Importing Libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading the dataset

```
df=pd.read_csv("hotel_bookings 2.csv")
```

Exploratory Data Analysis and Data Cleaning

```
df.head()

          hotel  is_canceled  lead_time  arrival_date_year 
arrival_date_month  \
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   

   arrival_date_week_number  arrival_date_day_of_month  \
0                        27                          1   
1                        27                          1   
2                        27                          1   
3                        27                          1   
4                        27                          1   

   stays_in_weekend_nights  stays_in_week_nights  adults  ... 
deposit_type  \
0                        0                     0       2  ...    No 
Deposit   
1                        0                     0       2  ...    No 
Deposit   
2                        0                     1       1  ...    No 
Deposit   
3                        0                     1       1  ...    No 
Deposit   
```

```
4                        0                     2       2  ...    No
Deposit

   agent company days_in_waiting_list customer_type   adr  \
0    NaN     NaN                    0     Transient   0.0
1    NaN     NaN                    0     Transient   0.0
2    NaN     NaN                    0     Transient  75.0
3  304.0     NaN                    0     Transient  75.0
4  240.0     NaN                    0     Transient  98.0

   required_car_parking_spaces  total_of_special_requests
reservation_status  \
0                            0                          0
Check-Out
1                            0                          0
Check-Out
2                            0                          0
Check-Out
3                            0                          0
Check-Out
4                            0                          1
Check-Out

  reservation_status_date
0                1/7/2015
1                1/7/2015
2                2/7/2015
3                2/7/2015
4                3/7/2015

[5 rows x 32 columns]
```

```
df.tail()
```

```
             hotel  is_canceled  lead_time  arrival_date_year  \
119385  City Hotel            0         23               2017
119386  City Hotel            0        102               2017
119387  City Hotel            0         34               2017
119388  City Hotel            0        109               2017
119389  City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
119385                         30                        2
```

```
119386                         31                        2
119387                         31                        2
119388                         31                        2
119389                         29                        2

        stays_in_week_nights  adults  ...  deposit_type  agent company  
\
119385                     5       2  ...    No Deposit  394.0     NaN  

119386                     5       3  ...    No Deposit    9.0     NaN  

119387                     5       2  ...    No Deposit    9.0     NaN  

119388                     5       2  ...    No Deposit   89.0     NaN  

119389                     7       2  ...    No Deposit    9.0     NaN  


       days_in_waiting_list customer_type     adr  \
119385                    0     Transient   96.14   
119386                    0     Transient  225.43   
119387                    0     Transient  157.71   
119388                    0     Transient  104.40   
119389                    0     Transient  151.20   

        required_car_parking_spaces  total_of_special_requests  \
119385                            0                          0   
119386                            0                          2   
119387                            0                          4   
119388                            0                          0   
119389                            0                          2   

        reservation_status reservation_status_date  
119385           Check-Out                6/9/2017  
119386           Check-Out                7/9/2017  
119387           Check-Out                7/9/2017  
119388           Check-Out                7/9/2017  
119389           Check-Out                7/9/2017  

[5 rows x 32 columns]
```

```
df.shape
```

```
(119390, 32)
```

```
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')
```

```
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
```

```python
df['reservation_status_date'] =
pd.to_datetime(df['reservation_status_date'], format='%d/%m/%Y')
df['reservation_status_date']
```

```
0        2015-07-01
1        2015-07-01
2        2015-07-02
3        2015-07-02
4        2015-07-03
            ...
119385   2017-09-06
119386   2017-09-07
119387   2017-09-07
119388   2017-09-07
119389   2017-09-07
Name: reservation_status_date, Length: 119390, dtype: datetime64[ns]
```

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

```
             hotel arrival_date_month    meal country
market_segment  \
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


       distribution_channel reserved_room_type assigned_room_type  \
count                119390             119390             119390
unique                    5                 10                 12
top                   TA/TO                  A                  A
freq                  97870              85994              74053

       deposit_type customer_type reservation_status
count        119390        119390             119390
unique            3             4                  3
top      No Deposit     Transient          Check-Out
freq         104641         89613              75166
```

```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']
```

```
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                               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
```

```
df.drop(['company','agent'],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
dtype: int64
```

```
df.describe()
```

```
         is_canceled      lead_time  arrival_date_year  \
count  118898.000000  118898.000000      118898.000000   
mean        0.371352     104.311435        2016.157656   
min         0.000000       0.000000        2015.000000   
25%         0.000000      18.000000        2016.000000   
50%         0.000000      69.000000        2016.000000   
75%         1.000000     161.000000        2017.000000   
max         1.000000     737.000000        2017.000000   
std         0.483168     106.903309           0.707459   

       arrival_date_week_number  arrival_date_day_of_month  \
count             118898.000000              118898.000000   
mean                  27.166555                  15.800880   
min                    1.000000                   1.000000   
25%                   16.000000                   8.000000   
50%                   28.000000                  16.000000   
75%                   38.000000                  23.000000   
max                   53.000000                  31.000000   
std                   13.589971                   8.780324   

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            118898.000000         118898.000000  118898.000000   
```

```
mean                  0.928897              2.502145       1.858391
min                   0.000000              0.000000       0.000000
25%                   0.000000              1.000000       2.000000
50%                   1.000000              2.000000       2.000000
75%                   2.000000              3.000000       2.000000
max                  16.000000             41.000000      55.000000
std                   0.996216              1.900168       0.578576

            children         babies  is_repeated_guest  \
count  118898.000000  118898.000000      118898.000000   
mean        0.104207       0.007948           0.032011   
min         0.000000       0.000000           0.000000   
25%         0.000000       0.000000           0.000000   
50%         0.000000       0.000000           0.000000   
75%         0.000000       0.000000           0.000000   
max        10.000000      10.000000           1.000000   
std         0.399172       0.097380           0.176029   

       previous_cancellations  previous_bookings_not_canceled  \
count           118898.000000                   118898.000000   
mean                 0.087142                        0.131634   
min                  0.000000                        0.000000   
25%                  0.000000                        0.000000   
50%                  0.000000                        0.000000   
75%                  0.000000                        0.000000   
max                 26.000000                       72.000000   
std                  0.845869                        1.484672   

       booking_changes  days_in_waiting_list            adr  \
count    118898.000000         118898.000000  118898.000000   
mean          0.221181              2.330754     102.003243   
min           0.000000              0.000000      -6.380000   
25%           0.000000              0.000000      70.000000   
50%           0.000000              0.000000      95.000000   
75%           0.000000              0.000000     126.000000   
max          21.000000            391.000000    5400.000000   
std           0.652785             17.630452      50.485862   

       required_car_parking_spaces  total_of_special_requests  \
count                118898.000000              118898.000000   
mean                      0.061885                   0.571683   
min                       0.000000                   0.000000   
25%                       0.000000                   0.000000   
50%                       0.000000                   0.000000   
75%                       0.000000                   1.000000   
max                       8.000000                   5.000000   
std                       0.244172                   0.792678   

             reservation_status_date  
count                         118898  
```

```
mean   2016-07-30 07:37:53.336809984
min              2014-10-17 00:00:00
25%              2016-02-02 00:00:00
50%              2016-08-08 00:00:00
75%              2017-02-09 00:00:00
max              2017-09-14 00:00:00
std                              NaN
```

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

```
df
```

```
               hotel  is_canceled  lead_time  arrival_date_year  \
0       Resort Hotel            0        342               2015
1       Resort Hotel            0        737               2015
2       Resort Hotel            0          7               2015
3       Resort Hotel            0         13               2015
4       Resort Hotel            0         14               2015
...              ...          ...        ...                ...
119385    City Hotel            0         23               2017
119386    City Hotel            0        102               2017
119387    City Hotel            0         34               2017
119388    City Hotel            0        109               2017
119389    City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
...                   ...                       ...
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                        0
1                               1                        0
2                               1                        0
3                               1                        0
4                               1                        0
...                           ...                      ...
119385                         30                        2
119386                         31                        2
119387                         31                        2
119388                         31                        2
119389                         29                        2
```

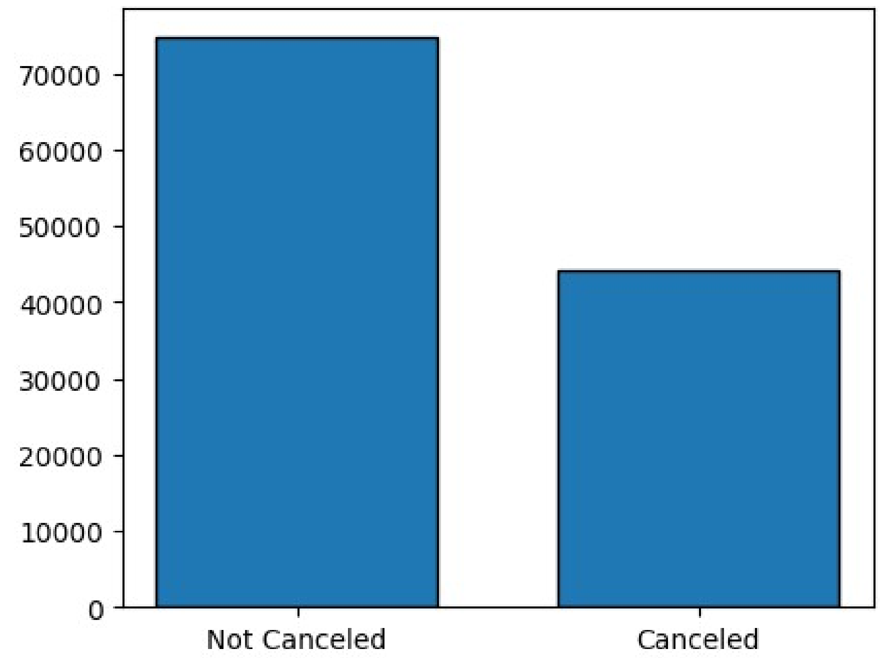
```
        stays_in_week_nights  adults  ...  assigned_room_type  \
0                          0       2  ...                   C
1                          0       2  ...                   C
2                          1       1  ...                   C
3                          1       1  ...                   A
4                          2       2  ...                   A
...                      ...     ...  ...                 ...
119385                     5       2  ...                   A
119386                     5       3  ...                   E
119387                     5       2  ...                   D
119388                     5       2  ...                   A
119389                     7       2  ...                   A

        booking_changes deposit_type days_in_waiting_list
customer_type  \
0                     3   No Deposit                    0
Transient
1                     4   No Deposit                    0
Transient
2                     0   No Deposit                    0
Transient
3                     0   No Deposit                    0
Transient
4                     0   No Deposit                    0
Transient
...                 ...          ...                  ...           ..
.
119385                0   No Deposit                    0
Transient
119386                0   No Deposit                    0
Transient
119387                0   No Deposit                    0
Transient
119388                0   No Deposit                    0
Transient
119389                0   No Deposit                    0
Transient

           adr  required_car_parking_spaces  total_of_special_requests
\
0         0.00                            0                          0

1         0.00                            0                          0

2        75.00                            0                          0

3        75.00                            0                          0

4        98.00                            0                          1
```

```
...        ...                          ...                        ...
119385   96.14                            0                          0
119386  225.43                            0                          2
119387  157.71                            0                          4
119388  104.40                            0                          0
119389  151.20                            0                          2

        reservation_status reservation_status_date
0                Check-Out              2015-07-01
1                Check-Out              2015-07-01
2                Check-Out              2015-07-02
3                Check-Out              2015-07-02
4                Check-Out              2015-07-03
...                    ...                     ...
119385           Check-Out              2017-09-06
119386           Check-Out              2017-09-07
119387           Check-Out              2017-09-07
119388           Check-Out              2017-09-07
119389           Check-Out              2017-09-07

[118897 rows x 30 columns]
```

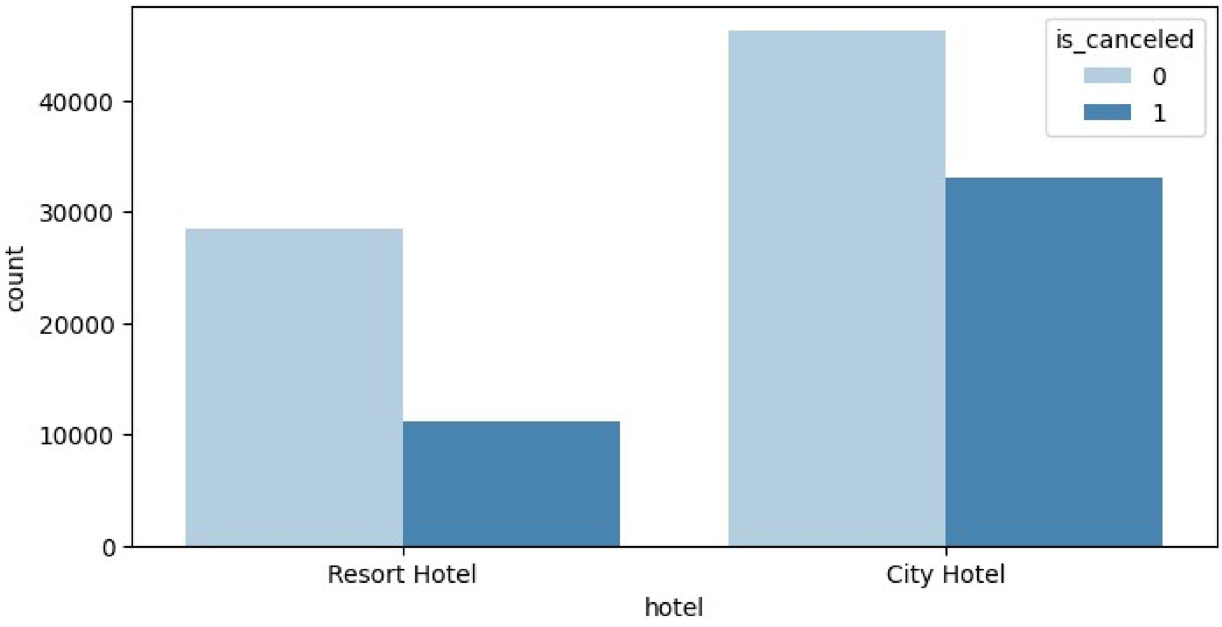
Data Analysis and Visualizations

```
cancelled_perc=df['is_canceled'].value_counts(normalize=True)
print(cancelled_perc)
plt.figure(figsize=(5,4))
plt.bar(['Not Canceled','Canceled'],df['is_canceled'].value_counts(),edgecolor='k',width=0.7)
plt.show()
```

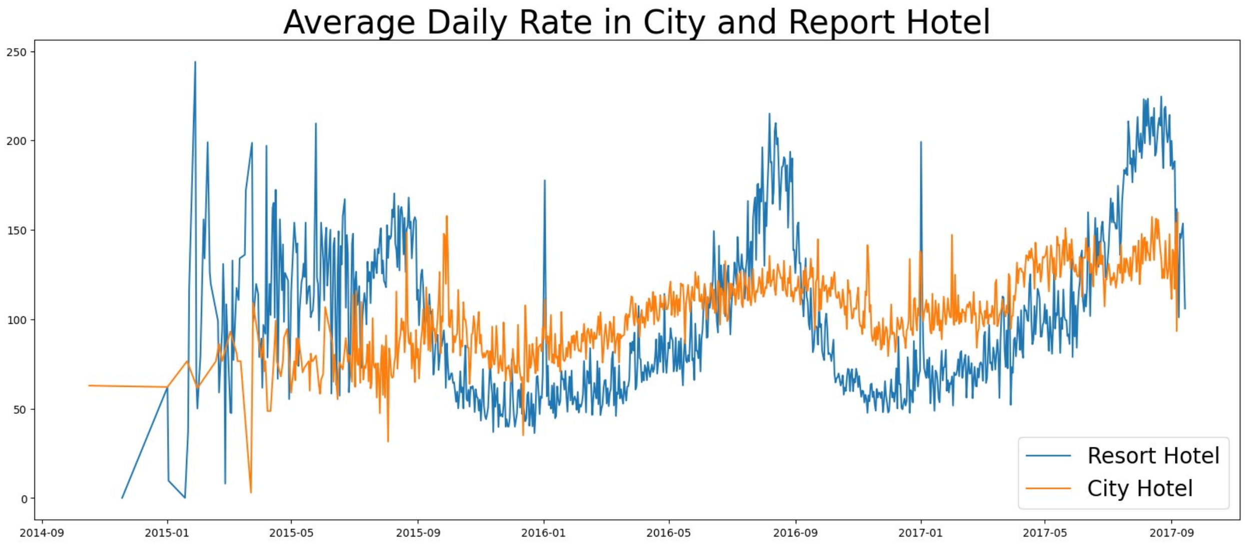
```
is_canceled
0    0.628653
1    0.371347
Name: proportion, dtype: float64
```



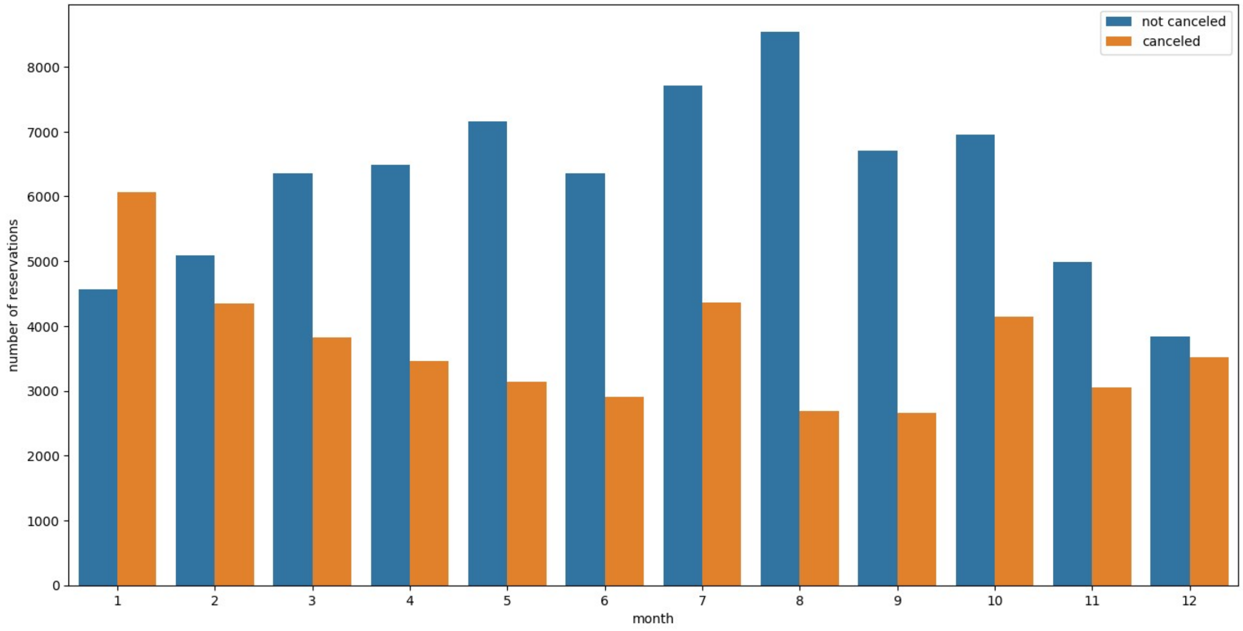
```
plt.figure(figsize=(8,4))
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,palette='Blues')
legend_labels,_=ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor(1,1))
plt.title('Reservation status in different hotels',size=20)
plt.xlabel('hotel')
plt.ylabel('number of reservations')

----------------------------------------------------------------------
-----
NameError                                 Traceback (most recent call
last)
Cell In[29], line 4
      2
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,palette='Blues')
      3 legend_labels,_=ax1.get_legend_handles_labels()
----> 4 ax1.legend(bbox_to_anchor(1,1))
      5 plt.title('Reservation status in different hotels',size=20)
      6 plt.xlabel('hotel')

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



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

is_canceled
0    0.72025
1    0.27975
Name: proportion, dtype: float64

city_hotel=df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)

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

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

plt.figure(figsize=(20,8))
plt.title('Average Daily Rate in City and Report 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()
```

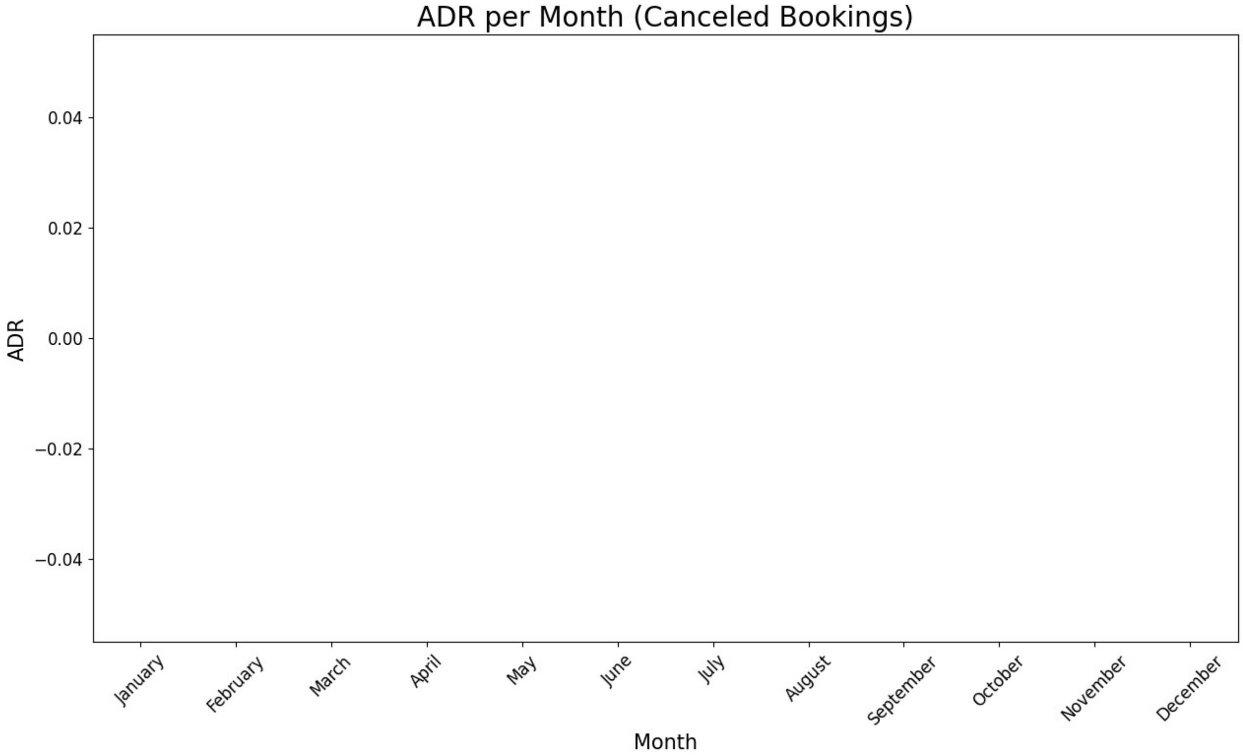


```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,8))
ax1=sns.countplot(x='month',hue='is_canceled',data=df)
plt.ylabel('number of reservations')
plt.legend(['not canceled','canceled'])
plt.show()
```



```
# Group and filter the data
filtered_data = df[df['is_canceled'] == 1].groupby('month')
[['adr']].sum().reset_index()
```

```python
# Plot the data
plt.figure(figsize=(15, 8))
plt.title('ADR per Month (Canceled Bookings)', fontsize=20)
sns.barplot(x='month', y='adr', data=filtered_data, palette='viridis')

# Enhance visualization
plt.xlabel('Month', fontsize=15)
plt.ylabel('ADR', fontsize=15)
plt.xticks(rotation=45, fontsize=12)
plt.yticks(fontsize=12)
plt.show()
```

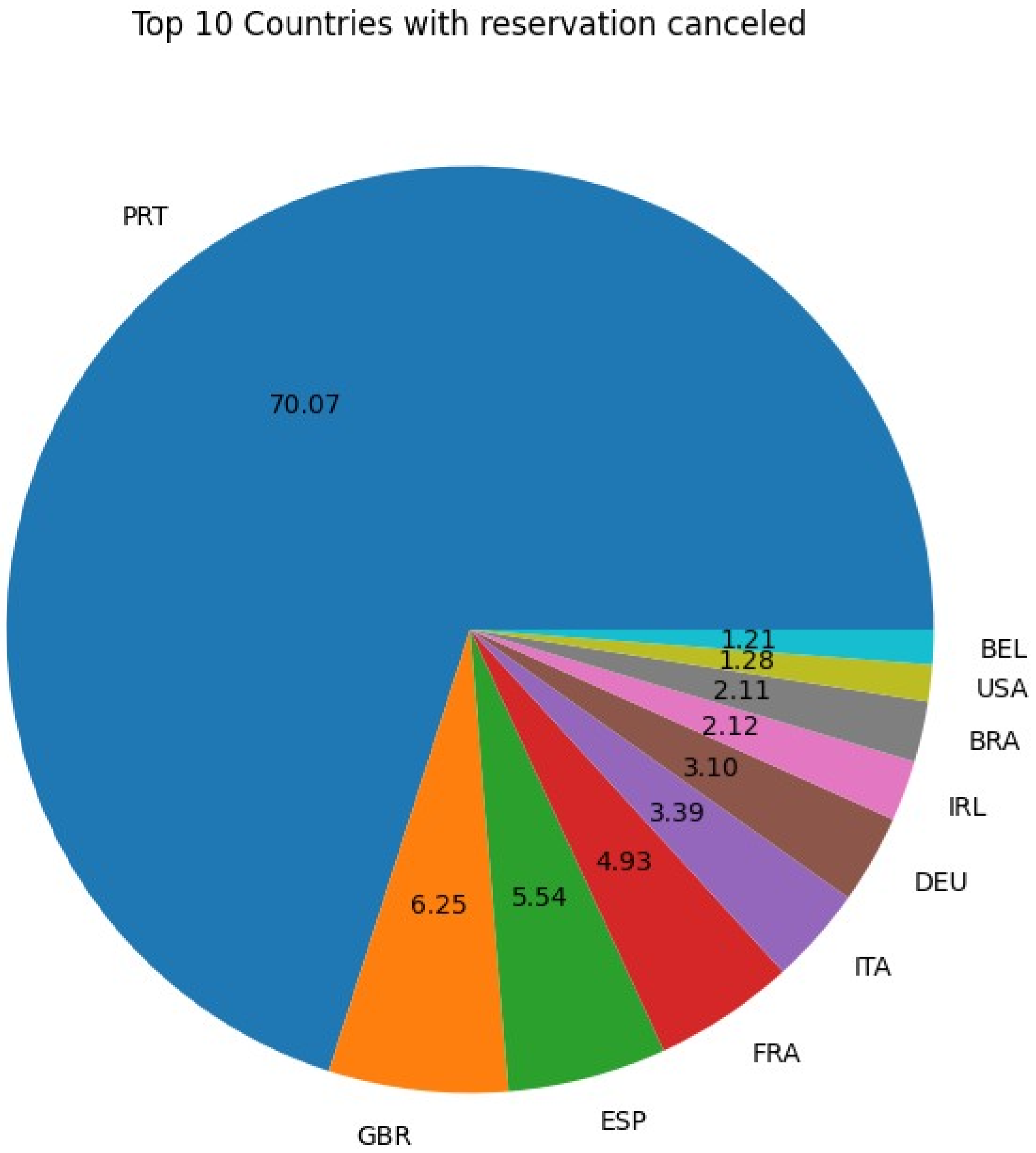


```python
print(df[df['is_canceled'] == 1].groupby('month')
[['adr']].sum().reset_index())
```

```
       month  adr
0    January  0.0
1   February  0.0
2      March  0.0
3      April  0.0
4        May  0.0
5       June  0.0
6       July  0.0
7     August  0.0
8  September  0.0
9    October  0.0
```

```
10   November  0.0
11   December  0.0
```

```
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()
```

```
market_segment
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: count, dtype: int64
```

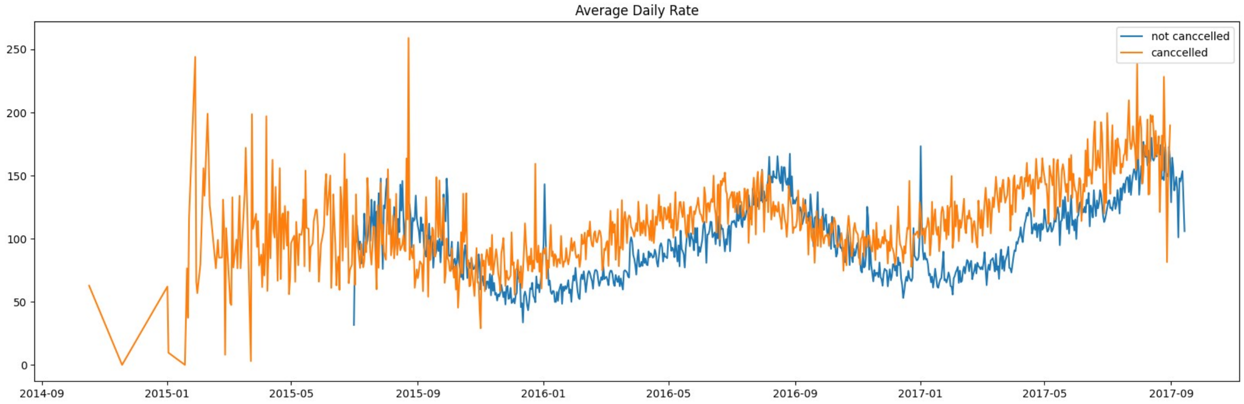
```
df['market_segment'].value_counts(normalize=True)
```

```
market_segment
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: proportion, dtype: float64
```

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

```
market_segment
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: proportion, 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')
plt.plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr['adr'],label='not canccelled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr[
'adr'],label='canccelled')
plt.legend()
plt.show()
```


Average Daily Rate
not canccelled
canccelled