# HotelCancellationDataAnalysis

May 24, 2024

## 1 Importing Libraries

[253]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filewarnings("ignore")
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In[253], line 5
      3 import seaborn as sns
      4 import warnings
----> 5 warnings.filewarnings("ignore")

AttributeError: module 'warnings' has no attribute 'filewarnings'
```

## 2 Loading the dataset

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

## 3 Explorartory Data Analysis and Data Cleaning

[285]:
```
df.head()
```

[285]:

| | 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  …  \
0                        0                     0       2  …
1                        0                     0       2  …
2                        0                     1       1  …
3                        0                     1       1  …
4                        0                     2       2  …

   booking_changes  deposit_type days_in_waiting_list customer_type   adr  \
0                3    No Deposit                    0     Transient   0.0
1                4    No Deposit                    0     Transient   0.0
2                0    No Deposit                    0     Transient  75.0
3                0    No Deposit                    0     Transient  75.0
4                0    No Deposit                    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 month
0              2015-01-07     1
1              2015-01-07     1
2              2015-02-07     2
3              2015-02-07     2
4              2015-03-07     3

[5 rows x 31 columns]
```

[287]:
```
df.shape
```

[287]:
```
(118897, 31)
```

[289]:
```
df.columns
```

[289]:
```
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',
       'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'month'],
      dtype='object')
```

[291]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 118897 entries, 0 to 119389
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  int64
 2   lead_time                       118897 non-null  int64
 3   arrival_date_year               118897 non-null  int64
 4   arrival_date_month              118897 non-null  object
 5   arrival_date_week_number        118897 non-null  int64
 6   arrival_date_day_of_month       118897 non-null  int64
 7   stays_in_weekend_nights         118897 non-null  int64
 8   stays_in_week_nights            118897 non-null  int64
 9   adults                          118897 non-null  int64
 10  children                        118897 non-null  float64
 11  babies                          118897 non-null  int64
 12  meal                            118897 non-null  object
 13  country                         118897 non-null  object
 14  market_segment                  118897 non-null  object
 15  distribution_channel            118897 non-null  object
 16  is_repeated_guest               118897 non-null  int64
 17  previous_cancellations          118897 non-null  int64
 18  previous_bookings_not_canceled  118897 non-null  int64
 19  reserved_room_type              118897 non-null  object
 20  assigned_room_type              118897 non-null  object
 21  booking_changes                 118897 non-null  int64
 22  deposit_type                    118897 non-null  object
 23  days_in_waiting_list            118897 non-null  int64
 24  customer_type                   118897 non-null  object
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64
 27  total_of_special_requests       118897 non-null  int64
 28  reservation_status              118897 non-null  object
 29  reservation_status_date         118897 non-null  datetime64[ns]
 30  month                           118897 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(1), int64(16), object(11)
memory usage: 28.6+ MB
```

[293]:
```
df['reservation_status_date'] = pd.
↪to_datetime(df['reservation_status_date'],format='mixed')
```

[295]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 118897 entries, 0 to 119389
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118897 non-null  object
 1   is_canceled                     118897 non-null  int64
 2   lead_time                       118897 non-null  int64
 3   arrival_date_year               118897 non-null  int64
 4   arrival_date_month              118897 non-null  object
 5   arrival_date_week_number        118897 non-null  int64
 6   arrival_date_day_of_month       118897 non-null  int64
 7   stays_in_weekend_nights         118897 non-null  int64
 8   stays_in_week_nights            118897 non-null  int64
 9   adults                          118897 non-null  int64
 10  children                        118897 non-null  float64
 11  babies                          118897 non-null  int64
 12  meal                            118897 non-null  object
 13  country                         118897 non-null  object
 14  market_segment                  118897 non-null  object
 15  distribution_channel            118897 non-null  object
 16  is_repeated_guest               118897 non-null  int64
 17  previous_cancellations          118897 non-null  int64
 18  previous_bookings_not_canceled  118897 non-null  int64
 19  reserved_room_type              118897 non-null  object
 20  assigned_room_type              118897 non-null  object
 21  booking_changes                 118897 non-null  int64
 22  deposit_type                    118897 non-null  object
 23  days_in_waiting_list            118897 non-null  int64
 24  customer_type                   118897 non-null  object
 25  adr                             118897 non-null  float64
 26  required_car_parking_spaces     118897 non-null  int64
 27  total_of_special_requests       118897 non-null  int64
 28  reservation_status              118897 non-null  object
 29  reservation_status_date         118897 non-null  datetime64[ns]
 30  month                           118897 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(1), int64(16), object(11)
memory usage: 28.6+ MB
```

[297]:
```
df.describe(include = 'object')
```

[297]:
```
              hotel arrival_date_month    meal country market_segment  \
count        118897             118897  118897  118897         118897
unique            2                 12       5     177              7
top     City Hotel             August      BB     PRT      Online TA
freq          79301              13852   91862   48585          56402

       distribution_channel reserved_room_type assigned_room_type  \
count                118897             118897             118897
unique                    5                 10                 12
top                   TA/TO                  A                  A
freq                  97729              85600              73862

       deposit_type customer_type reservation_status
count        118897        118897             118897
unique            3             4                  3
top      No Deposit     Transient          Check-Out
freq         104163         89173              74745
```

[299]:
```
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' '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'
 'Aviation']
--------------------------------------------------
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
--------------------------------------------------
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'B' 'P']
--------------------------------------------------
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'L' 'K' 'P']
--------------------------------------------------
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
--------------------------------------------------
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
--------------------------------------------------
reservation_status
['Check-Out' 'Canceled' 'No-Show']
--------------------------------------------------
```

[301]:
```
df.isnull().sum()
```

[301]:
```
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
month                           0
dtype: int64
```

[305]:
```
# remove columns company and agent as they have less dependence on␣
 ↪cancellations and too many null values that cannot be handled.
# remove the rows where children values are null
df.drop(columns = ['agent', 'company'], axis = 1, inplace = True, errors =␣
 ↪'ignore')
df.dropna(inplace = True)
```

[307]:
```
df.isnull().sum()
```

[307]:
```
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
month                           0
dtype: int64
```

[309]:
```
df.describe()
```

[309]:

| | is_canceled | lead_time | arrival_date_year |
|---|---|---|---|
| count | 118897.000000 | 118897.000000 | 118897.000000 |
| mean | 0.371347 | 104.312018 | 2016.157657 |
| 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 | 0.483167 | 106.903570 | 0.707462 |

| | arrival_date_week_number | arrival_date_day_of_month |
|---|---|---|
| count | 118897.000000 | 118897.000000 |
| mean | 27.166674 | 15.800802 |
| 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 | 13.589966 | 8.780321 |

| | stays_in_weekend_nights | stays_in_week_nights | adults |
|---|---|---|---|
| count | 118897.000000 | 118897.000000 | 118897.000000 |
| mean | 0.928905 | 2.502157 | 1.858390 |
| 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 | 0.996217 | 1.900171 | 0.578578 |

| | children | babies | is_repeated_guest |
|---|---|---|---|
| count | 118897.000000 | 118897.000000 | 118897.000000 |
| mean | 0.104208 | 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 |

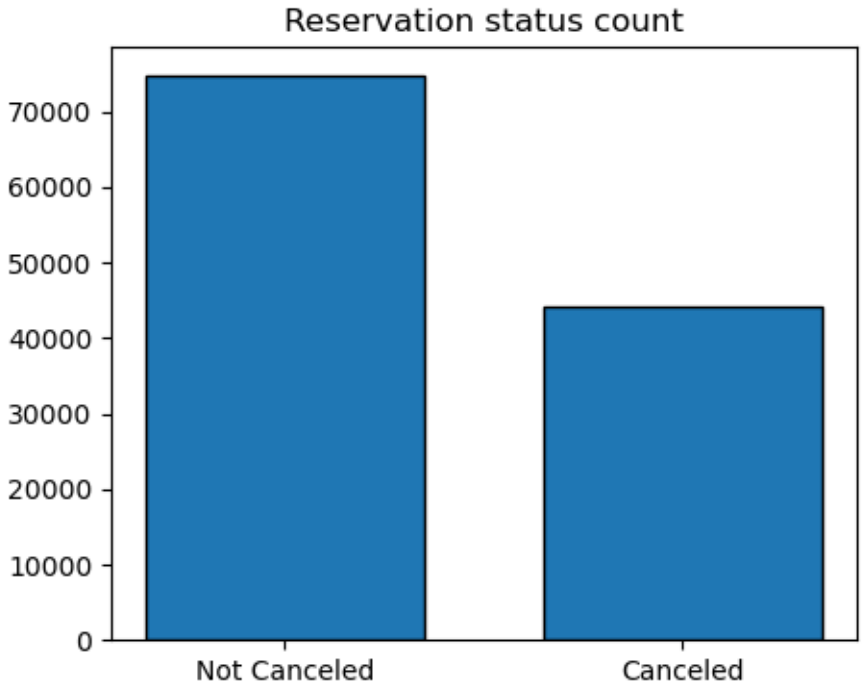
```
75%          0.000000       0.000000           0.000000
max         10.000000      10.000000           1.000000
std          0.399174       0.097381           0.176030

       previous_cancellations  previous_bookings_not_canceled  \
count           118897.000000                   118897.000000
mean                 0.087143                        0.131635
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                  0.845872                        1.484678

       booking_changes  days_in_waiting_list            adr  \
count    118897.000000         118897.000000  118897.000000
mean          0.221175              2.330774     101.958683
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     510.000000
std           0.652784             17.630525      48.091199

       required_car_parking_spaces  total_of_special_requests  \
count                118897.000000              118897.000000
mean                     0.061885                   0.571688
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                      0.244173                   0.792680

            reservation_status_date          month
count                        118897  118897.000000
mean   2016-07-31 20:51:35.775334912       6.386393
min              2014-10-17 00:00:00       1.000000
25%              2016-02-03 00:00:00       3.000000
50%              2016-08-05 00:00:00       6.000000
75%              2017-02-16 00:00:00       9.000000
max              2017-12-09 00:00:00      12.000000
std                              NaN       3.390731
```

[311]:
```
#removing outlier
df = df[df['adr']<5000]
```

# 4 Data Analysis and Visualizations

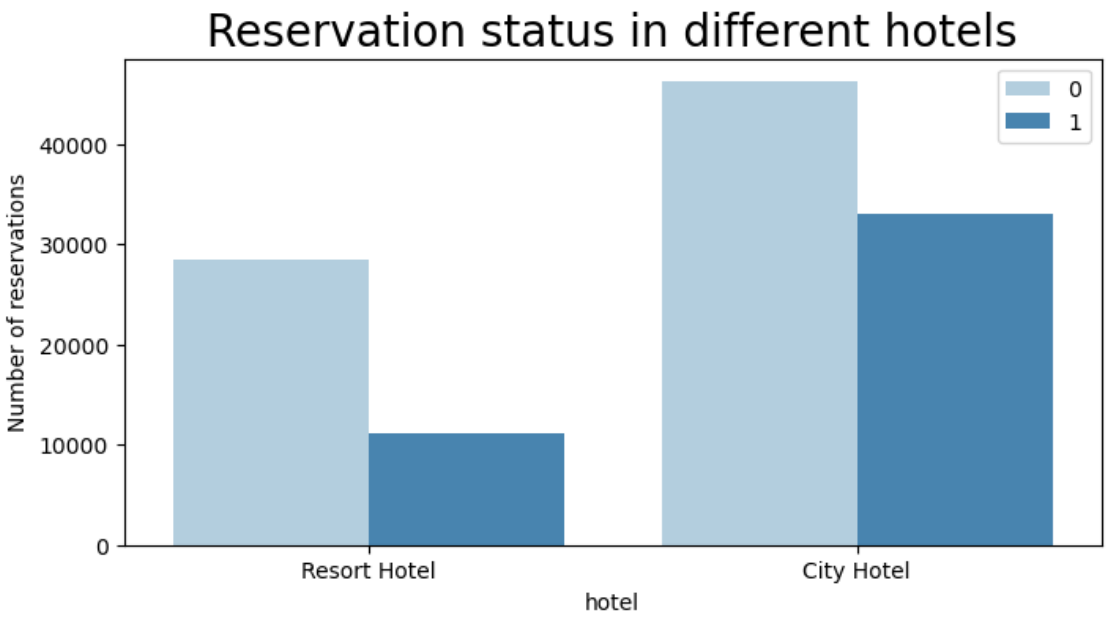
[313]:
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)
plt.figure(figsize = (5,4))
plt.title('Reservation status count')
plt.bar(['Not Canceled', 'Canceled'],df['is_canceled'].value_counts(),␣
 ↪edgecolor = 'k', width = 0.7)
plt.show()
```

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




[315]:
```
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')
```

[315]:
```
Text(0, 0.5, 'Number of reservations')
```




[317]:
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

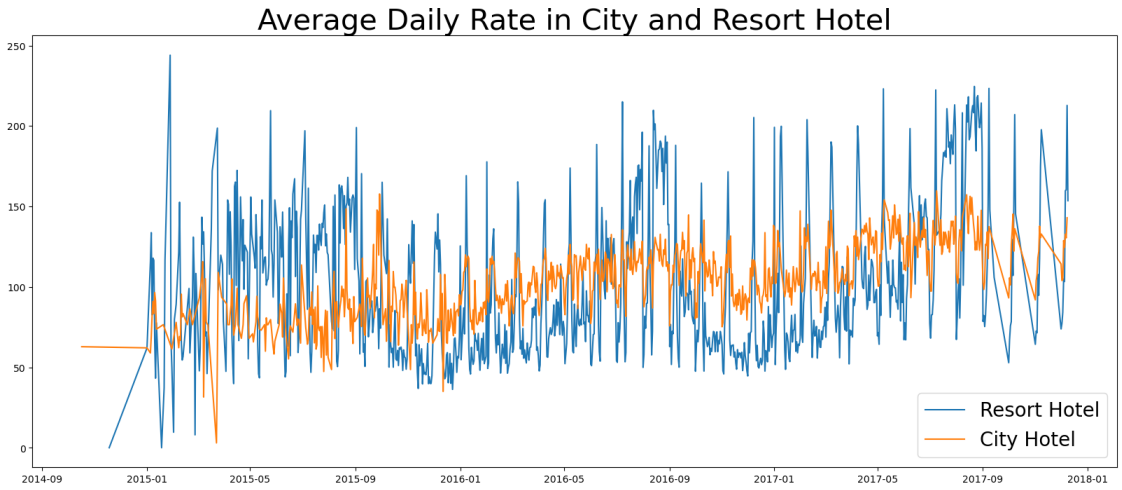
[317]:
```
is_canceled
0    0.72025
1    0.27975
Name: proportion, dtype: float64
```

[319]:
```
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

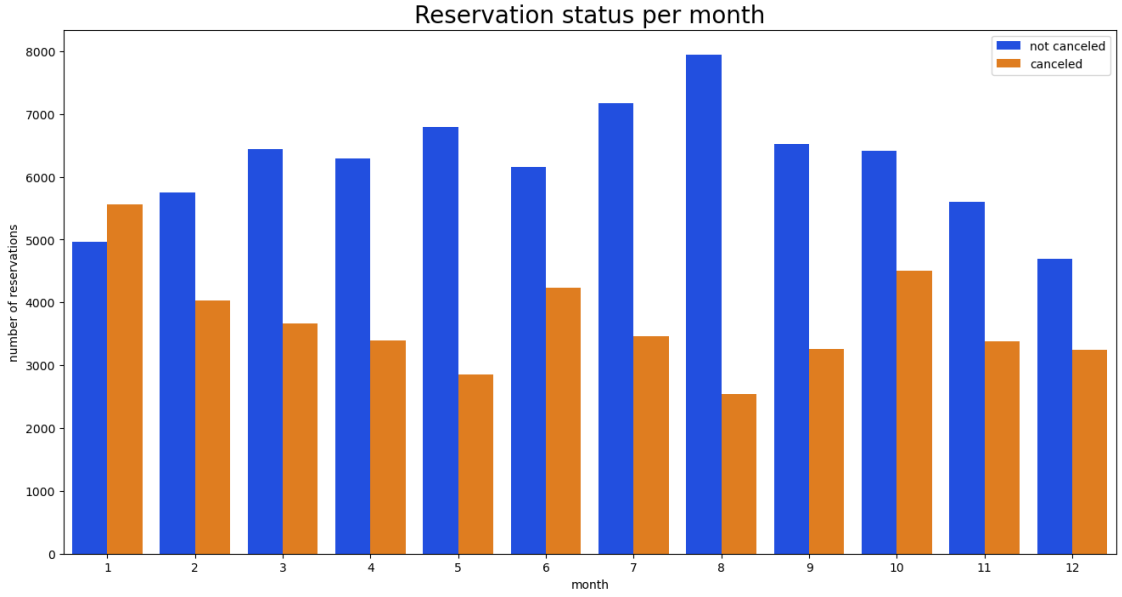
[319]:
```
is_canceled
0    0.582918
1    0.417082
Name: proportion, dtype: float64
```

[321]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

[323]:
```
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 = 'City Hotel')
plt.legend(fontsize = 20)
plt.show()
```

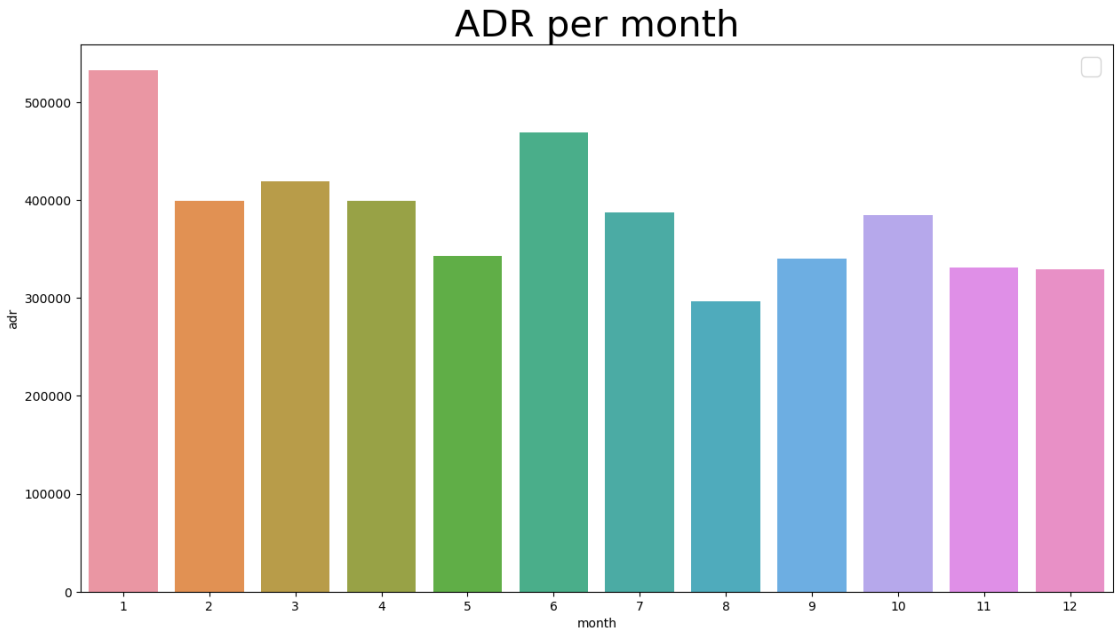



[327]:
```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x = 'month',hue = 'is_canceled', data = df, palette =␣
↪'bright')
legend_labels,_ = ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per month', size = 20)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.legend(['not canceled', 'canceled'])
plt.show()
```

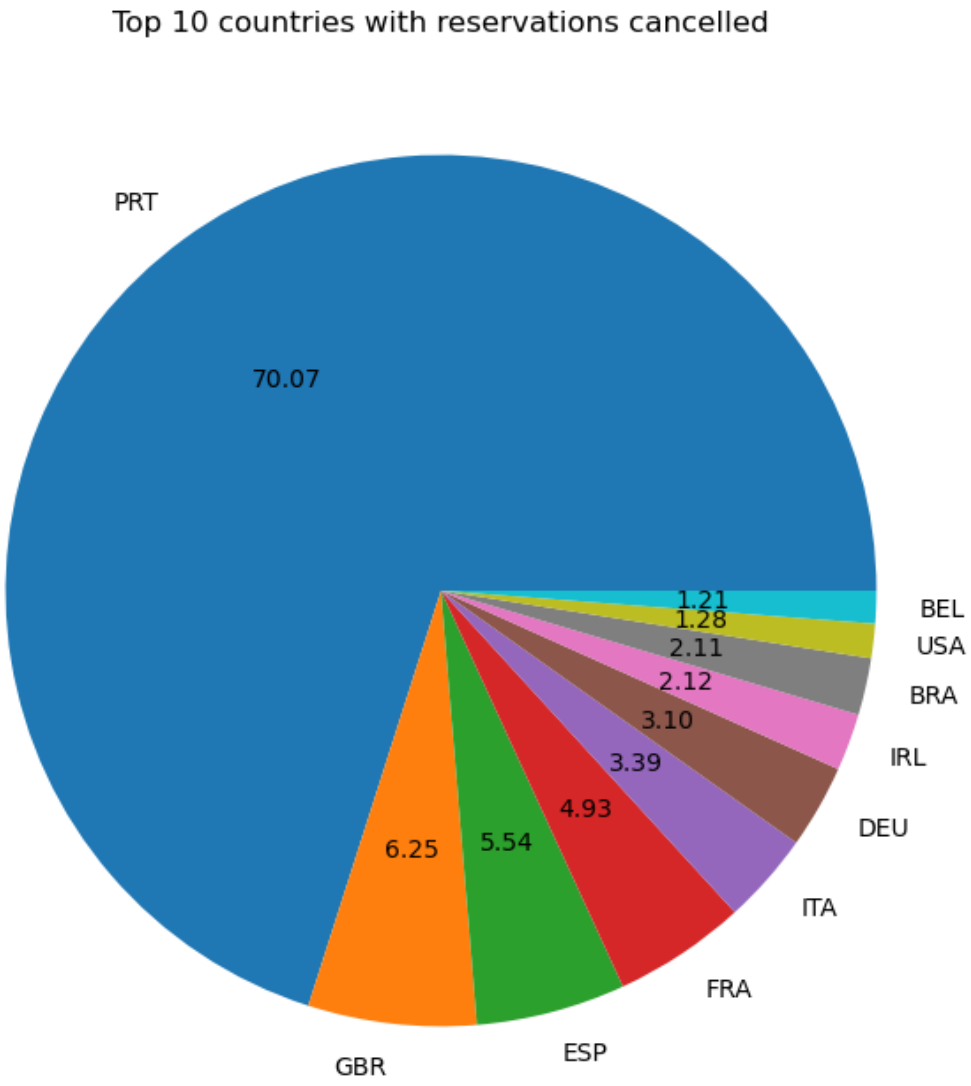


[329]:
```
plt.figure(figsize = (15,8))
plt.title('ADR per month', fontsize = 30)
sns.barplot(x ='month', y = 'adr', data = df[df['is_canceled'] == 1].
 ↪groupby('month')[['adr']].sum().reset_index())
plt.legend(fontsize = 20)
plt.show()
```

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```




[331]:
```
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 cancelled')
plt.pie(top_10_country, autopct = '%.2f', labels = top_10_country.index)
plt.show()
```




[333]:
```
df['market_segment'].value_counts()
```

[333]:
```
market_segment
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: count, dtype: int64
```

[335]:
```
df['market_segment'].value_counts(normalize = True)
```

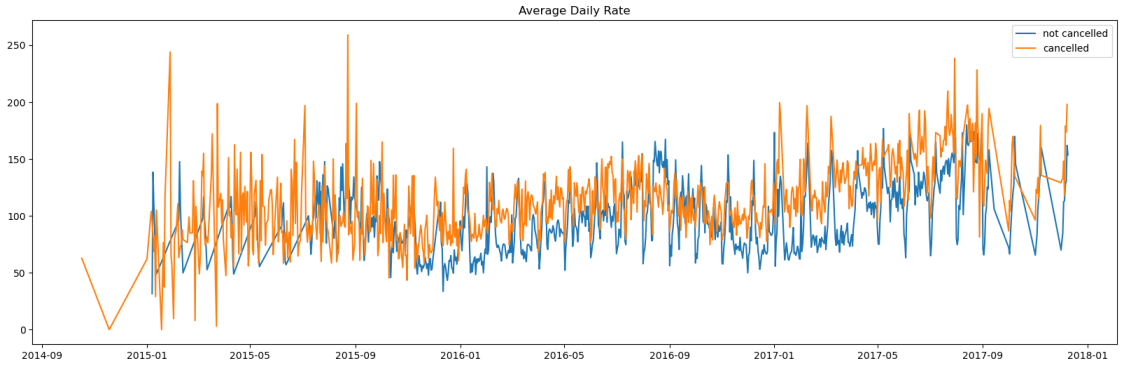
[335]:
```
market_segment
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: proportion, dtype: float64
```

[337]:
```
cancelled_data['market_segment'].value_counts(normalize = True)
```

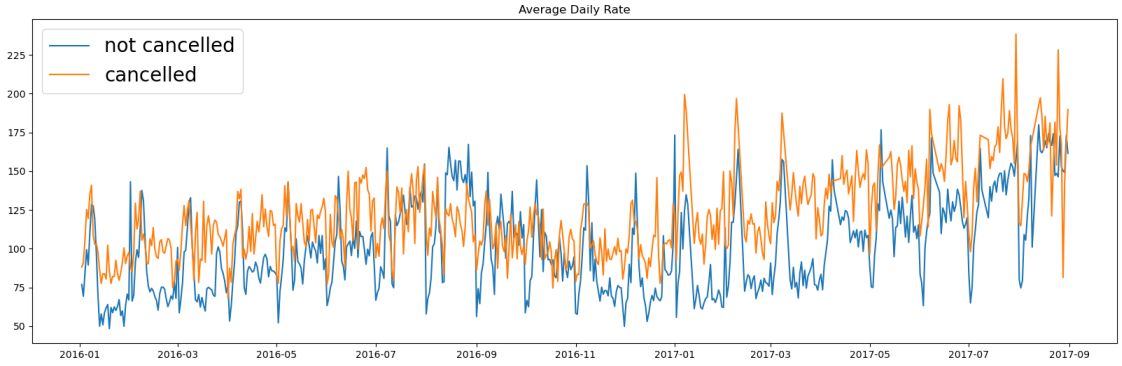
[337]:
```
market_segment
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: proportion, dtype: float64
```

[339]:
```
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_df = df[df['is_canceled'] == 0]
not_cancelled_df_adr = not_cancelled_df.
 ↪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 cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label = 'cancelled')
plt.legend()
```

[339]:
```
<matplotlib.legend.Legend at 0x7fa471b0f8d0>
```




[341]:
```
cancelled_df_adr =␣
 ↪cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') &␣
 ↪(cancelled_df_adr['reservation_status_date']<'2017-09')]
not_cancelled_df_adr =␣
 ↪not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016')␣
 ↪& (not_cancelled_df_adr['reservation_status_date']<'2017-09')]

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 cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label = 'cancelled')
plt.legend(fontsize = 20)
plt.show()
```




[ ]:

[ ]: