In [1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```python
df = pd.read_csv('hotel_booking.csv')
```

In [3]:
```python
df.head()
```

Out[3]:

	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	...
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...

5 rows × 32 columns

Exploratory data analysis

In [4]:
```python
df.shape
```

Out[4]:
```
(119390, 32)
```

In [5]:
```python
df.columns
```

Out[5]:
```
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 [6]:
```python
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 [7]:
```python
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

In [8]:
```python
df.describe(include = 'object')
```

Out[8]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166

In [9]:
```python
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']
--------------------------------------------------
```

In [10]:
```python
df.isnull().sum()
```

Out[10]:
```
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 [11]:
```python
df.drop(['company','agent'], axis = 1 , inplace = True)
df.dropna(inplace = True)
```

In [12]:
```python
df.isnull().sum()
```

Out[12]:
```
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
dtype: int64
```

In [13]:
```python
df.describe()
```

Out[13]:

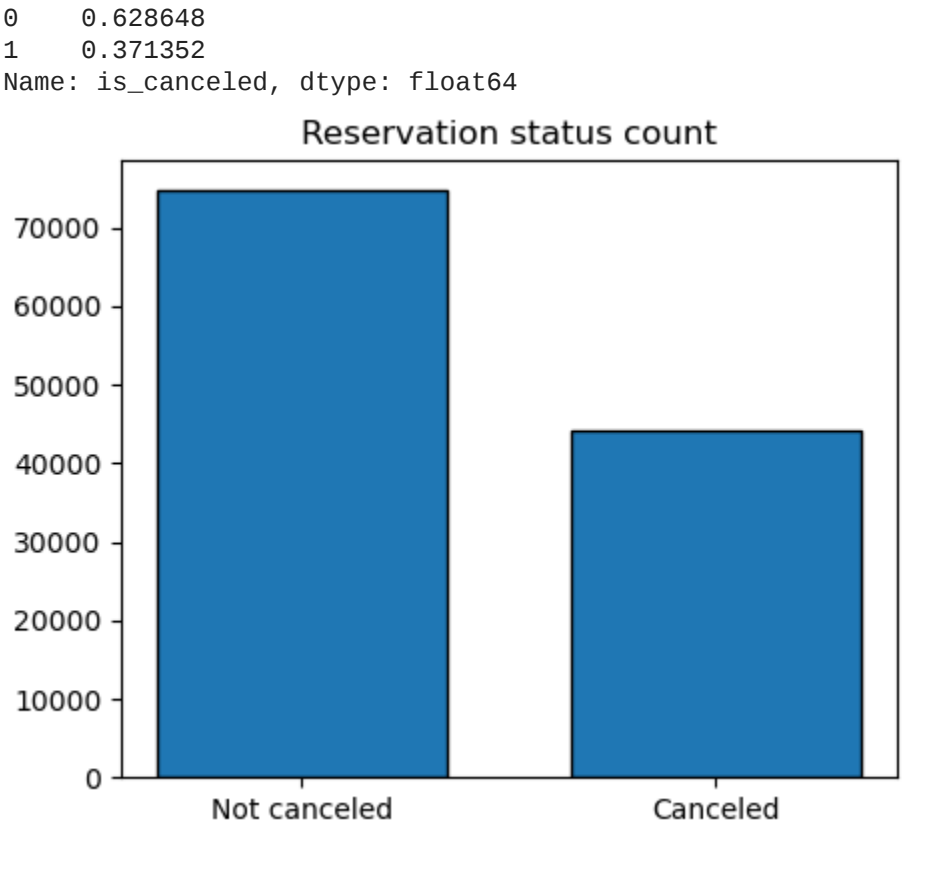
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	childre
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.00000
mean	0.371352	104.311435	2016.157656	27.166555	15.800880	0.928897	2.502145	1.858391	0.10420
std	0.483168	106.903309	0.707459	13.589971	8.780324	0.996216	1.900168	0.578576	0.39917
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.00000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.00000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.00000
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.00000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	41.000000	55.000000	10.00000

Data analysis and Visualization

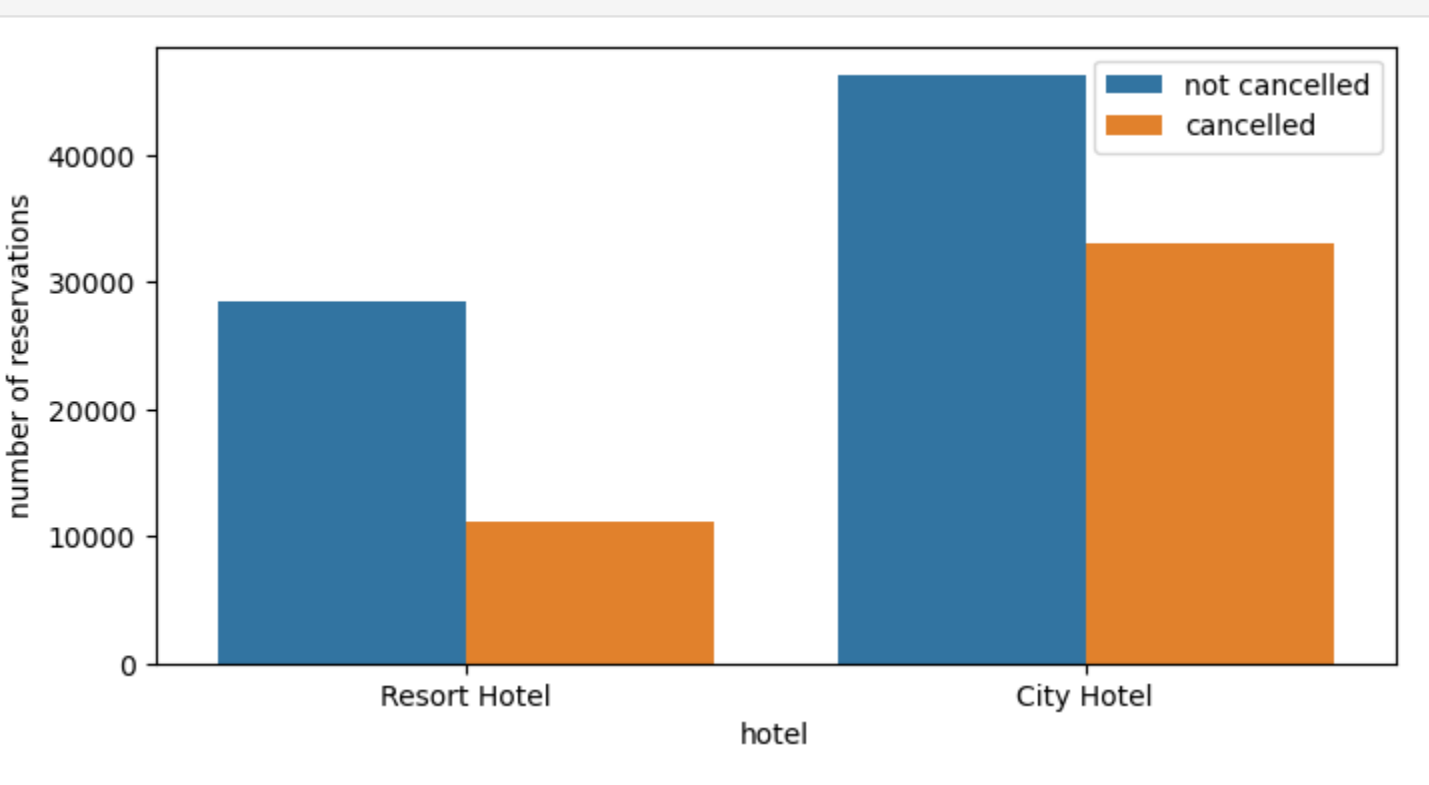
In [14]:
```python
cancelled_percentage  = df['is_canceled'].value_counts(normalize = True)
print(cancelled_percentage)

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    0.628648
1    0.371352
Name: is_canceled, dtype: float64
```



In [15]:
```python
plt.figure(figsize = (8,4))
axl = sns.countplot(x = 'hotel' , hue = 'is_canceled',data= df)
legend_labels = axl.get_legend_handles_labels()
axl.legend(bbox_to_anchor = (1,1))
plt.xlabel('hotel')
plt.ylabel('number of reservations')
plt.legend(['not cancelled', 'cancelled'])
plt.show()
```



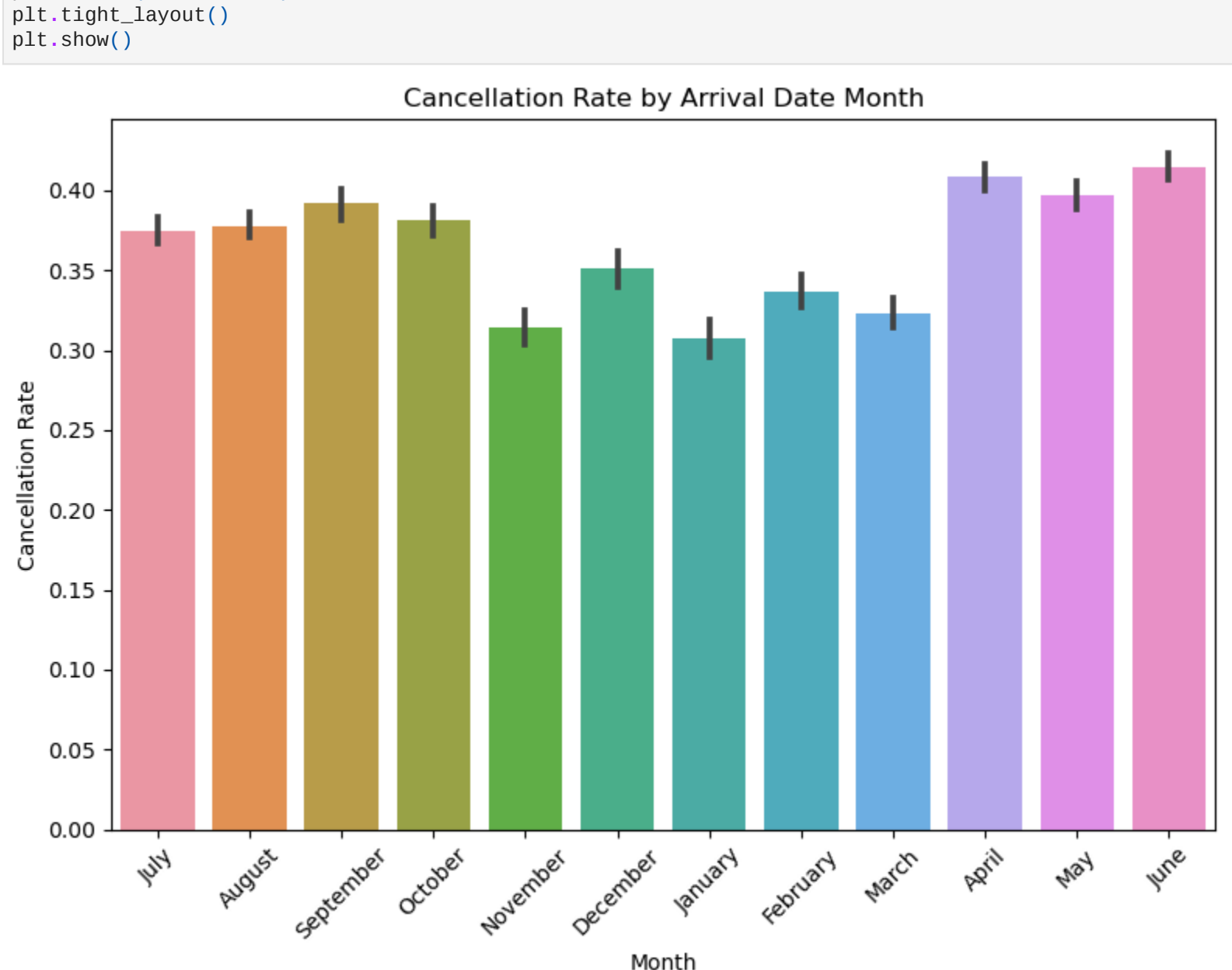
In [16]:
```python
resort_hotel = df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

Out[16]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

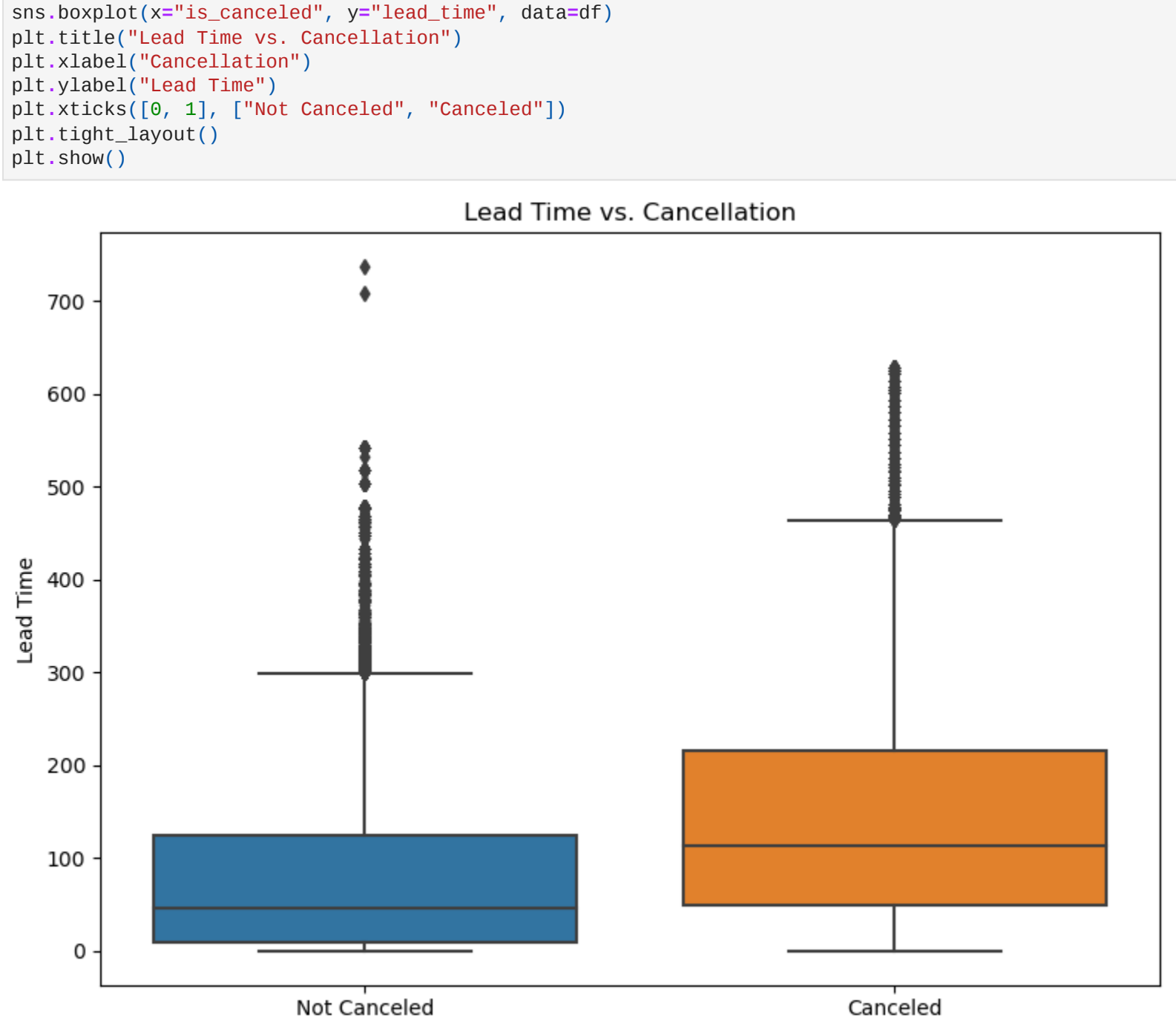
In [17]:
```python
city_hotel = df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

Out[17]:
```
0    0.582911
1    0.417089
Name: is_canceled, dtype: float64
```

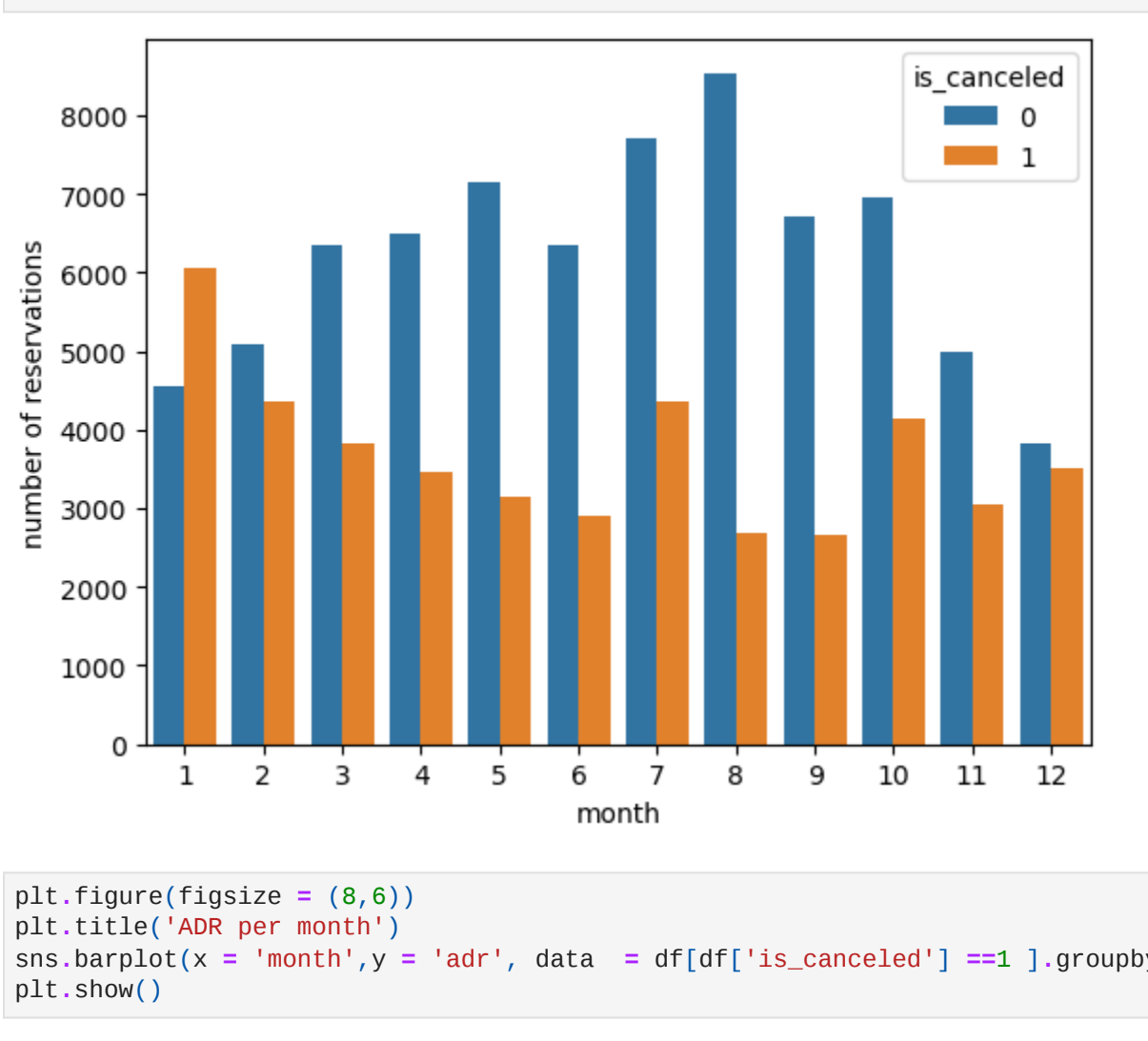
In [18]:
```python
# Bar plot: Cancellation rate by arrival date month
plt.figure(figsize=(8, 6))
sns.barplot(x="arrival_date_month", y="is_canceled", data=df)
plt.title("Cancellation Rate by Arrival Date Month")
plt.xlabel("Month")
plt.ylabel("Cancellation Rate")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



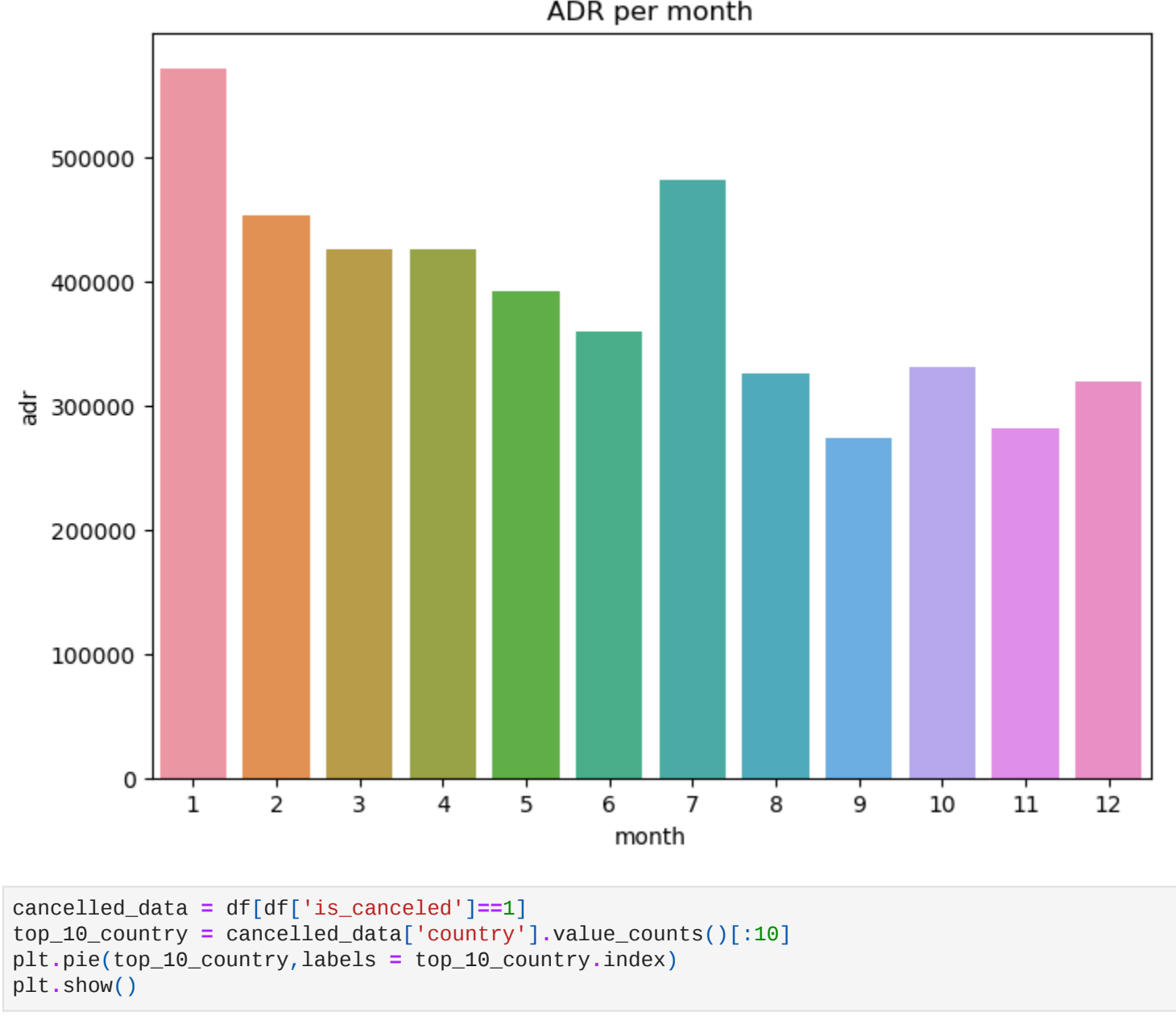
In [19]:
```python
# Box plot: Lead time vs. Cancellation
plt.figure(figsize=(8, 6))
sns.boxplot(x="is_canceled", y="lead_time", data=df)
plt.title("Lead Time vs. Cancellation")
plt.xlabel("Cancellation")
plt.ylabel("Lead Time")
plt.xticks([0, 1], ["Not Canceled", "Canceled"])
plt.tight_layout()
plt.show()
```



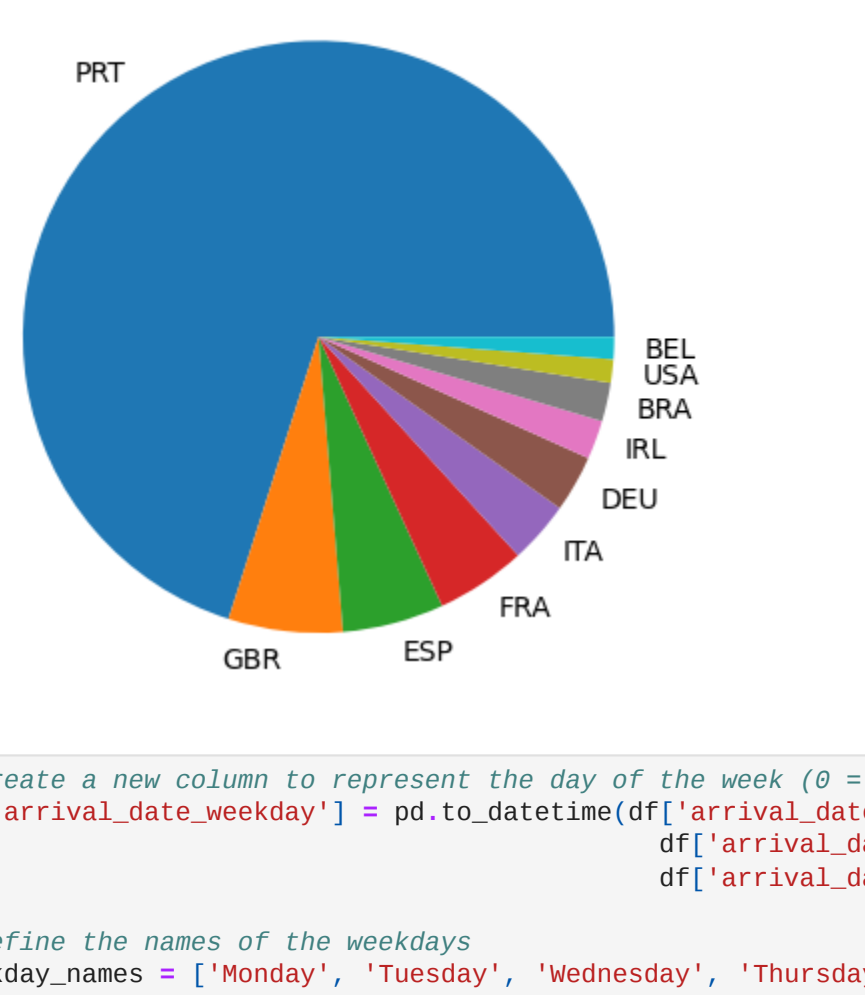
In [21]:
```python
df['month'] = df['reservation_status_date'].dt.month
axl = sns.countplot(x = 'month', hue = 'is_canceled',data = df)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.show()
```



In [32]:
```python
plt.figure(figsize = (8,6))
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 [24]:
```python
cancelled_data = df[df['is_canceled']==1]
top_10_country = cancelled_data['country'].value_counts()[:10]
plt.pie(top_10_country,labels = top_10_country.index)
plt.show()
```

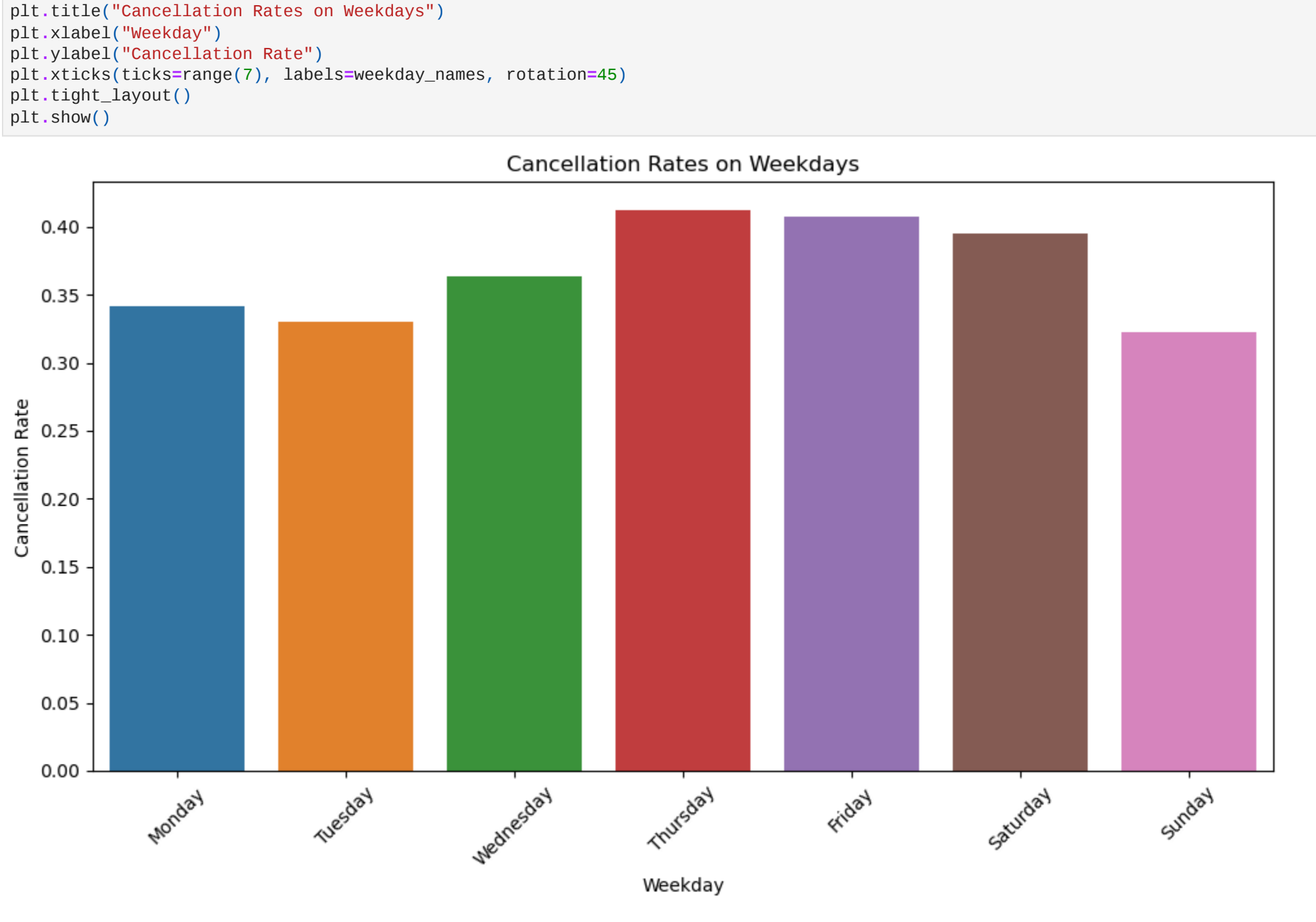


In [26]:
```python
# Create a new column to represent the day of the week (0 = Monday, 6 = Sunday)
df['arrival_date_weekday'] = pd.to_datetime(df['arrival_date_year'].astype(str) + '-' +
                                            df['arrival_date_month'] + '-' +
                                            df['arrival_date_day_of_month'].astype(str)).dt.dayofweek

# Define the names of the weekdays
weekday_names = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']

# Calculate cancellation rates for each day of the week
cancellation_rates = df.groupby('arrival_date_weekday')['is_canceled'].mean()

# Create a bar plot to visualize cancellation rates for weekdays
plt.figure(figsize=(10, 6))
sns.barplot(x=cancellation_rates.index, y=cancellation_rates.values)
plt.title("Cancellation Rates on Weekdays")
plt.xlabel("Weekday")
plt.ylabel("Cancellation Rate")
plt.xticks(ticks=range(7), labels=weekday_names, rotation=45)
plt.tight_layout()
plt.show()
```



In [27]:
```python
# Filter the data for repeated guests
repeated_guests = df[df['is_repeated_guest'] == 1]

# Calculate the total number of repeated guests
total_repeated_guests = len(repeated_guests)

# Calculate the number of canceled bookings among repeated guests
canceled_repeated_guests = repeated_guests[repeated_guests['is_canceled'] == 1]
total_canceled_repeated_guests = len(canceled_repeated_guests)

# Calculate the percentage of canceled bookings among repeated guests
percentage_canceled = (total_canceled_repeated_guests / total_repeated_guests) * 100

print(f"Percentage of repeated guests who canceled their bookings: {percentage_canceled:.2f}%")
```

```
Percentage of repeated guests who canceled their bookings: 14.50%
```

In [ ]: