Hotel Bookings Data Analysis

The dataset of hotel booking is picked from kaggle.com. The dataset contains data of customers from Jan 2015 to Dec 2017 and having 119390 rows and 36 columns.

In this project, we will be analysing various factors which are responsible for cancellations of booked hotels.

Importing Libraries

In [10]:

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

Loading Dataset

In [12]:

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

Exploratory Data Analysis and Data Cleaning

In [14]:

```
df.head(5)
```

Out[14]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_numbe
0	Resort Hotel	0	342	2015	July	2
1	Resort Hotel	0	737	2015	July	2
2	Resort Hotel	0	7	2015	July	2
3	Resort Hotel	0	13	2015	July	2
4	Resort Hotel	0	14	2015	July	2

5 rows × 36 columns

In [15]:

```
df.tail(5)
```

Out[15]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_nu
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 × 36 columns

In [16]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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 
 32  name                            119390 non-null  object 
 33  email                           119390 non-null  object 
 34  phone-number                    119390 non-null  object 
 35  credit_card                     119390 non-null  object 
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

In [17]:

```
df.shape
```

Out[17]:

```
(119390, 36)
```

In [18]:

```
df.drop(columns = ['name', 'phone-number', 'email', 'credit_card'], inplace
```

In [19]:

```
df.shape
```

Out[19]:

```
(119390, 32)
```

In [20]:

```
df.info()
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```
<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 [21]: print(df['reservation_status_date'].dtypes)
```

```
object
```

```
In [22]: df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date']
```

```
In [23]: print(df['reservation_status_date'].dtypes)
```

```
datetime64[ns]
```

```
In [24]: df.describe()
```

Out[24]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_d
count	119390.000000	119390.000000	119390.000000	119390.000000	1
mean	0.370416	104.011416	2016.156554	27.165173	
std	0.482918	106.863097	0.707476	13.605138	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	160.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

In [25]:

```
df.describe(include = "object").columns
```

Out[25]:

```
Index(['hotel', 'arrival_date_month', 'meal', 'country', 'market_segment',
       'distribution_channel', 'reserved_room_type', 'assigned_room_type',
       'deposit_type', 'customer_type', 'reservation_status'],
      dtype='object')
```

In [26]:

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

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```
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 [27]:

```
df.isna().sum()
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[27]:

```
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 [28]:

```
df.drop(['agent', 'company'], axis = 1, inplace = True)
```

In [29]:

```
df.dropna(inplace = True)
```

In [30]:

```
df.isna().sum()
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[30]:

```
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 [31]:

```
df.describe()
```

Out[31]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_d
count	118898.000000	118898.000000	118898.000000	118898.000000	1
mean	0.371352	104.311435	2016.157656	27.166555	
std	0.483168	106.903309	0.707459	13.589971	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	161.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

In [32]:

```
df['adr'].value_counts()
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```
Out[32]: 62.00     3753
         75.00     2710
         90.00     2471
         65.00     2397
         0.00      1938
                   ...
         96.09        1
         48.03        1
         89.43        1
         63.07        1
         157.71       1
         Name: adr, Length: 8870, dtype: int64
```

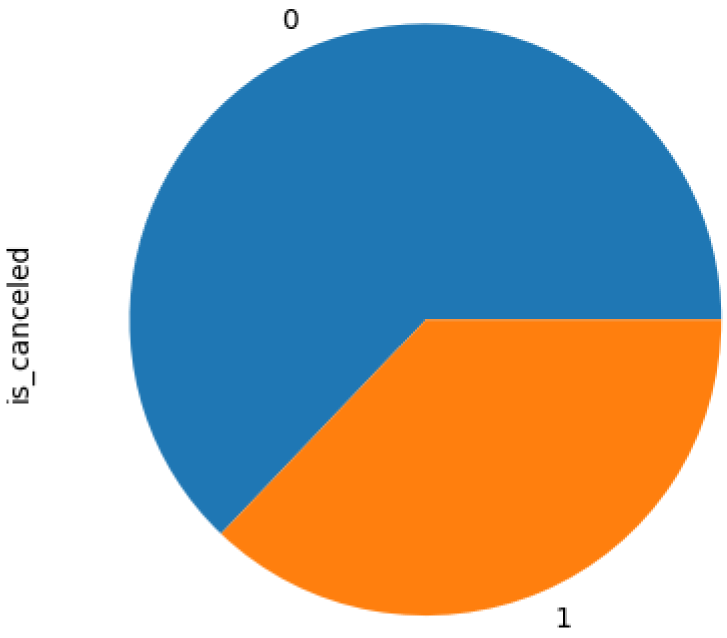
Data Analysis

```
In [37]: cancel_perc = df['is_canceled'].value_counts(normalize = True)
         cancel_perc
```

```
Out[37]: 0    0.628648
         1    0.371352
         Name: is_canceled, dtype: float64
```

```
In [39]: cancel_perc.plot(kind="pie")
```

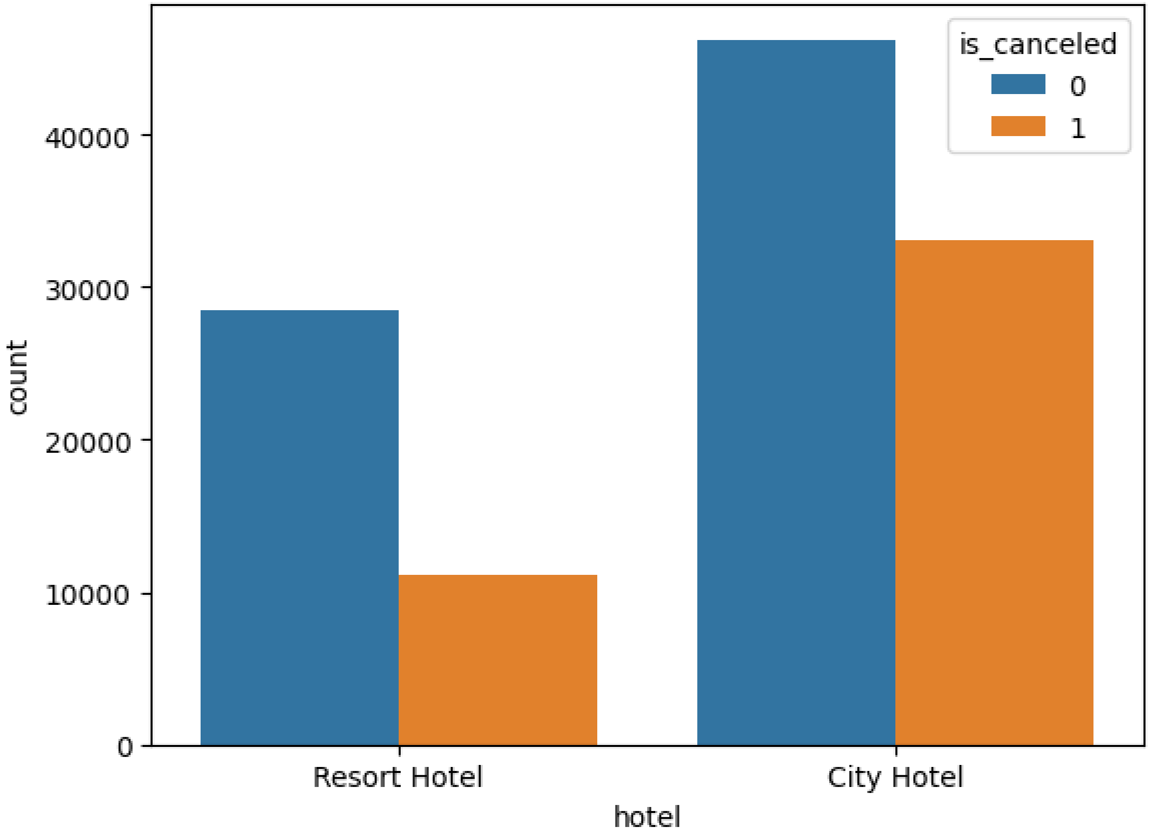
```
Out[39]: <AxesSubplot: ylabel='is_canceled'>
```



```
In [42]: cancel_plot = sns.countplot(x='hotel', hue = "is_canceled",  data = df)
         cancel_plot
```

```
Out[42]: <AxesSubplot: xlabel='hotel', ylabel='count'>
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



```
In [43]: resort_hotel = df[df['hotel'] == 'Resort Hotel']
         resort_hotel['is_canceled'].value_counts(normalize = True)
```

```
Out[43]: 0    0.72025
         1    0.27975
         Name: is_canceled, dtype: float64
```

```
In [44]: city_hotel = df[df['hotel'] == 'City Hotel']
         city_hotel['is_canceled'].value_counts(normalize = True)
```

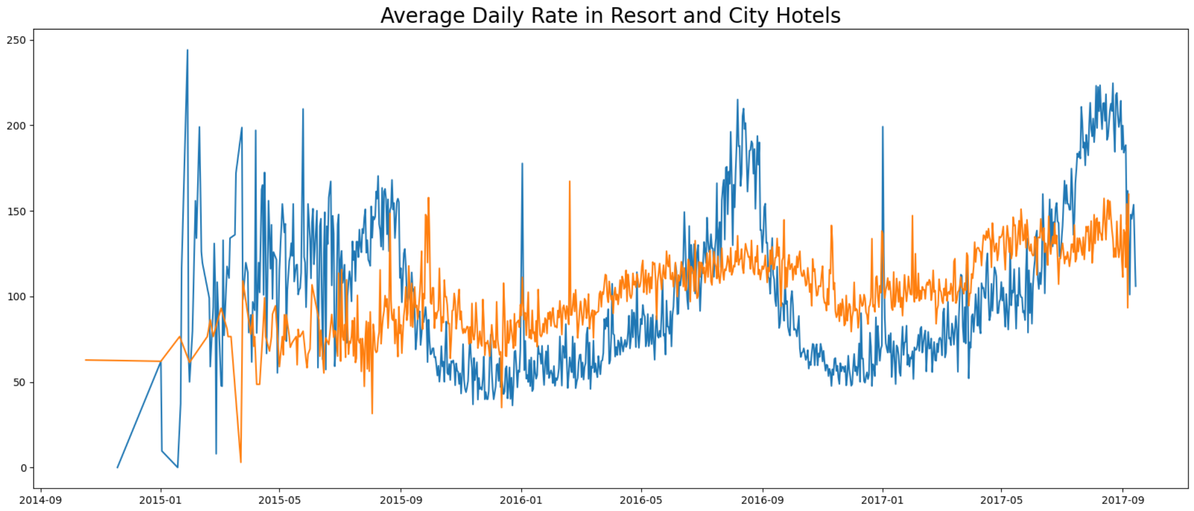
```
Out[44]: 0    0.582911
         1    0.417089
         Name: is_canceled, dtype: float64
```

```
In [46]: resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mear
         city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

```
In [52]: plt.figure(figsize = (20,8))
         plt.title("Average Daily Rate in Resort and City Hotels", fontsize = 20)
         plt.plot(resort_hotel.index, resort_hotel['adr'], label = "Resort Hotel")
         plt.plot(city_hotel.index, city_hotel['adr'], label = "City Hotel")
```

```
Out[52]: [<matplotlib.lines.Line2D at 0x1523e3b2110>]
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

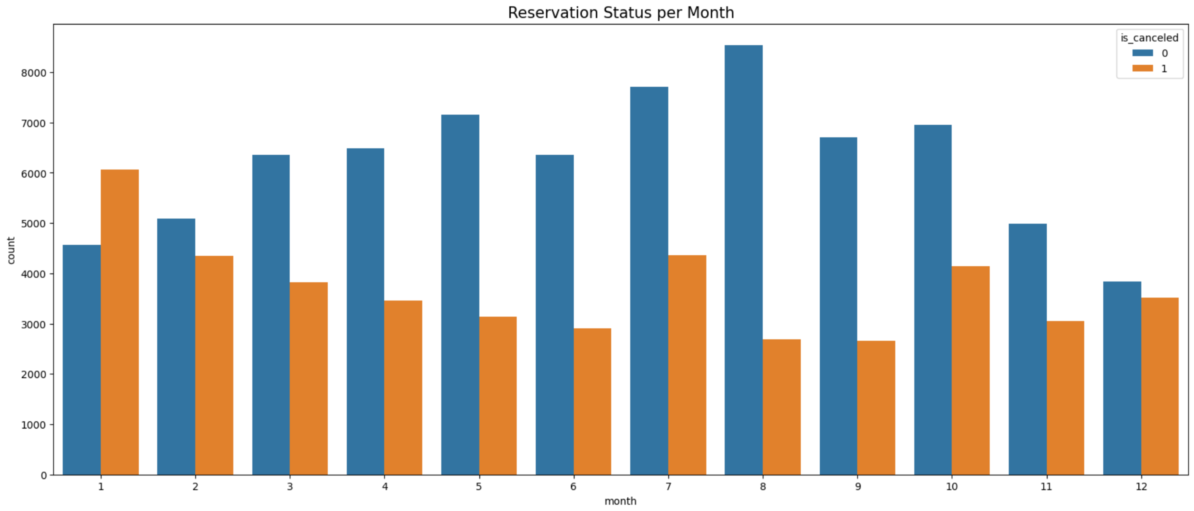


```
In [56]: df['month'] = df['reservation_status_date'].dt.month
         df['month'].value_counts()
```

```
Out[56]: 7     12074
         8     11223
         10    11095
         1     10622
         5     10294
         3     10177
         4      9957
         2      9436
         9      9359
         6      9255
         11     8052
         12     7354
         Name: month, dtype: int64
         <Figure size 2000x800 with 0 Axes>
```

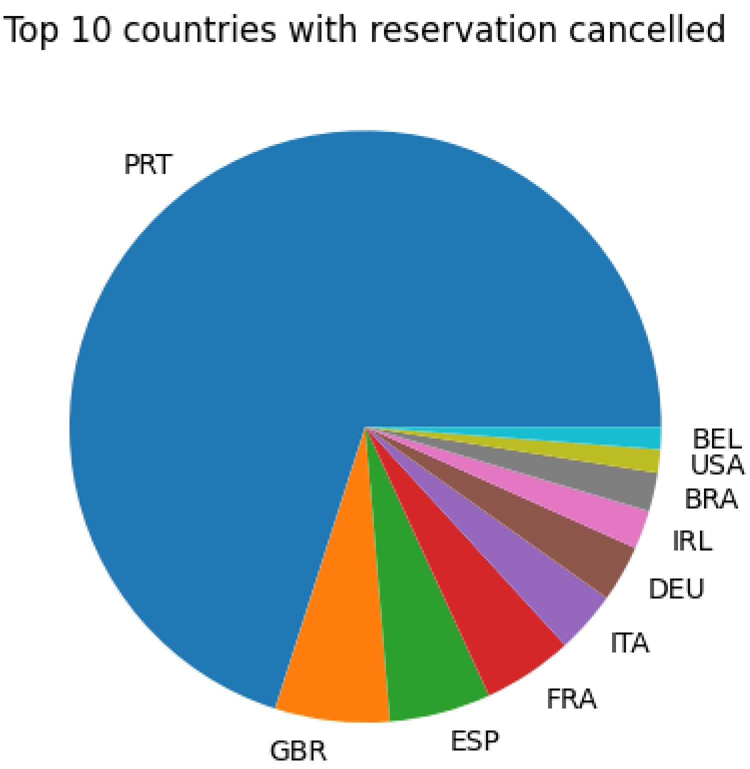
```
In [59]: plt.figure(figsize = (20, 8))
         month_plot = sns.countplot(x='month', hue='is_canceled', data = df)
         plt.title("Reservation Status per Month", fontsize = 15)
         month_plot
```

```
Out[59]: <AxesSubplot: title={'center': 'Reservation Status per Month'}, xlabel='mon
         th', ylabel='count'>
```



```
In [64]: cancelled_data = df[df['is_canceled'] == 1]
         top_10_country = cancelled_data.country.value_counts()[:10]
         plt.title("Top 10 countries with reservation cancelled")
         plt.pie(top_10_country, labels = top_10_country.index)
```

```
Out[64]: ([<matplotlib.patches.Wedge at 0x1523d870890>,
           <matplotlib.patches.Wedge at 0x1523d88bf10>,
           <matplotlib.patches.Wedge at 0x1523d89cf50>,
           <matplotlib.patches.Wedge at 0x1523d89ded0>,
           <matplotlib.patches.Wedge at 0x1523d89ed90>,
           <matplotlib.patches.Wedge at 0x1523d89fe50>,
           <matplotlib.patches.Wedge at 0x1523d8a8e50>,
           <matplotlib.patches.Wedge at 0x1523d8aa050>,
           <matplotlib.patches.Wedge at 0x1523d89ee90>,
           <matplotlib.patches.Wedge at 0x1523d8b4750>],
          [Text(-0.6485627932914347, 0.8884628878900974, 'PRT'),
           Text(-0.12448208506709475, -1.0929337630878408, 'GBR'),
           Text(0.27961841679380417, -1.0638672572223127, 'ESP'),
           Text(0.6082658179076615, -0.9165220645271579, 'FRA'),
           Text(0.8244472353575597, -0.7282079072018354, 'ITA'),
           Text(0.9549296655412358, -0.5459938954505844, 'DEU'),
           Text(1.031263324511448, -0.38274790073571413, 'IRL'),
           Text(1.0729039635595403, -0.24264600754644422, 'BRA'),
           Text(1.0926165756675046, -0.1272360741952387, 'USA'),
           Text(1.0992091152526196, -0.0417051668927687, 'BEL')])
```



```
In [65]: df.market_segment.value_counts(normalize = True)
```

```
Out[65]: Online TA        0.474373
         Offline TA/TO    0.203199
         Groups           0.166580
         Direct           0.104695
         Corporate        0.042986
         Complementary    0.006173
         Aviation         0.001993
         Name: market_segment, dtype: float64
```

```
In [66]: cancelled_data['market_segment'].value_counts(normalize = True)
```

```
Out[66]: Online TA        0.469685
         Groups           0.273979
         Offline TA/TO    0.187484
         Direct           0.043485
         Corporate        0.022150
         Complementary    0.002038
         Aviation         0.001178
         Name: market_segment, dtype: float64
```

Insights:

1. About 37% of clients had cancelled their reservatiom, which makes a significant impact on hotel industry.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

2. In comparison to resort hotels, city hotels have more bookings. It's possible that resort hotels are costlier than city hotels.

3. The number of confirmed and cancelled reservations were largest in the month of August. In january, mos cancelled reservations.

4. Cancellations are more common when the price are higher as it increases the cost of living.

5. The country - Portugal - has the highest number of hotel cancellations.

Suggestions:

1. Hotels should review their price, and try to lower rates for specific locations. Thy can offer some discounts to customers.

2. The hotels should provide reasonable discounts on weekends and holidays to reduce the number of cancellations.

3. In January, hotels can start extensive marketing campaigns to attract customers.

4. In Portugal, hotels can upgrade the quality of services and prices so that it can attract more customers with less cancellations.

In [ ]: