Hotel Booking Data Analysis

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
```

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

In [3]:
```
df
```

Out[3]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays
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	
...	...	...	...	...	...	...	...	...
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	

119390 rows × 32 columns

In [4]:
```
df.head()
df.head(10)
```

Out[4]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_w
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	Resort Hotel	0	14	2015	July	27	1	
6	Resort Hotel	0	0	2015	July	27	1	
7	Resort Hotel	0	9	2015	July	27	1	
8	Resort Hotel	1	85	2015	July	27	1	
9	Resort Hotel	1	75	2015	July	27	1	

10 rows × 32 columns

In [5]:
```
df.tail()
df.tail(8)
```

Out[5]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_
119382	City Hotel	0	135	2017	August	35	30	
119383	City Hotel	0	164	2017	August	35	31	
119384	City Hotel	0	21	2017	August	35	30	
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	

8 rows × 32 columns

In [6]:
```
df.columns
```

Out[6]:
```
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 [7]:
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'], format='%d/%m/%Y', dayfirst=True
```

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

Out[8]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	de
count	119390	119390	119390	118902	119390	119390	119390	119390	
unique	2	12	5	177	8	5	10	12	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	
freq	79330	13877	92310	48590	56477	97870	85994	74053	

In [9]:
```
df.isnull()
```

Out[9]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	
119385	False	False	False	False	False	False	False	
119386	False	False	False	False	False	False	False	
119387	False	False	False	False	False	False	False	
119388	False	False	False	False	False	False	False	
119389	False	False	False	False	False	False	False	

119390 rows × 32 columns

In [10]:
```
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]:
```
print(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'],
      dtype='object')
```

In [ ]:
```
```

In [18]:
```
df.dropna(inplace = True)
```

In [19]:
```
df.isnull().sum()
```

Out[19]:
```
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
```

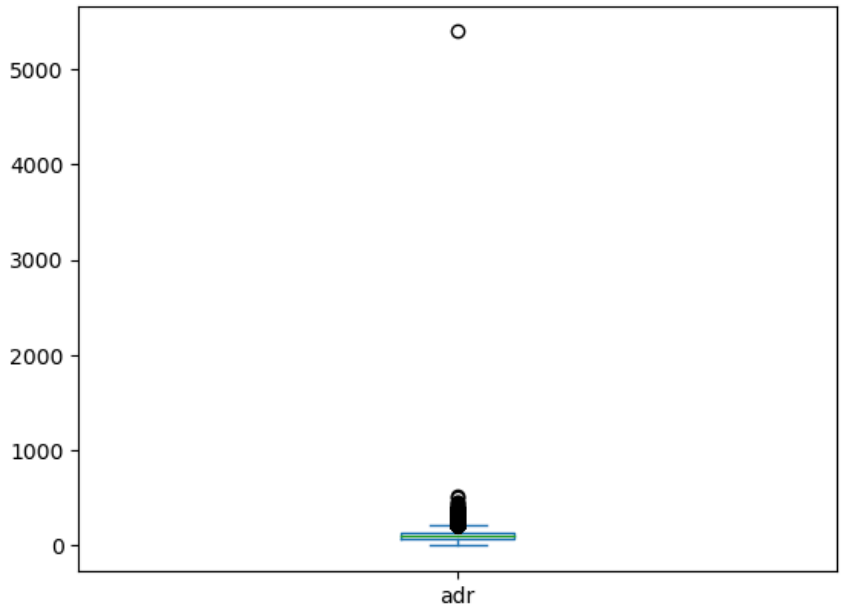
In [20]:
```
df.describe()
```

Out[20]:

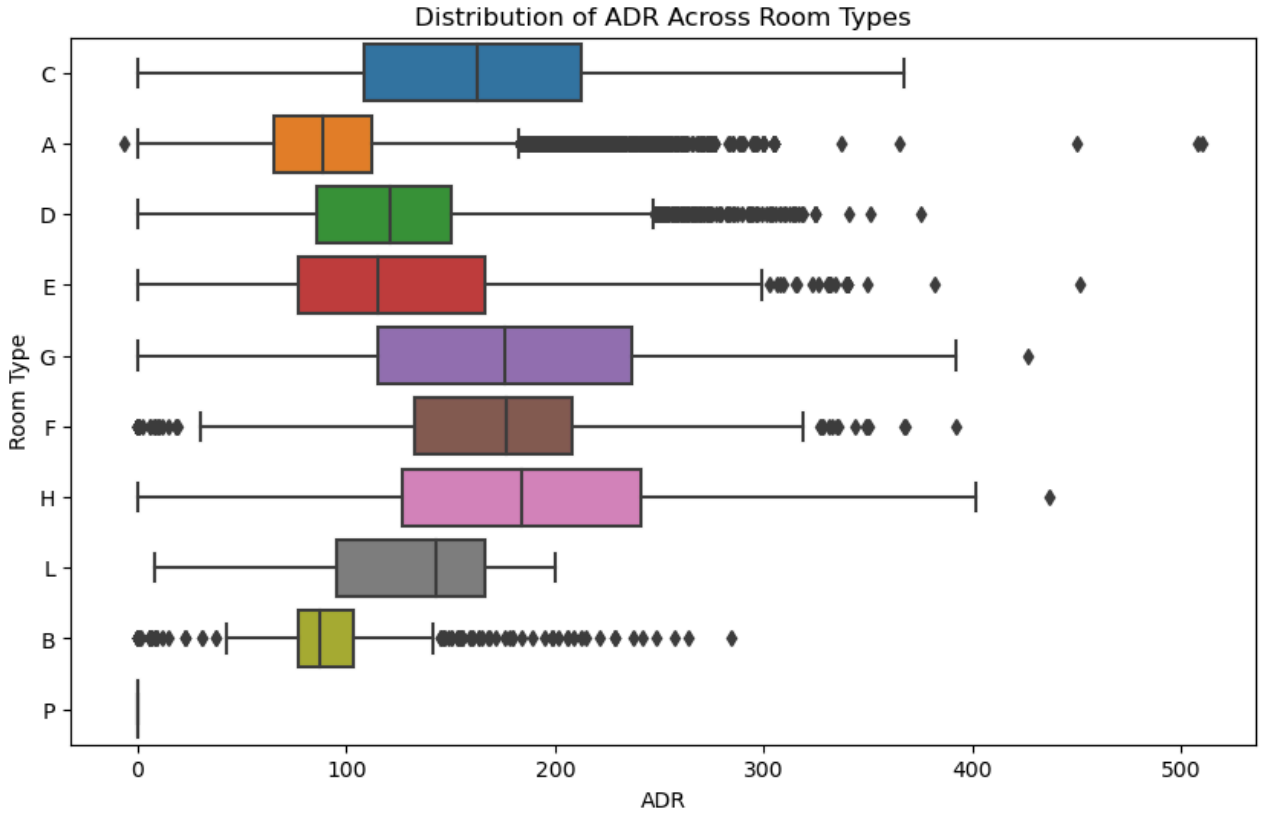
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	s
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	
mean	0.371352	104.311435	2016.157656	27.166555	15.800880	0.928897	
std	0.483168	106.903309	0.707459	13.589971	8.780324	0.996216	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	

In [21]:
```
df['adr'].plot(kind = 'box')
```

Out[21]: <Axes: >



In [39]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='adr', y='reserved_room_type')
plt.title('Distribution of ADR Across Room Types')
plt.xlabel('ADR')
plt.ylabel('Room Type')

plt.show()
```



In [22]:
```
df= df[df['adr']<5000]
```

In [23]:
```
df.describe()
```

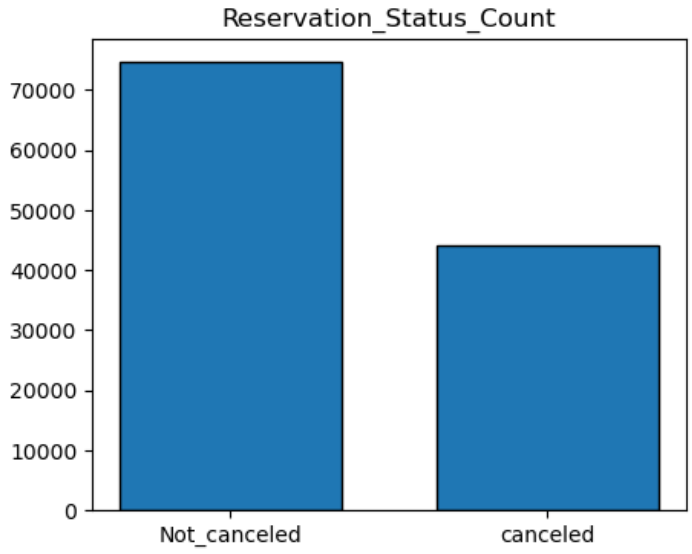
Out[23]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	s
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	
mean	0.371347	104.312018	2016.157657	27.166674	15.800802	0.928905	
std	0.483167	106.903570	0.707462	13.589966	8.780321	0.996217	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	

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

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

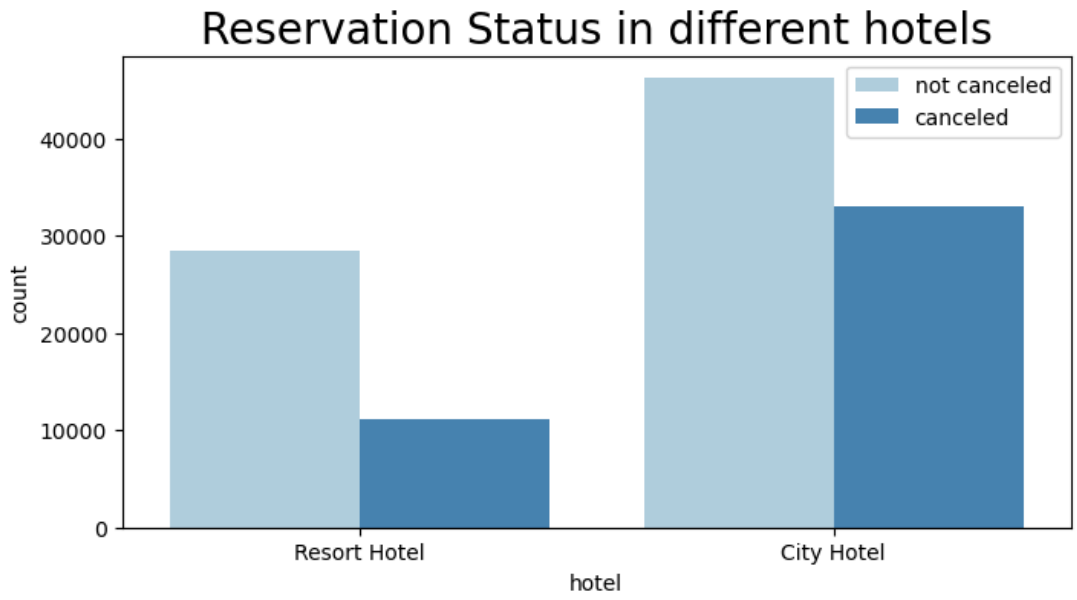
In [25]:
```
plt.figure(figsize = (5,4))
plt.title('Reservation_Status_Count')
plt.bar(['Not_canceled','canceled'],df['is_canceled'].value_counts(), edgecolor = 'k', width = 0.7)
plt.show()
```



In [26]:
```
df['is_canceled'] = df['is_canceled'].replace({0: 'not canceled', 1: 'canceled'})

# Plotting the count of reservation status in different hotels
plt.figure(figsize=(8, 4))
ax1 = sns.countplot(x='hotel', hue='is_canceled', data=df, palette='Blues')
ax1.legend(bbox_to_anchor=(1, 1))
legend_labels, _ = ax1.get_legend_handles_labels()
plt.title('Reservation Status in different hotels', size=20)
```

```
C:\Users\Powad\AppData\Local\Temp\ipykernel_13632\2156168252.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df['is_canceled'] = df['is_canceled'].replace({0: 'not canceled', 1: 'canceled'})
```

Out[26]:
```
Text(0.5, 1.0, 'Reservation Status in different hotels')
```



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

Out[27]:
```
not canceled    0.72025
canceled        0.27975
Name: is_canceled, dtype: float64
```

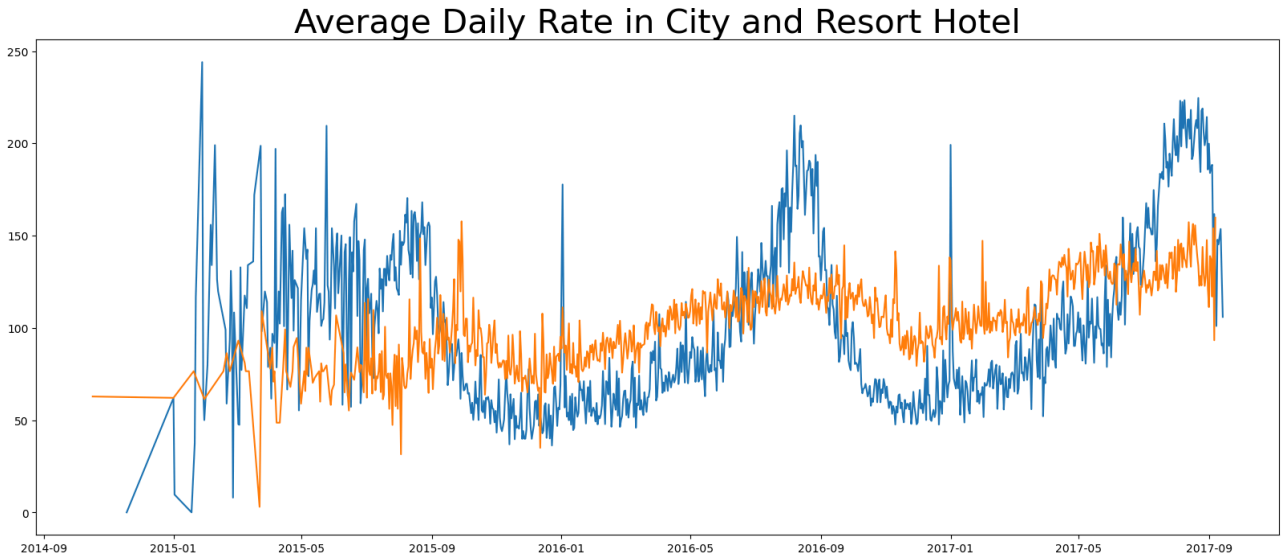
In [28]:
```
City_hotel = df[df['hotel'] == 'City Hotel']
City_hotel['is_canceled'].value_counts(normalize = True)
```

Out[28]:
```
not canceled    0.582918
canceled        0.417082
Name: is_canceled, dtype: float64
```

In [31]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()

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

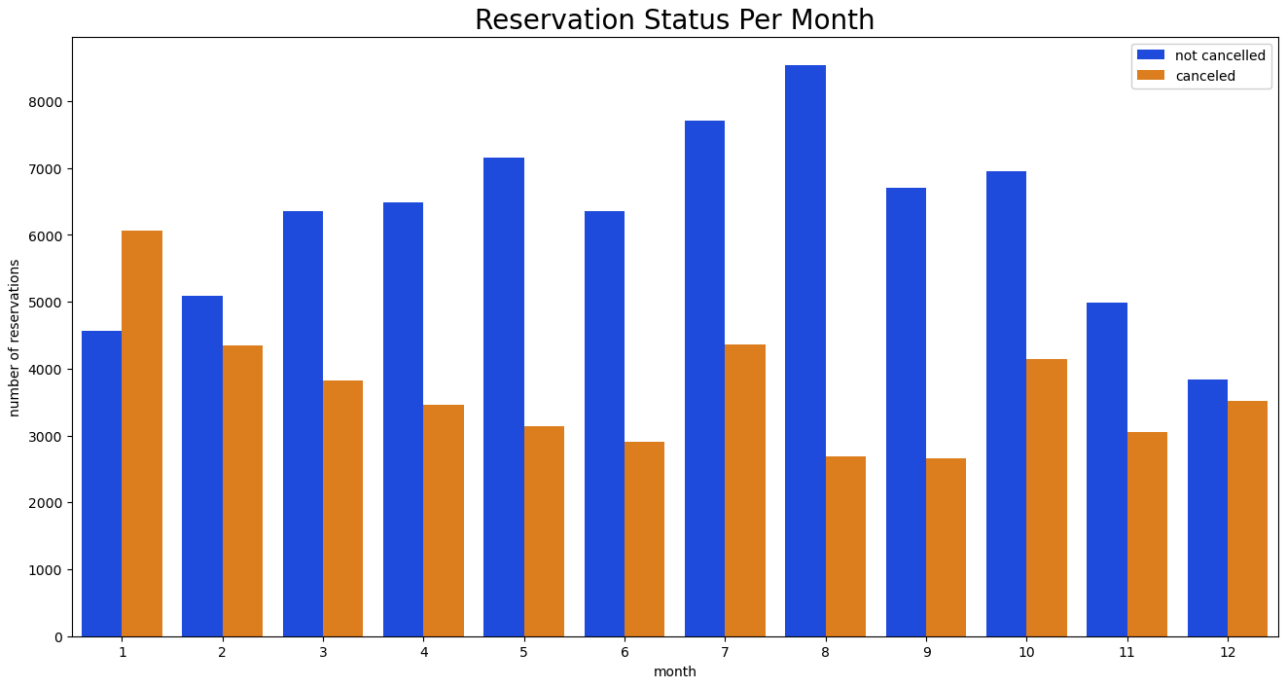
In [32]:
```
plt.figure(figsize=(20,8))

plt.title('Average Daily Rate in 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 = 'resort hotel')
```

Out[32]: [<matplotlib.lines.Line2D at 0x23805402a50>]



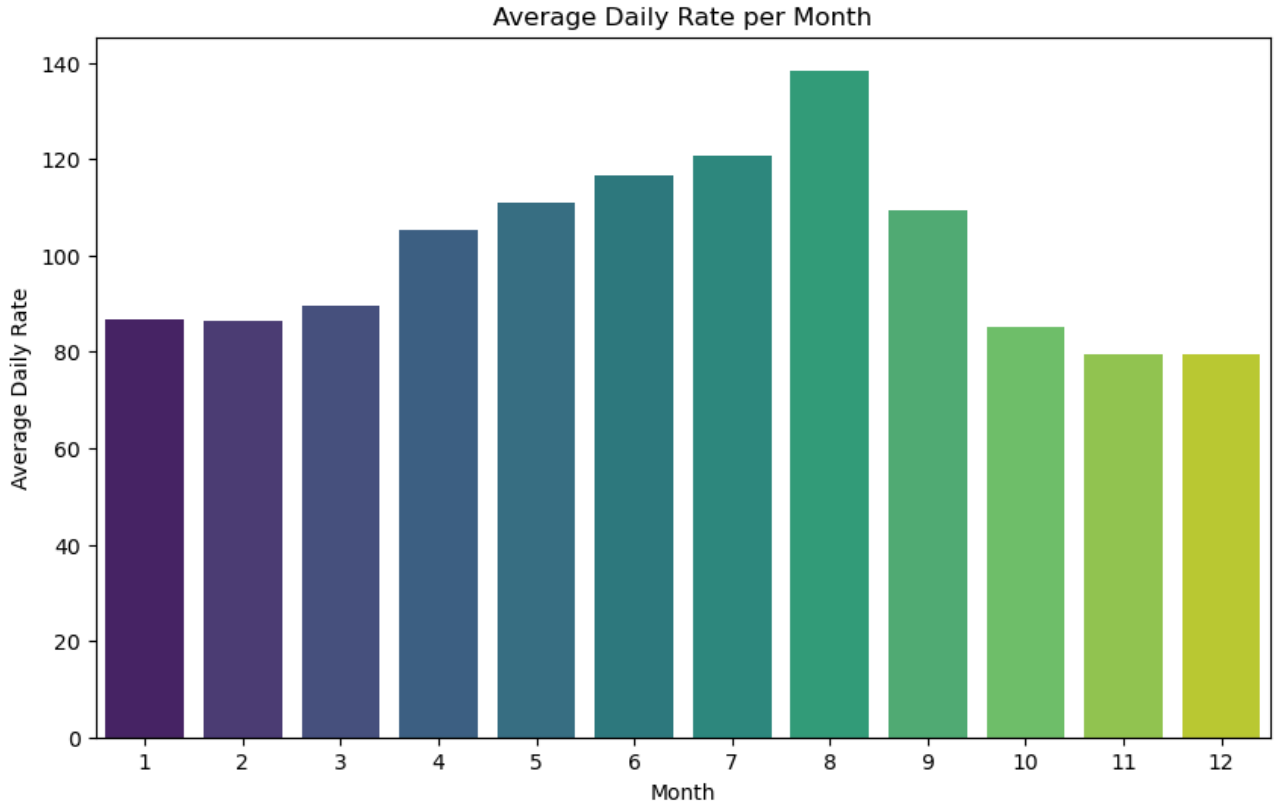
In [33]:
```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x= 'month',hue = 'is_canceled',data = df, palette= 'bright')
legend_labels,_ = ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor = (1,1))
plt.title('Reservation Status Per Month', size = 20)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.legend(['not cancelled','canceled'])
plt.show()
```

```
C:\Users\Powad\AppData\Local\Temp\ipykernel_13632\3476278942.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df['month'] = df['reservation_status_date'].dt.month
```



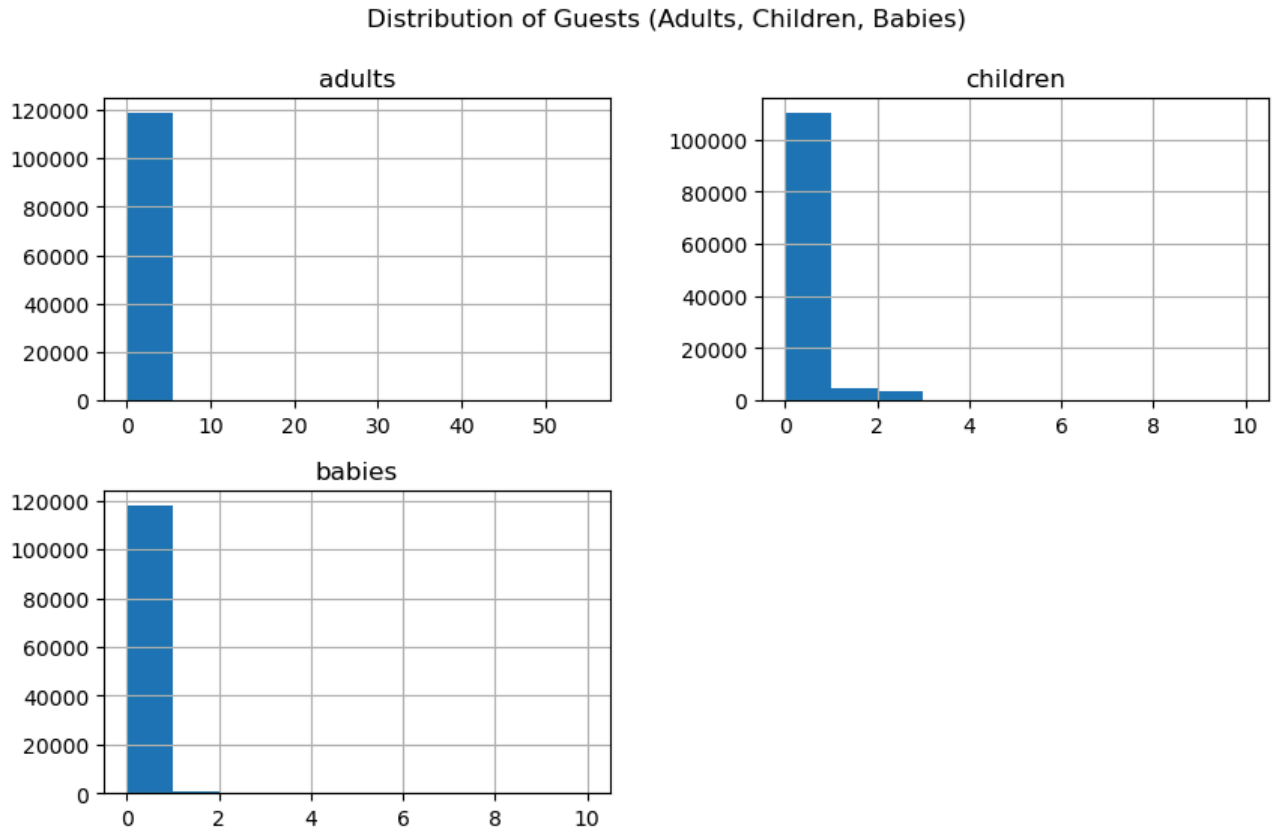
In [34]:

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

# Calculate average daily rate for each month
average_daily_rate = df.groupby('month')['adr'].mean().reset_index()

# Plotting average daily rate for each month with Seaborn
plt.figure(figsize=(10, 6))
sns.barplot(x='month', y='adr', data=average_daily_rate, palette='viridis')
plt.xlabel('Month')
plt.ylabel('Average Daily Rate')
plt.title('Average Daily Rate per Month')
plt.show()
```

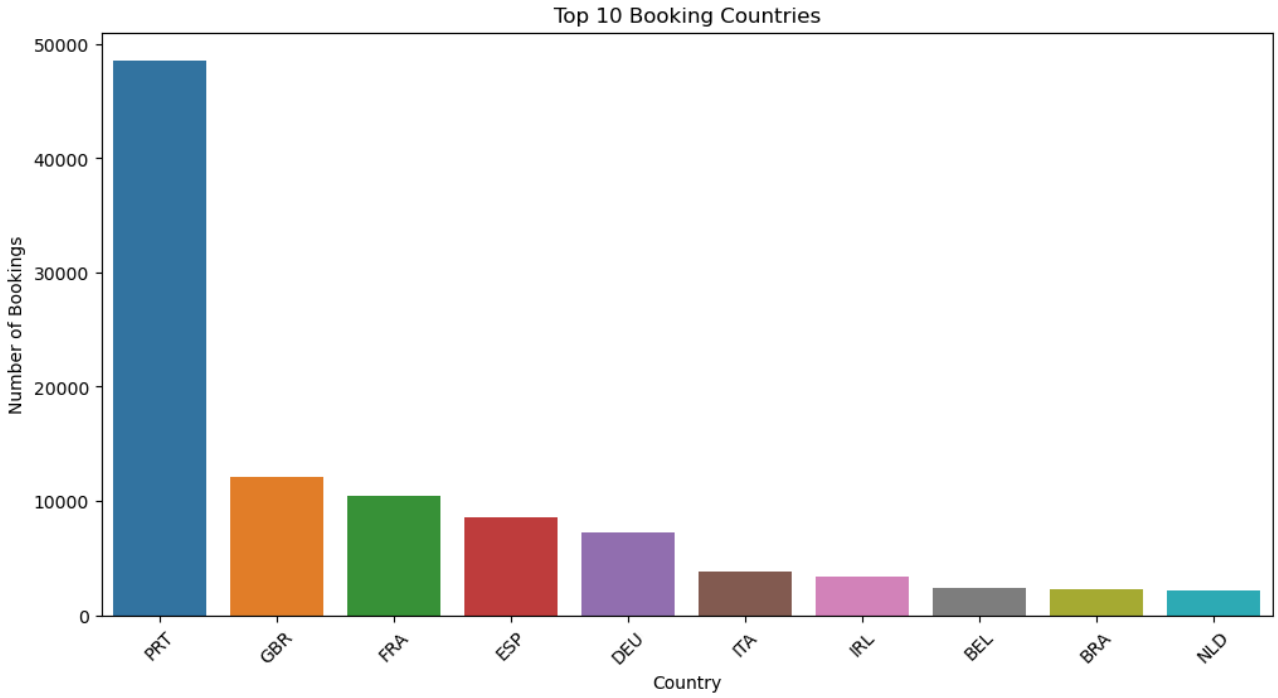


###Distribution of Adults, Children, and Babies

In [35]:
```
df[['adults', 'children', 'babies']].hist(bins=10, figsize=(10, 6))
plt.suptitle('Distribution of Guests (Adults, Children, Babies)')
plt.xlabel('Number of Guests')
plt.ylabel('Frequency')
plt.show()
```



In [36]:
```
top_countries = df['country'].value_counts().head(10)
plt.figure(figsize=(12, 6))
sns.barplot(x=top_countries.index, y=top_countries.values)
plt.title('Top 10 Booking Countries')
plt.xlabel('Country')
plt.ylabel('Number of Bookings')
plt.xticks(rotation=45)

plt.show()
```



In [ ]:

In [ ]: