In [92]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [93]:

```
import pandas as pd

df = pd.read_csv('hotel_booking.csv')
```

In [94]:

```
df.head()
```

Out[94]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
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

In [95]:

```
df.tail()
```

Out[95]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num
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 × 32 columns

In [96]:

```
df.shape
```

Out[96]:

```
(119390, 32)
```

In [97]:

```
df.columns
```

Out[97]:

```
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')
```

In [98]:

```
df.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
```

In [100…

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

In [101…

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

Out[101]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reser
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	

In [102…

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

```
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']
```

In [103…

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

Out[103]:

```
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
```

In [104…

```
df.describe()
```

is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weel
9390.000000	119390.000000	119390.000000	119390.000000	119390.000000	11!
0.370416	104.011416	2016.156554	27.165173	15.798241	
0.482918	106.863097	0.707476	13.605138	8.780829	
0.000000	0.000000	2015.000000	1.000000	1.000000	
0.000000	18.000000	2016.000000	16.000000	8.000000	
0.000000	69.000000	2016.000000	28.000000	16.000000	
1.000000	160.000000	2017.000000	38.000000	23.000000	
1.000000	737.000000	2017.000000	53.000000	31.000000	

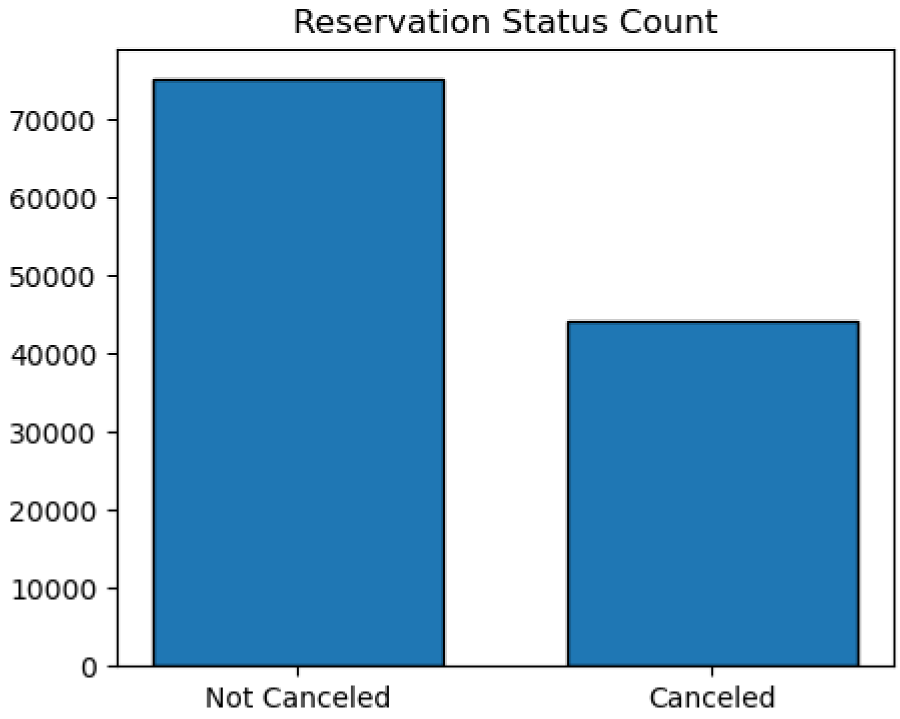
In [105…

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

In [106…

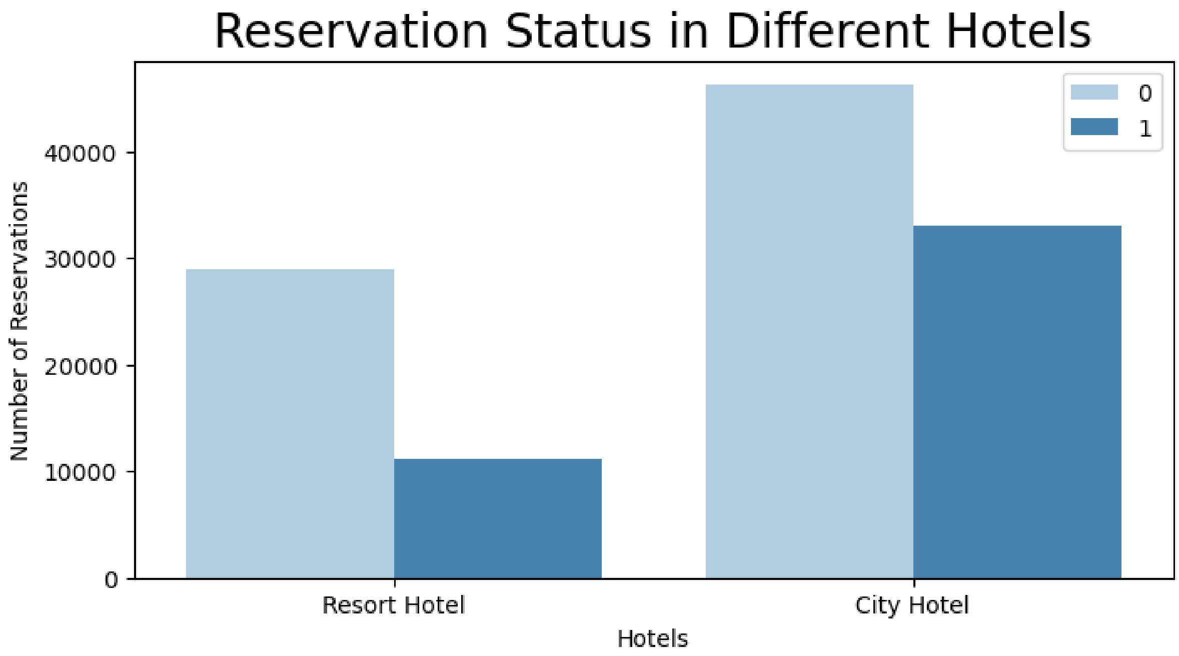
```
import pandas as pd
import matplotlib.pyplot as plt

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='
plt.show()
```

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



In [107…

```
import seaborn as sns
import matplotlib.pyplot as plt

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('Hotels')
plt.ylabel('Number of Reservations')
plt.show()
```



In [108…

```
df['hotel'] = df['hotel'].str.lower()
resort_hotel = df[df['hotel'] == 'resort hotel']
resort_cancelled_perc = resort_hotel['is_canceled'].value_counts(normalize=True)
print(resort_cancelled_perc)
```

```
0    0.722366
1    0.277634
Name: is_canceled, dtype: float64
```

In [109…

```
df['hotel'] = df['hotel'].str.lower()
city_hotel = df[df['hotel'] == 'city hotel']
city_cancelled_perc = city_hotel['is_canceled'].value_counts(normalize=True)
print(city_cancelled_perc)
```

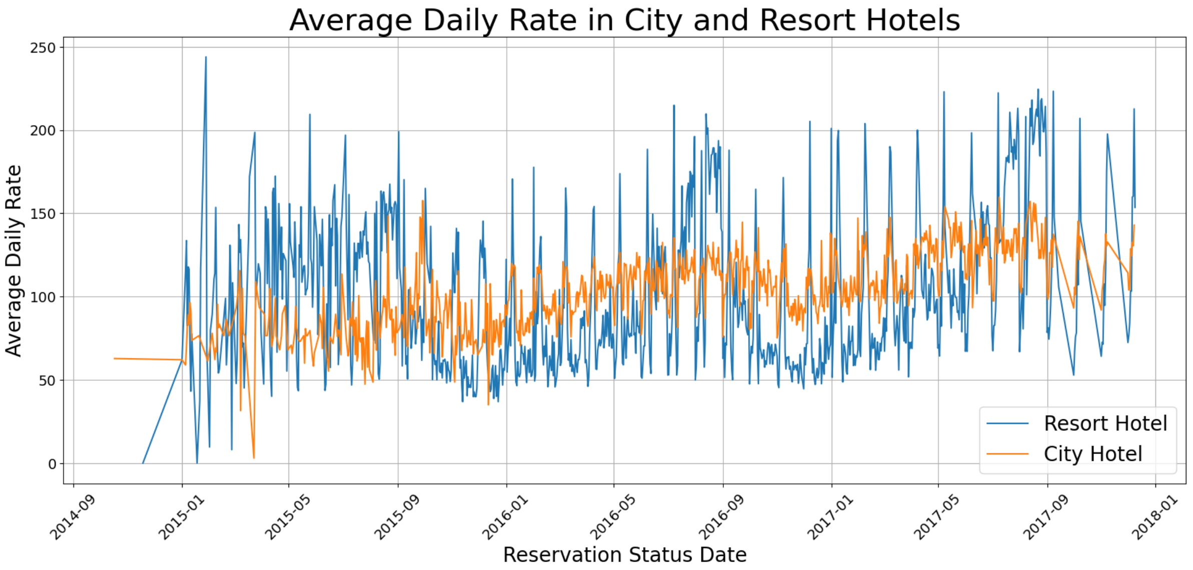
```
0    0.582738
1    0.417262
Name: is_canceled, dtype: float64
```

In [110…

```
resort_hotel = df[df['hotel'] == 'resort hotel'].groupby('reservation_status_date')
city_hotel = df[df['hotel'] == 'city hotel'].groupby('reservation_status_date')[['a
```

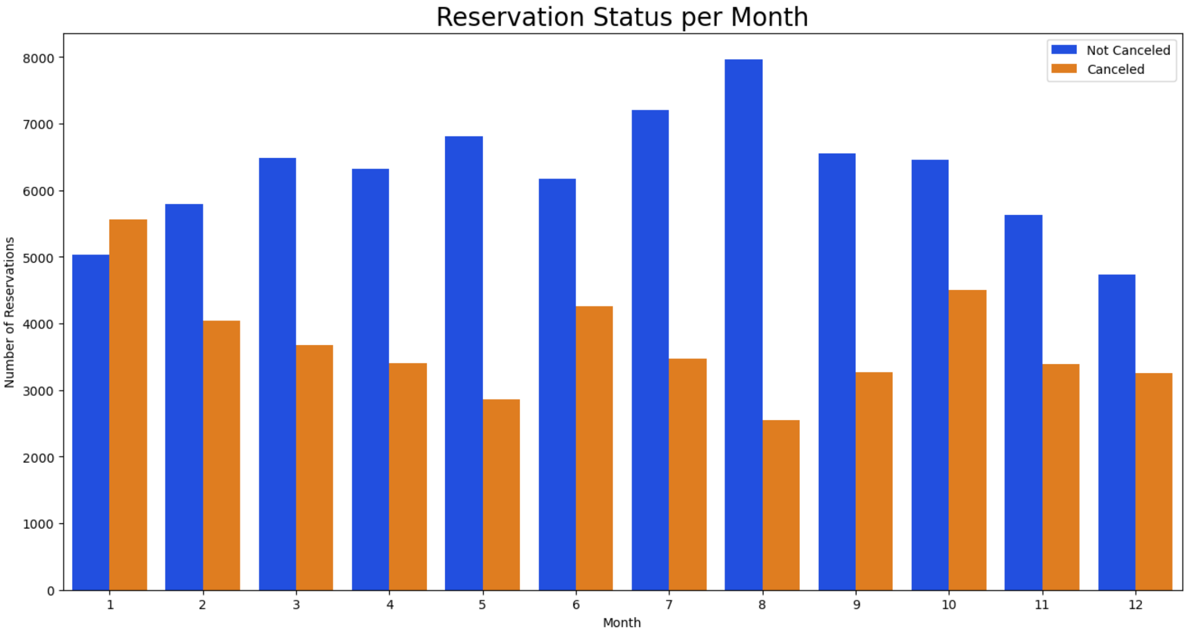
In [111…

```
import matplotlib.pyplot as plt

plt.figure(figsize=(20, 8))
plt.title('Average Daily Rate in City and Resort Hotels', 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.xlabel('Reservation Status Date', fontsize=20)
plt.ylabel('Average Daily Rate', fontsize=20)
plt.xticks(fontsize=14, rotation=45)
plt.yticks(fontsize=14)
plt.grid(True)
plt.show()
```



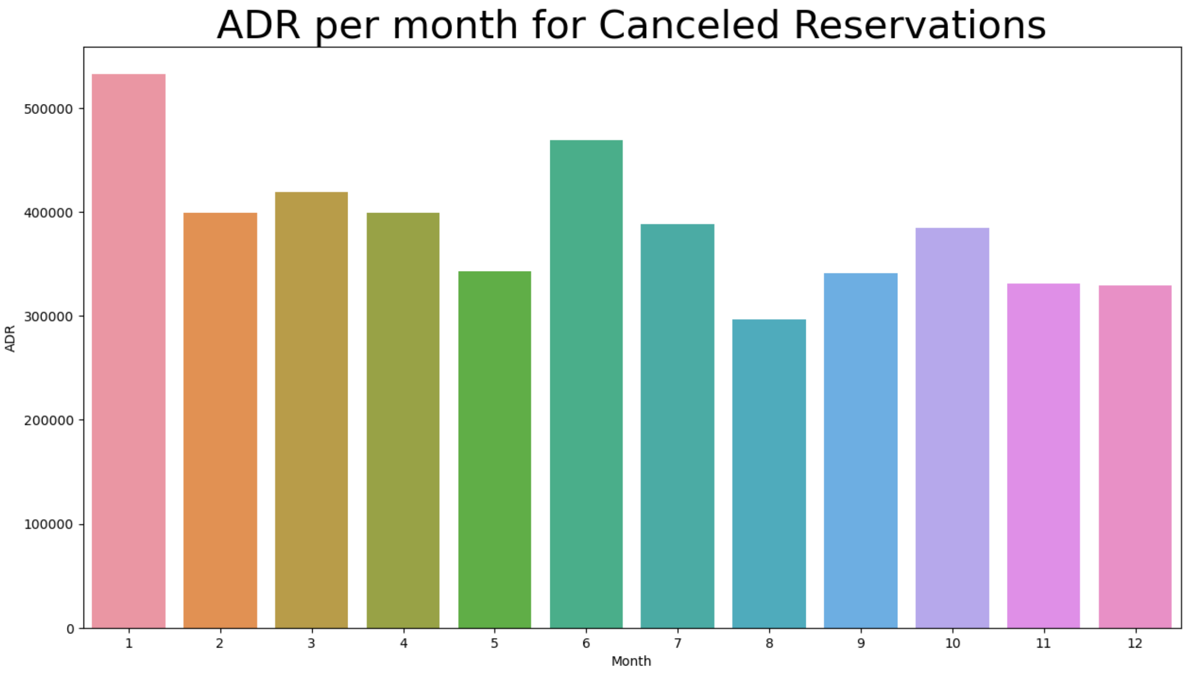
In [112…

```
import matplotlib.pyplot as plt
import seaborn as sns

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()
```



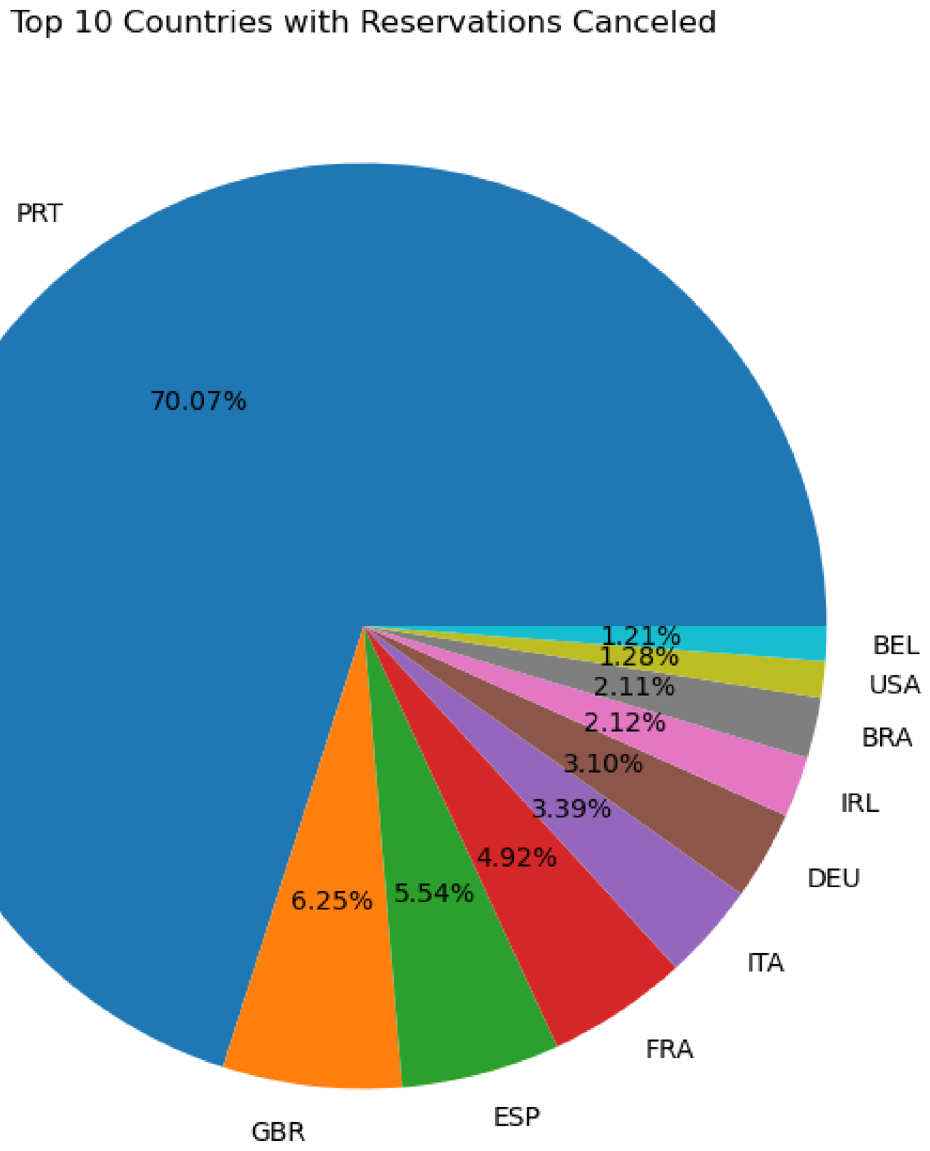
In [113…

```
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')[['
plt.xlabel('Month')
plt.ylabel('ADR')
plt.show()
```



In [114…

```
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 Reservations Canceled')
plt.pie(top_10_country, autopct='%.2f%%', labels=top_10_country.index)
plt.show()
```



Top 10 Countries with Reservations Canceled

```
In [115…  df['market_segment'].value_counts()
```

```
Out[115]: Online TA        56477
          Offline TA/TO    24218
          Groups           19811
          Direct           12606
          Corporate         5295
          Complementary      743
          Aviation           237
          Undefined            2
          Name: market_segment, dtype: int64
```

```
In [116…  df['market_segment'].value_counts(normalize = True)
```

```
Out[116]: Online TA        0.473050
          Offline TA/TO    0.202850
          Groups           0.165937
          Direct           0.105588
          Corporate        0.044351
          Complementary    0.006223
          Aviation         0.001985
          Undefined        0.000017
          Name: market_segment, dtype: float64
```

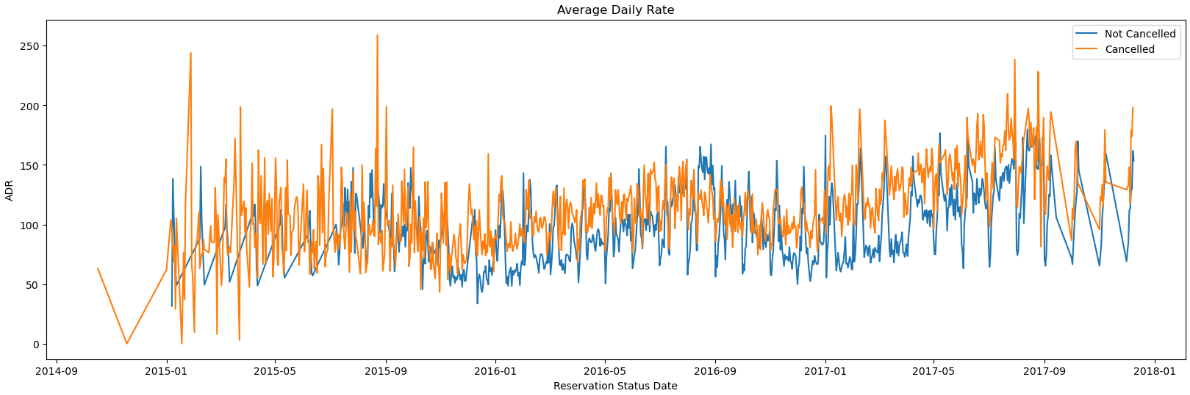
```
In [117…  cancelled_data['market_segment'].value_counts(normalize = True)
```

Out[117]:

```
Online TA        0.468964
Groups           0.273545
Offline TA/TO    0.187911
Direct           0.043733
Corporate        0.022432
Complementary    0.002193
Aviation         0.001176
Undefined        0.000045
Name: market_segment, dtype: float64
```

In [118…

```
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_data_adr = not_cancelled_data.groupby('reservation_status_date')[['ad
not_cancelled_data_adr.reset_index(inplace=True)
not_cancelled_data_adr.sort_values('reservation_status_date', inplace=True)

plt.figure(figsize=(20, 6))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_data_adr['reservation_status_date'], not_cancelled_data_adr[
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], labe
plt.xlabel('Reservation Status Date')
plt.ylabel('ADR')
plt.legend()
plt.show()
```



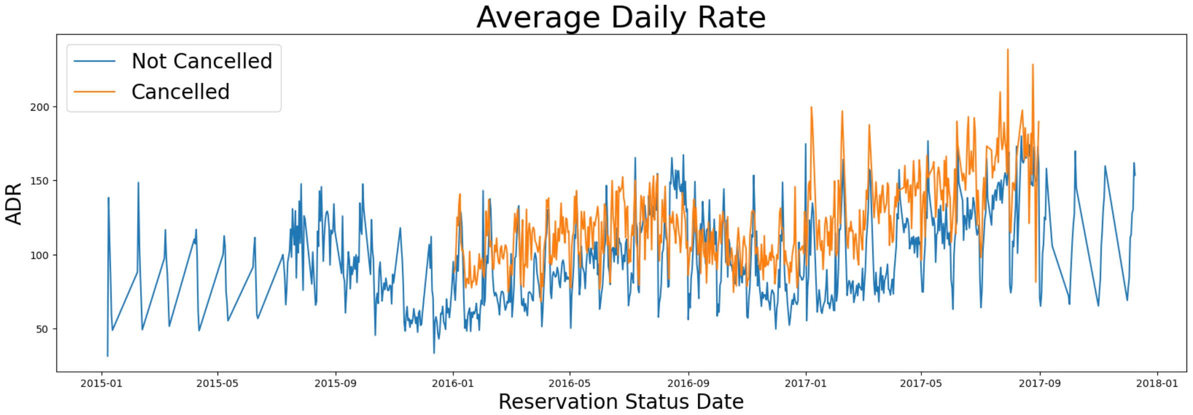
In [121…

```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] >
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_stat
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[121], line 2
      1 cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_
date'] > '2016') & (cancelled_df_adr['reservation_status_date'] <'2017-09')]
----> 2 not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reserva
tion_status_date'] > '2016') & (not_cancelled_df_adr['reservation_status_date']
<'2017-09')]

NameError: name 'not_cancelled_df_adr' is not defined
```

In [123…

```
plt.figure(figsize=(20, 6))
plt.title('Average Daily Rate', fontsize=30)
plt.plot(not_cancelled_data_adr['reservation_status_date'], not_cancelled_data_adr[
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], labe
plt.xlabel('Reservation Status Date', fontsize=20)
plt.ylabel('ADR', fontsize=20)
plt.legend(fontsize=20)
plt.show()
```


Average Daily Rate
Not Cancelled
Cancelled
ADR
200
150
100
50
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
Reservation Status Date

In [ ]: