date']=pd.to_datetime(df['reservation_status_date'])
df.head()
```

Out[16]:

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

5 rows × 29 columns

In [17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 29 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  meal                            119390 non-null  object
 11  country                         118902 non-null  object
 12  market_segment                  119390 non-null  object
 13  distribution_channel            119390 non-null  object
 14  is_repeated_guest               119390 non-null  int64
 15  previous_cancellations          119390 non-null  int64
 16  previous_bookings_not_canceled  119390 non-null  int64
 17  reserved_room_type              119390 non-null  object
 18  assigned_room_type              119390 non-null  object
 19  booking_changes                 119390 non-null  int64
 20  deposit_type                    119390 non-null  object
 21  agent                           103050 non-null  float64
 22  days_in_waiting_list            119390 non-null  int64
 23  customer_type                   119390 non-null  object
 24  adr                             119390 non-null  float64
 25  required_car_parking_spaces     119390 non-null  int64
 26  total_of_special_requests       119390 non-null  int64
 27  reservation_status              119390 non-null  object
 28  reservation_status_date         119390 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(15), object(11)
memory usage: 26.4+ MB
```

In [18]:
```
df.describe(include='object')
```

Out[18]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_ty
count	119390	119390	119390	118902	119390	119390	119390	119
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	74

In [19]:
```
df.describe()
```

Out[19]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_ni
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.00
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.92
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.00
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.00
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.00
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.00
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.00
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.99

In [20]:
```
df=df[df['adr']<5000]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 119389 entries, 0 to 119389
Data columns (total 29 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119389 non-null  object
 1   is_canceled                     119389 non-null  int64
 2   lead_time                       119389 non-null  int64
 3   arrival_date_year               119389 non-null  int64
 4   arrival_date_month              119389 non-null  object
 5   arrival_date_week_number        119389 non-null  int64
 6   arrival_date_day_of_month       119389 non-null  int64
 7   stays_in_weekend_nights         119389 non-null  int64
 8   stays_in_week_nights            119389 non-null  int64
 9   adults                          119389 non-null  int64
 10  meal                            119389 non-null  object
 11  country                         118901 non-null  object
 12  market_segment                  119389 non-null  object
 13  distribution_channel            119389 non-null  object
 14  is_repeated_guest               119389 non-null  int64
 15  previous_cancellations          119389 non-null  int64
 16  previous_bookings_not_canceled  119389 non-null  int64
 17  reserved_room_type              119389 non-null  object
 18  assigned_room_type              119389 non-null  object
 19  booking_changes                 119389 non-null  int64
 20  deposit_type                    119389 non-null  object
 21  agent                           103049 non-null  float64
 22  days_in_waiting_list            119389 non-null  int64
 23  customer_type                   119389 non-null  object
 24  adr                             119389 non-null  float64
 25  required_car_parking_spaces     119389 non-null  int64
 26  total_of_special_requests       119389 non-null  int64
 27  reservation_status              119389 non-null  object
 28  reservation_status_date         119389 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(15), object(11)
memory usage: 27.3+ MB
```

In [21]:
```
df.describe()
```

Out[21]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_ni
count	119389.000000	119389.000000	119389.000000	119389.000000	119389.000000	119389.00
mean	0.370411	104.011994	2016.156555	27.165292	15.798164	0.92
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.00
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.00
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.00
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.00
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.00
std	0.482917	106.863358	0.707479	13.605134	8.780826	0.99

In [22]:
```
df['agent'].fillna(0,inplace=True)
df['country'].fillna('x' ,inplace=True)
```

In [23]:
```
df.head(100)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 119389 entries, 0 to 119389
Data columns (total 29 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119389 non-null  object
 1   is_canceled                     119389 non-null  int64
 2   lead_time                       119389 non-null  int64
 3   arrival_date_year               119389 non-null  int64
 4   arrival_date_month              119389 non-null  object
 5   arrival_date_week_number        119389 non-null  int64
 6   arrival_date_day_of_month       119389 non-null  int64
 7   stays_in_weekend_nights         119389 non-null  int64
 8   stays_in_week_nights            119389 non-null  int64
 9   adults                          119389 non-null  int64
 10  meal                            119389 non-null  object
 11  country                         119389 non-null  object
 12  market_segment                  119389 non-null  object
 13  distribution_channel            119389 non-null  object
 14  is_repeated_guest               119389 non-null  int64
 15  previous_cancellations          119389 non-null  int64
 16  previous_bookings_not_canceled  119389 non-null  int64
 17  reserved_room_type              119389 non-null  object
 18  assigned_room_type              119389 non-null  object
 19  booking_changes                 119389 non-null  int64
 20  deposit_type                    119389 non-null  object
 21  agent                           119389 non-null  float64
 22  days_in_waiting_list            119389 non-null  int64
 23  customer_type                   119389 non-null  object
 24  adr                             119389 non-null  float64
 25  required_car_parking_spaces     119389 non-null  int64
 26  total_of_special_requests       119389 non-null  int64
 27  reservation_status              119389 non-null  object
 28  reservation_status_date         119389 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(15), object(11)
memory usage: 27.3+ MB
```

Data Visualization

In [24]:
```
canceled= df['is_canceled'].value_counts(normalize=True)
print(canceled)
```

```
is_canceled
0    0.629589
1    0.370411
Name: proportion, dtype: float64
```

In [25]:
```
canceled=sns.catplot(data=df, kind='count', x='is_canceled',hue='hotel', height=4, aspect=2, palette='Accent')
canceled.set_xticklabels(['not_canceled','canceled'])
plt.title("Cancelation Rate")
plt.show()
```

In [26]:
```
City_hotel_cancelation= df[df['hotel']=='City Hotel']
City_hotel_cancelation['is_canceled'].value_counts(normalize=True)
```

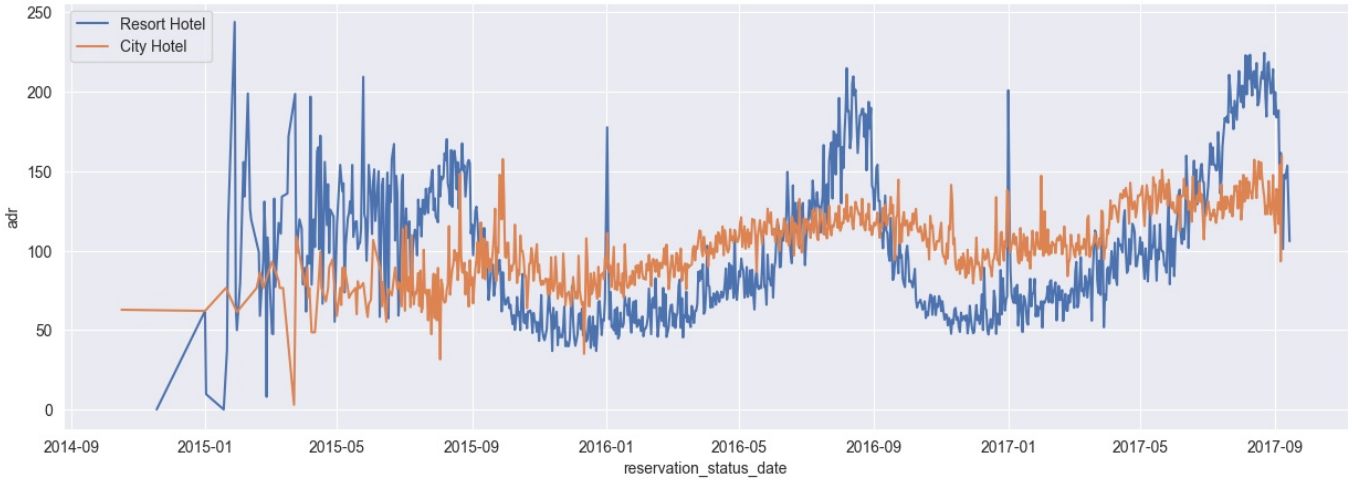
Out[26]:
```
is_canceled
0    0.582738
1    0.417262
Name: proportion, dtype: float64
```

In [27]:
```
Resort_hotel_cancelation= df[df['hotel']=='Resort Hotel']
Resort_hotel_cancelation['is_canceled'].value_counts(normalize=True)
```

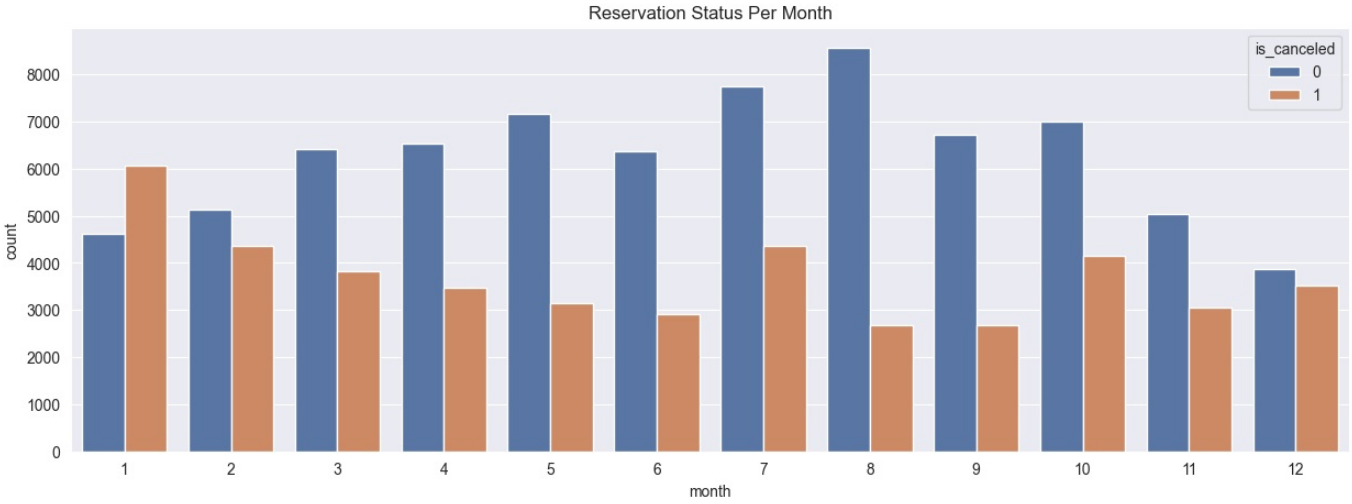
Out[27]:
```
is_canceled
0    0.722366
1    0.277634
Name: proportion, dtype: float64
```

In [28]:
```
Resort_hotel = Resort_hotel_cancelation.groupby('reservation_status_date')[['adr']].mean()
City_hotel = City_hotel_cancelation.groupby('reservation_status_date')[['adr']].mean()
```

In [29]:
```
sns.set({'figure.figsize': (15,5)})
sns.lineplot(data=Resort_hotel, x=Resort_hotel.index, y=Resort_hotel['adr'], label='Resort Hotel')
sns.lineplot(data=City_hotel, x=City_hotel.index, y=City_hotel['adr'], label='City Hotel')
plt.legend(fontsize=10)
plt.show()
```



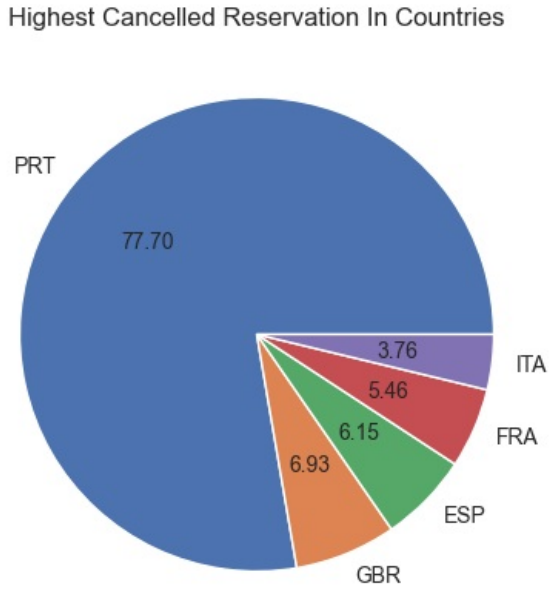
In [30]:
```
plt.title('Reservation Status Per Month')
df['month']= df['reservation_status_date'].dt.month
sns.countplot(x='month', hue='is_canceled', data=df)
plt.show()
```



In [31]:
```
plt.title('ADR per month')
sns.barplot(x='month', y='adr', data=df[df['is_canceled']==1].groupby('month')[['adr']].sum().reset_index())
plt.show()
```

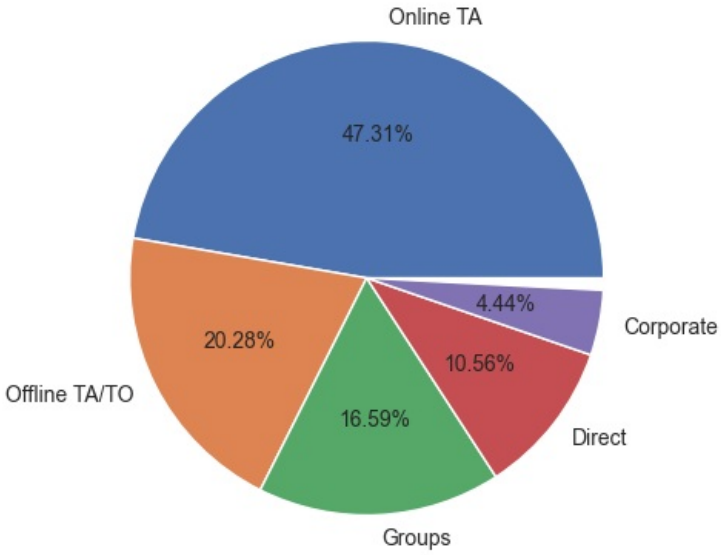
In [45]:
```
sns.set({'figure.figsize' :(5,5)})
plt.title("Highest Cancelled Reservation In Countries")
canceled_data = df[df['is_canceled']==1]
top_10_country=canceled_data['country'].value_counts()[:5]
patches, texts, autotexts=plt.pie(top_10_country, autopct='%.2f', labels=top_10_country.index)

plt.show()
```



In [41]:
```
sns.set({'figure.figsize' :(5,5)})
market_segment_percent=((df['market_segment'].value_counts()/df['market_segment'].value_counts().sum())*100)
print(market_segment_percent[:5])
patches, texts, autotexts=plt.pie(market_segment_percent, autopct='%.2f%%', labels= market_segment_percent.index
patches[6].set_visible(False)
texts[6].set_visible(False)
autotexts[6].set_visible(False)
patches[5].set_visible(False)
texts[5].set_visible(False)
autotexts[5].set_visible(False)
patches[7].set_visible(False)
texts[7].set_visible(False)
autotexts[7].set_visible(False)
plt.show()
```

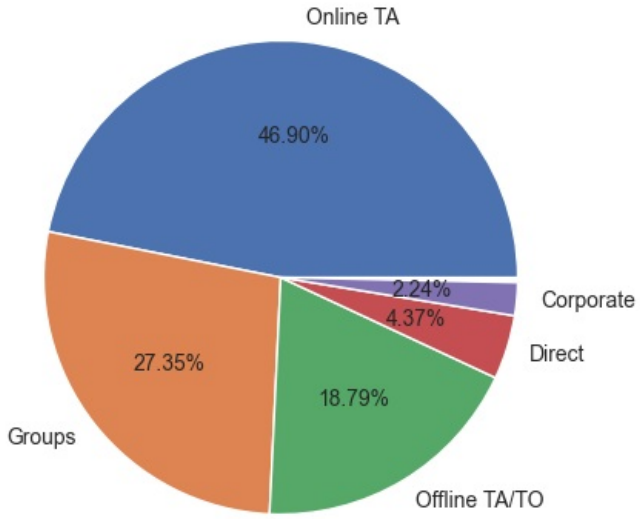
```
market_segment
Online TA        47.305028
Offline TA/TO    20.284951
Groups           16.593656
Direct           10.558762
Corporate         4.435082
Name: count, dtype: float64
```



In [42]:
```
sns.set({'figure.figsize' :(5,5)})
market_segment_percent=(canceled_data['market_segment'].value_counts()/canceled_data['market_segment'].value_co
print(market_segment_percent[:5])
patches, texts, autotexts=plt.pie(market_segment_percent, autopct='%.2f%%', labels= market_segment_percent.inde
patches[6].set_visible(False)
texts[6].set_visible(False)
autotexts[6].set_visible(False)
patches[5].set_visible(False)
texts[5].set_visible(False)
autotexts[5].set_visible(False)
patches[7].set_visible(False)
texts[7].set_visible(False)
autotexts[7].set_visible(False)

plt.show()
```

```
market_segment
Online TA        46.896411
Groups           27.354544
Offline TA/TO    18.791127
Direct            4.373290
Corporate         2.243177
Name: count, dtype: float64
```

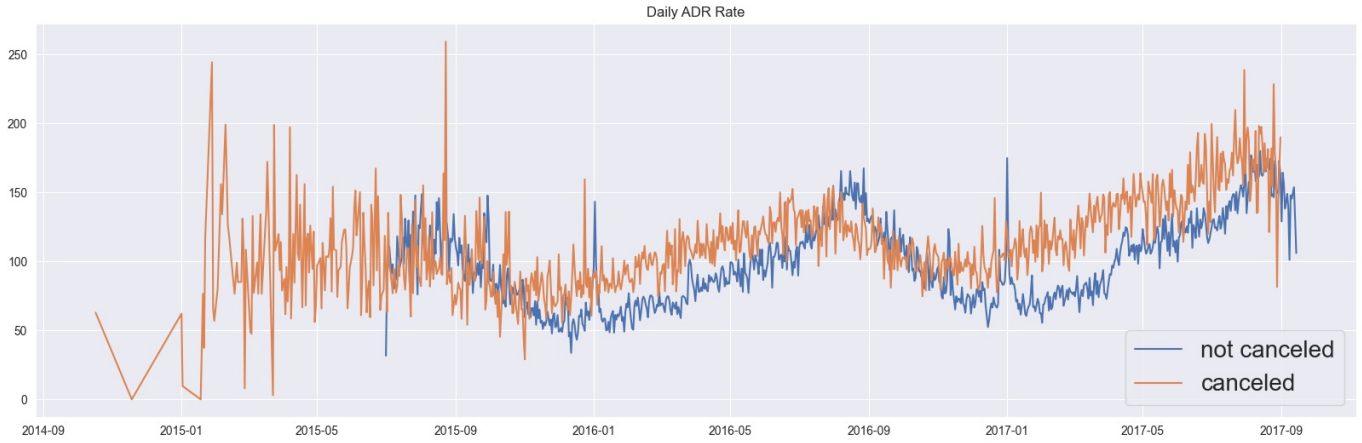


In [37]:
```
not_canceled_data = df[df['is_canceled']==0]
not_canceled_data_adr= not_canceled_data.groupby('reservation_status_date')[['adr']].mean()
not_canceled_data_adr.reset_index(inplace=True)
```

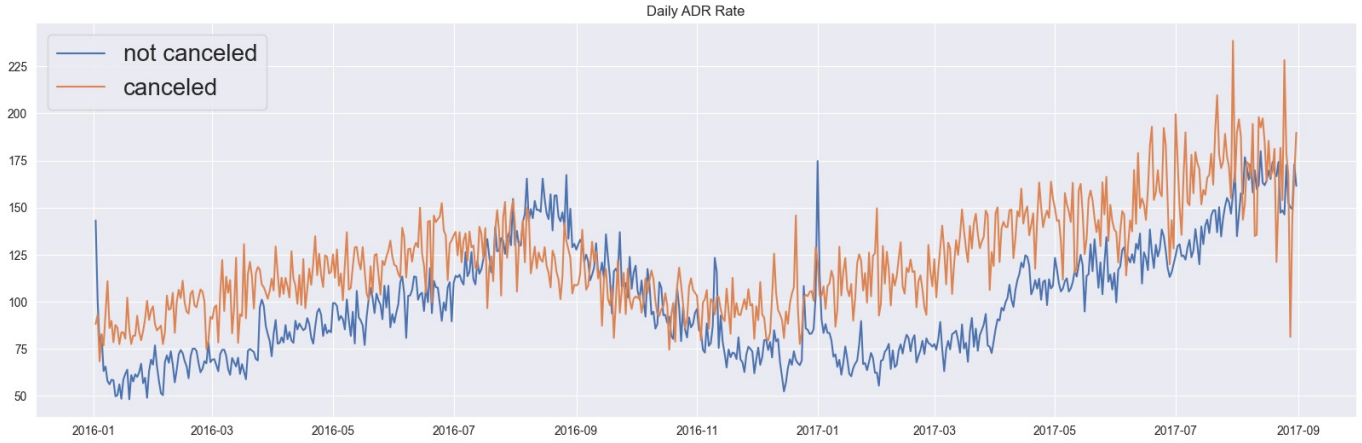
```
not_canceled_data_adr.sort_values('reservation_status_date', inplace=True)

canceled_data = df[df['is_canceled']==1]
canceled_data_adr= canceled_data.groupby('reservation_status_date')[['adr']].mean()
canceled_data_adr.reset_index(inplace=True)
canceled_data_adr.sort_values('reservation_status_date', inplace=True)

plt.figure(figsize=(20,6))
plt.title("Daily ADR Rate")
plt.plot(not_canceled_data_adr['reservation_status_date'], not_canceled_data_adr['adr'], label='not canceled')
plt.plot(canceled_data_adr['reservation_status_date'], canceled_data_adr['adr'], label='canceled')
plt.legend(fontsize=20)
plt.show()
```



In [38]:
```
not_canceled_data_adr=not_canceled_data_adr[(not_canceled_data_adr['reservation_status_date']>'2016') & (not_ca
canceled_data_adr=canceled_data_adr[(canceled_data_adr['reservation_status_date']>'2016') & (canceled_data_adr[
plt.figure(figsize=(20,6))
plt.title("Daily ADR Rate")
plt.plot(not_canceled_data_adr['reservation_status_date'], not_canceled_data_adr['adr'], label='not canceled')
plt.plot(canceled_data_adr['reservation_status_date'], canceled_data_adr['adr'], label='canceled')
plt.legend(fontsize=20)
plt.show()
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: