In [82…

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

Loading the dataset

In [83…

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

In [84…

```
df.head()
```

Out[84]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arriva
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	

5 rows × 32 columns

In [85…

```
df.tail()
```

Out[85]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month
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 × 32 columns

In [86…

```
df.shape
```

Out[86]:

```
(119390, 32)
```

In [87…

```
df.columns
```

Out[87]:

```
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', 'mea
l',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'age
nt',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [88…

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           119390 non-null  object 
 1   is_canceled                     119390 non-null  int64  
 2   lead_time                       119390 non-null  int64  
 3   arrival_date_year               119390 non-null  int64  
 4   arrival_date_month              119390 non-null  object 
 5   arrival_date_week_number        119390 non-null  int64  
 6   arrival_date_day_of_month       119390 non-null  int64  
 7   stays_in_weekend_nights         119390 non-null  int64  
 8   stays_in_week_nights            119390 non-null  int64  
 9   adults                          119390 non-null  int64  
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64  
 12  meal                            119390 non-null  object 
 13  country                         118902 non-null  object 
 14  market_segment                  119390 non-null  object 
 15  distribution_channel            119390 non-null  object 
 16  is_repeated_guest               119390 non-null  int64  
 17  previous_cancellations          119390 non-null  int64  
 18  previous_bookings_not_canceled  119390 non-null  int64  
 19  reserved_room_type              119390 non-null  object 
 20  assigned_room_type              119390 non-null  object 
 21  booking_changes                 119390 non-null  int64  
 22  deposit_type                    119390 non-null  object 
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64  
 26  customer_type                   119390 non-null  object 
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

In [89…

```
# Here we have to perform our analysis on 'reservation_status_date' i
# this into 'date-time'
```

In [90…

```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status
```

In [91…

```
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  datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

In [92…

```
df.describe()
```

Out[92]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_nun
count	119390.000000	119390.000000	119390.000000	119390.00
mean	0.370416	104.011416	2016.156554	27.16
std	0.482918	106.863097	0.707476	13.60
min	0.000000	0.000000	2015.000000	1.00
25%	0.000000	18.000000	2016.000000	16.00
50%	0.000000	69.000000	2016.000000	28.00
75%	1.000000	160.000000	2017.000000	38.00
max	1.000000	737.000000	2017.000000	53.00

In [93…

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

Out[93]:

	hotel	arrival_date_month	meal	country	market_segment	distri
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	

In [94…

```
for col in df.describe(include='object').columns:
    print (col)
    print (df[col].unique())
    print ('-'*50)
```

```
hotel
['Resort Hotel' 'City Hotel']
--------------------------------------------------
arrival_date_month
['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
--------------------------------------------------
meal
['BB' 'FB' 'HB' 'SC' 'Undefined']
--------------------------------------------------
country
['PRT' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' nan 'ROU' 'NOR' 'OMN' 'ARG' 'POL'
 'DEU' 'BEL' 'CHE' 'CN' 'GRC' 'ITA' 'NLD' 'DNK' 'RUS' 'SWE' 'AUS' 'ES
T'
 'CZE' 'BRA' 'FIN' 'MOZ' 'BWA' 'LUX' 'SVN' 'ALB' 'IND' 'CHN' 'MEX' 'MA
R'
 'UKR' 'SMR' 'LVA' 'PRI' 'SRB' 'CHL' 'AUT' 'BLR' 'LTU' 'TUR' 'ZAF' 'AG
O'
 'ISR' 'CYM' 'ZMB' 'CPV' 'ZWE' 'DZA' 'KOR' 'CRI' 'HUN' 'ARE' 'TUN' 'JA
M'
 'HRV' 'HKG' 'IRN' 'GEO' 'AND' 'GIB' 'URY' 'JEY' 'CAF' 'CYP' 'COL' 'GG
Y'
 'KWT' 'NGA' 'MDV' 'VEN' 'SVK' 'FJI' 'KAZ' 'PAK' 'IDN' 'LBN' 'PHL' 'SE
N'
 'SYC' 'AZE' 'BHR' 'NZL' 'THA' 'DOM' 'MKD' 'MYS' 'ARM' 'JPN' 'LKA' 'CU
B'
 'CMR' 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SYR' 'SGP' 'BD
I'
 'SAU' 'VNM' 'PLW' 'QAT' 'EGY' 'PER' 'MLT' 'MWI' 'ECU' 'MDG' 'ISL' 'UZ
B'
 'NPL' 'BHS' 'MAC' 'TGO' 'TWN' 'DJI' 'STP' 'KNA' 'ETH' 'IRQ' 'HND' 'RW
A'
 'KHM' 'MCO' 'BGD' 'IMN' 'TJK' 'NIC' 'BEN' 'VGB' 'TZA' 'GAB' 'GHA' 'TM
P'
 'GLP' 'KEN' 'LIE' 'GNB' 'MNE' 'UMI' 'MYT' 'FRO' 'MMR' 'PAN' 'BFA' 'LB
Y'
 'MLI' 'NAM' 'BOL' 'PRY' 'BRB' 'ABW' 'AIA' 'SLV' 'DMA' 'PYF' 'GUY' 'LC
A'
 'ATA' 'GTM' 'ASM' 'MRT' 'NCL' 'KIR' 'SDN' 'ATF' 'SLE' 'LAO']
--------------------------------------------------
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Gro
ups'
 '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 [95…

```
df.isnull().sum()
```

Out[95]:

```
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 [96…

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

In [97…

```
df.isnull().sum()
```

```
Out[97]: 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
         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 [98…  df.dropna(inplace=True)
```

```
In [99…  df.describe()
```

Out[99]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_nun
count	118898.000000	118898.000000	118898.000000	118898.00
mean	0.371352	104.311435	2016.157656	27.16
std	0.483168	106.903309	0.707459	13.58
min	0.000000	0.000000	2015.000000	1.00
25%	0.000000	18.000000	2016.000000	16.00
50%	0.000000	69.000000	2016.000000	28.00
75%	1.000000	161.000000	2017.000000	38.00
max	1.000000	737.000000	2017.000000	53.00

There are lots of outlier here. We will not remove all of them as we are not going to use it further (like in children column, there are values like 10 or 0).

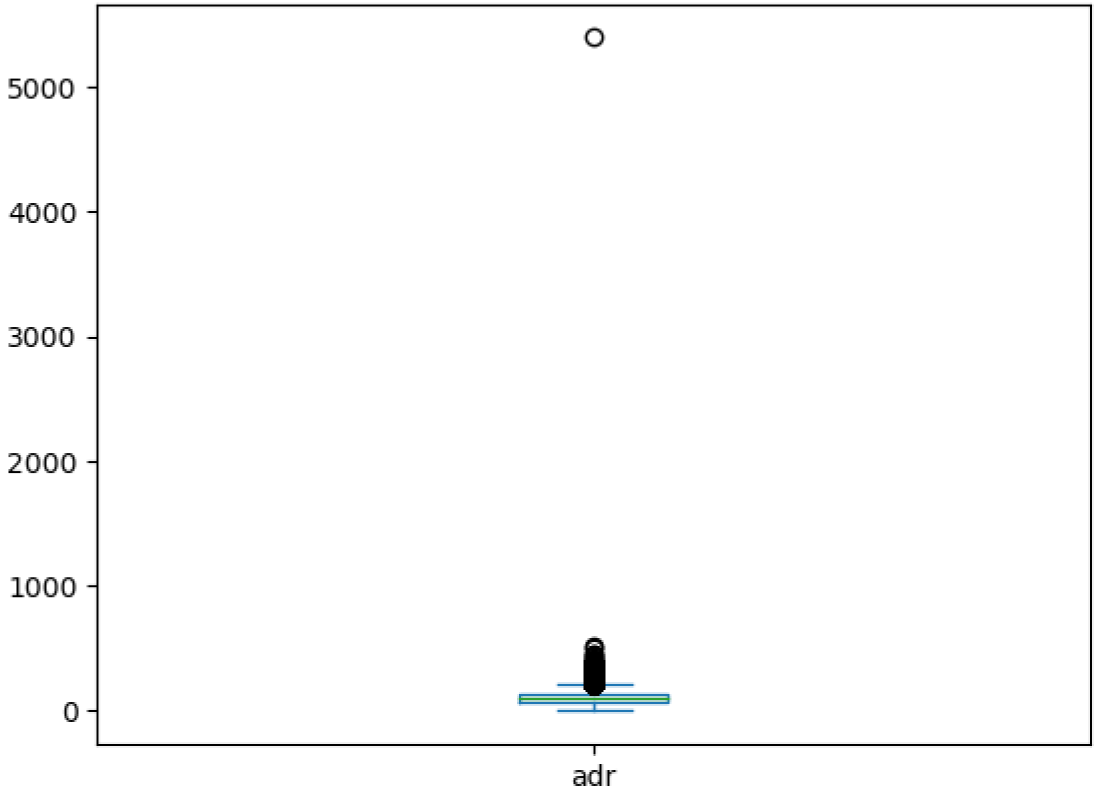
But we will remove outliers in adr (Average Daily Rate), there are values like 5400, or -6.38, these are vast outlier and we have to remove them. We can also see them using box plot shown below.

In [10…

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

Out[100]:

```
<Axes: >
```



In [10…

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

In [10…

```
df.describe()
```

Out[102]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_nu
count	118897.000000	118897.000000	118897.000000	118897.0
mean	0.371347	104.312018	2016.157657	27.1
std	0.483167	106.903570	0.707462	13.5
min	0.000000	0.000000	2015.000000	1.0
25%	0.000000	18.000000	2016.000000	16.0
50%	0.000000	69.000000	2016.000000	28.0
75%	1.000000	161.000000	2017.000000	38.0
max	1.000000	737.000000	2017.000000	53.0

Data Analysis and Visualizations

First, we will see amount of reservations cancelled and amount of reservations not cancelled

In [10…

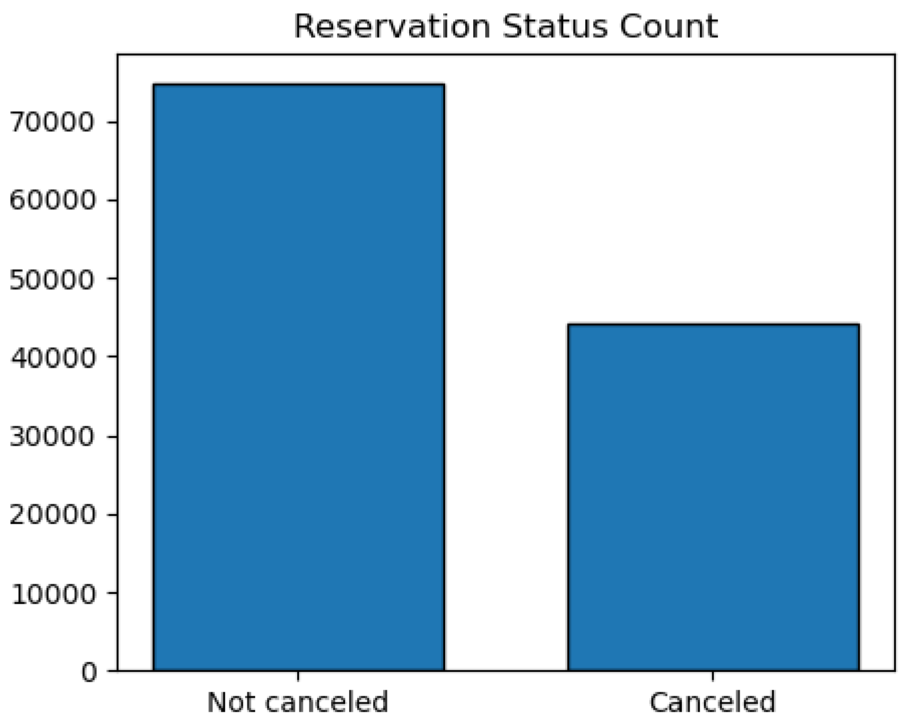
```
cancelled_perc = df['is_canceled'].value_counts(normalize=True)
cancelled_perc
```

Out[103]:

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

In [10…

```
plt.figure(figsize=(5,4))
plt.bar(['Not canceled','Canceled'],df['is_canceled'].value_counts(),
plt.title('Reservation Status Count')
plt.show()
```

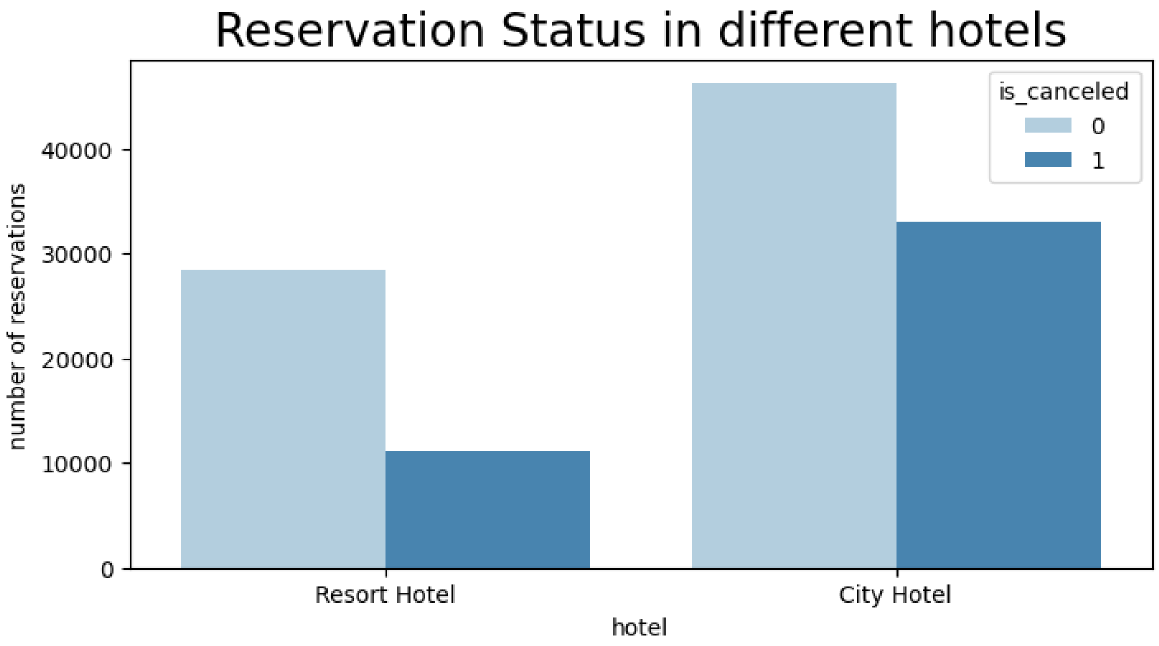


Then we will find which hotel has more cancellation rate.

In [10…

```
plt.figure(figsize=(8,4))
ax1 = sns.countplot(x='hotel',hue='is_canceled', data=df, palette='Bl
plt.title('Reservation Status in different hotels',size=20)
plt.xlabel('hotel')
plt.ylabel('number of reservations')

plt.show()
```



In the above figure, we draw a conclusion that cancellation rate is more in City Hotel as compared to Resort Hotel but ratio of not cancelled reservation to cancelled reservation seems more is more in Resort Hotel as compared to City hotel.

The reason behind this can be because Resort Hotel has price greater than City Hotel generally.

Conclusions: 1.) Cancellation in Resort Hotel can be due to high ADR (Average Daily Rate). 2.) Cancellation in City Hotel can be due to lack of maintenance and other cancellations.

In [10…

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

Out[106]:

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

In [10…

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

Out[107]:

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

In [10…

```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr'
resort_hotel
```

Out[108]:

	adr
reservation_status_date	
2014-11-18	0.000000
2015-01-01	61.966667
2015-01-05	115.363333
2015-01-06	133.677143
2015-01-07	82.485455
...	...
2017-12-05	103.287534
2017-12-06	159.808929
2017-12-07	160.306275
2017-12-08	212.767222
2017-12-09	153.570000

913 rows × 1 columns

In [10…

```
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].m
city_hotel
```

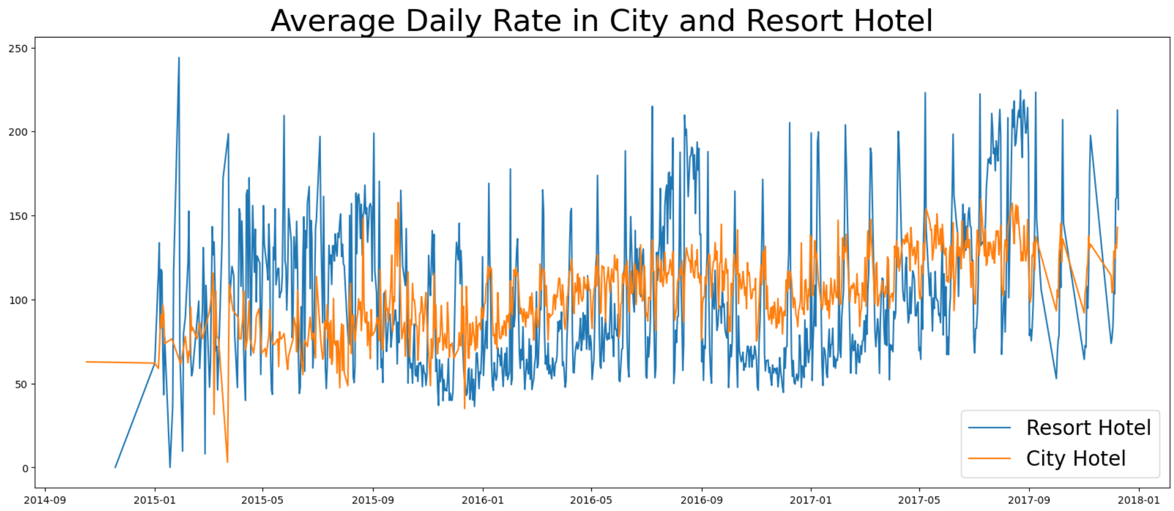
Out[109]:

	adr
reservation_status_date	
2014-10-17	62.800000
2015-01-01	62.063158
2015-01-05	58.900000
2015-01-06	69.216667
2015-01-07	82.877500
...	...
2017-12-04	128.755465
2017-12-05	124.544536
2017-12-06	132.725882
2017-12-07	130.473617
2017-12-08	142.949080

864 rows × 1 columns

In [11…

```
plt.figure(figsize = (20,8))
plt.title('Average Daily Rate in City and Resort Hotel', fontsize = 3
plt.plot(resort_hotel.index, resort_hotel['adr'], label = 'Resort Hot
plt.plot(city_hotel.index, city_hotel['adr'], label = 'City Hotel')
plt.legend(fontsize = 20)
plt.show()
```



We can visualise that orange line (City Hotel) is in the middle, that is ADR of City Hotel is in between ADR of Resort Hotel, that means ADR of City Hotel is less than ADR of Resort Hotel generally.

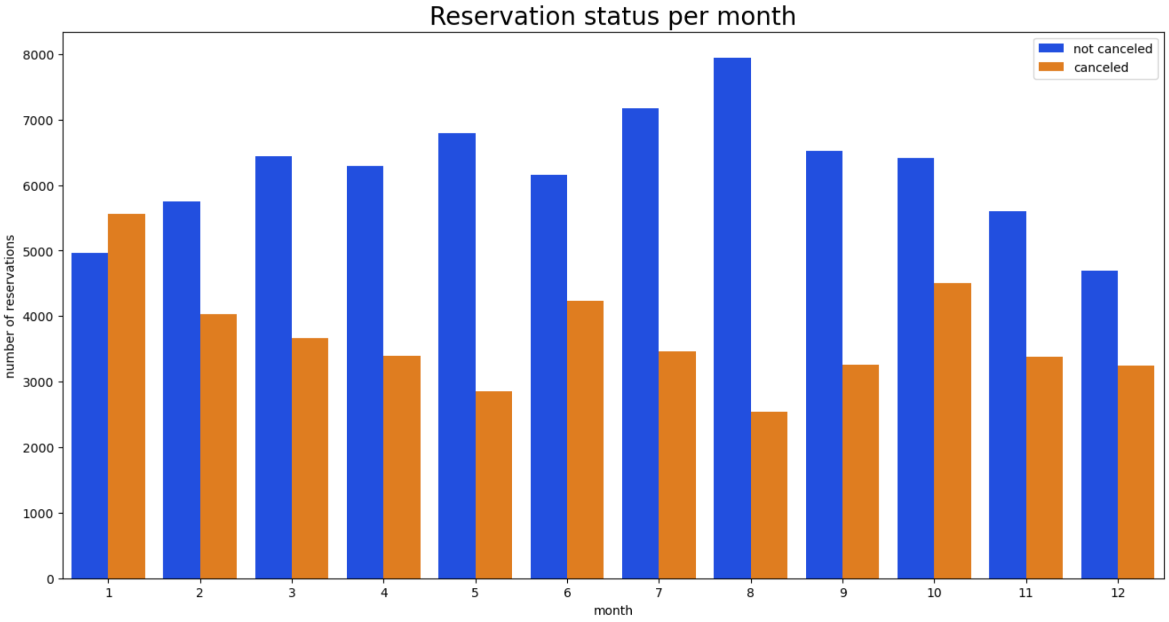
Next, we can see some spikes here, concluding that ADR of both City Hotel and Resort Hotel is high on weekends.

For some period of time, ADR of City Hotel is greater than ADR of Resort Hotel.

Now, I want to see which months have more reservations and cancellation rates.

In [11…

```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x = 'month', hue = 'is_canceled', data = df, pale
plt.title('Reservation status per month', size = 20)
plt.xlabel('month')
plt.ylabel('number of reservations')
plt.legend(['not canceled', 'canceled'])
plt.show()
```



We can visualise from the above graph that in January, larger number of cancellations are performed while in August, smaller number of cancellations are performed.

In August, there are larger number of reservations are done while in January, smaller number of reservations are done.

Now, it seems a bit confusing as when there are larger number of reservations (Auguts), cancellations are less while in case of January, reservations are less but cancellations are more.

So, one of the probable reason for such conclusion would be that ADR of Hotels in August is quite less and ADR of Hotels in January is quite high (therefore owing to more number of cancellations).

Now, we will see ADR for each month.

In [11…

```
df_grouped = df[df['is_canceled'] == 1].groupby('month')[['adr']].sum
```

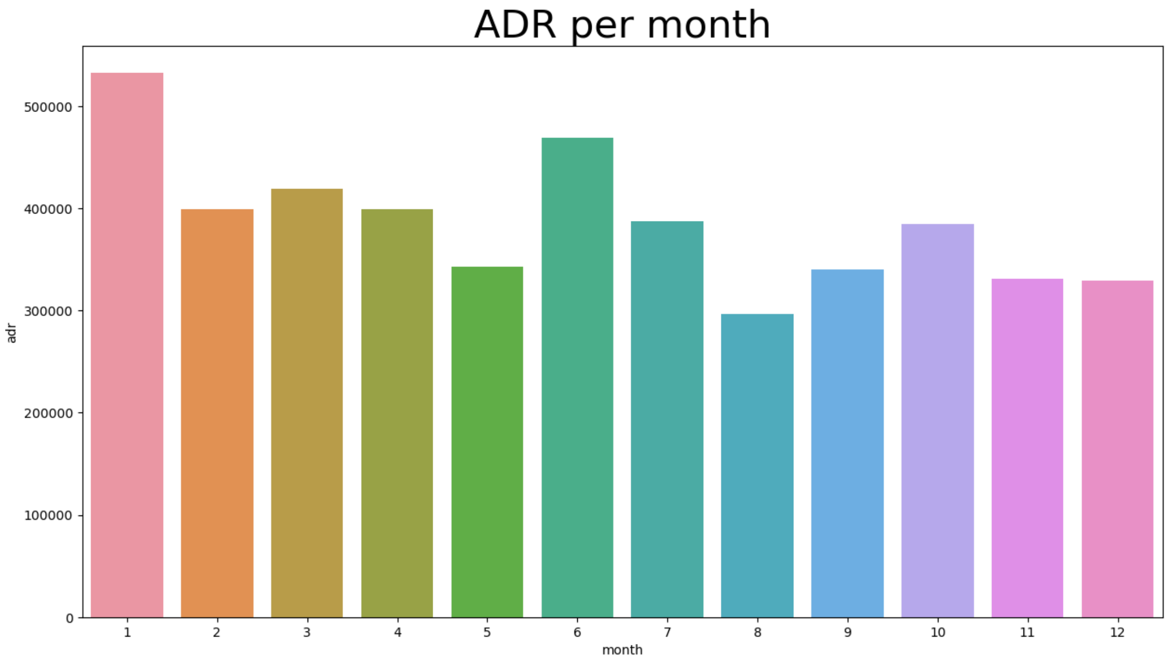
In [11…

```
df1 = df_grouped.reset_index()
```

In [11…

```
plt.figure(figsize = (15,8))

plt.title('ADR per month', fontsize = 30)
sns.barplot(x='month', y='adr', data = df1)
plt.show()
```

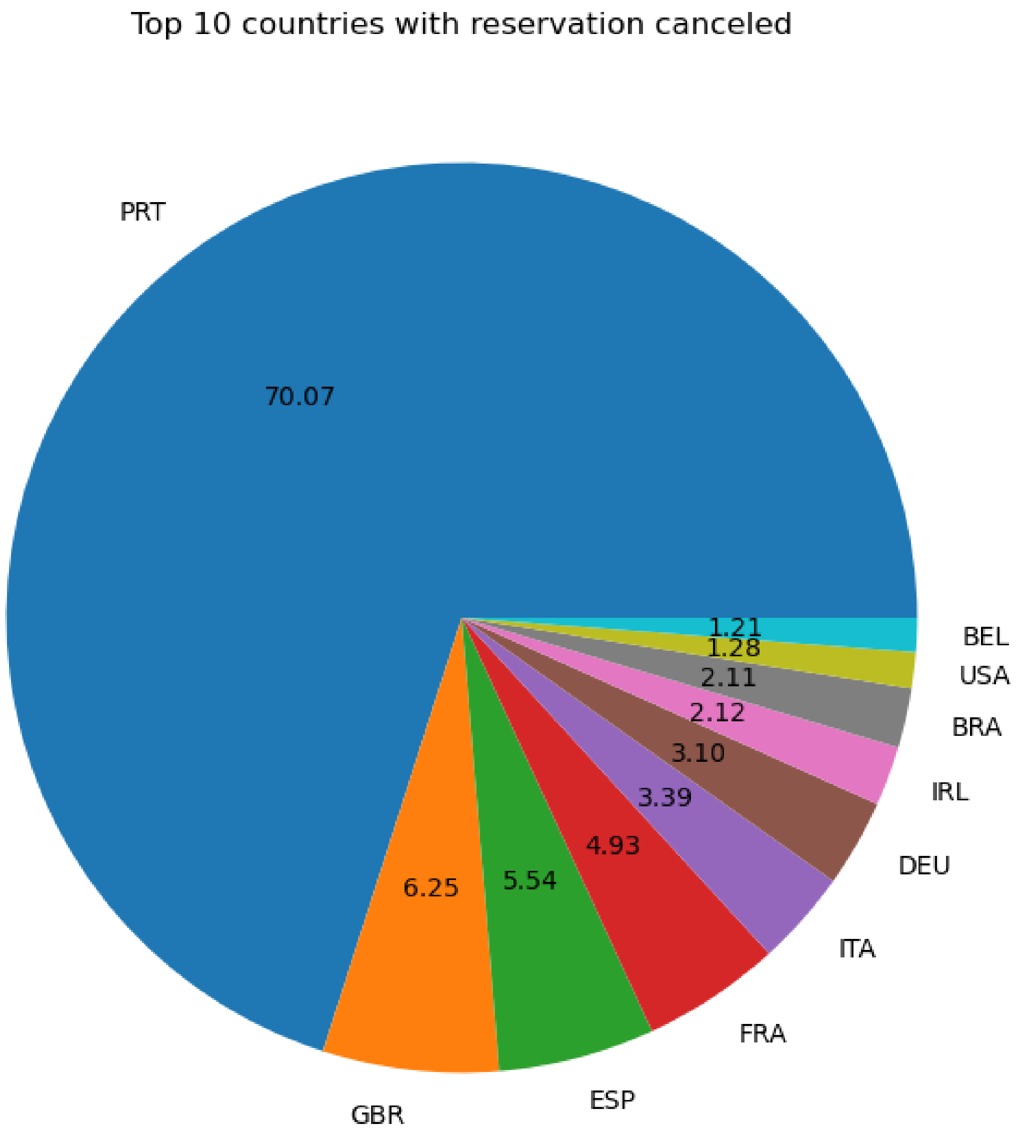


Thus, we can see that ADR of August is lowest of all, which leads to more reservations and less cancellations and ADR of January is highest among all, which leads to less reservations and more cancellations.

This proves our hypothesis that when prices are higher, eventually cancellations will be more (Because they will book the hotel but at last hour, they will think that it is costly, therefore leading to cancellation).

In [11…

```
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='%0.2f', labels=top_10_country.index)
plt.show()
```



We can visualise from above graph that, Portugal has witnessed huge number of cancellations.

Therefore, my suggestions to the hotels would be to enhance their facilities in Portugal, manage their prices, offering promotional discounts and doing advertisements.

```
In [11…  df['market_segment'].value_counts()
```

```
Out[117]: Online TA        56402
          Offline TA/TO    24159
          Groups           19806
          Direct           12448
          Corporate         5111
          Complementary      734
          Aviation           237
          Name: market_segment, dtype: int64
```

```
In [11…  df['market_segment'].value_counts(normalize=True)
```

Out[118]:

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

In [11…

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

Out[119]:

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

We can analyse that, majority of reservations are coming from Online TA(Travel Agent) but cancellations are also more from Online TA.

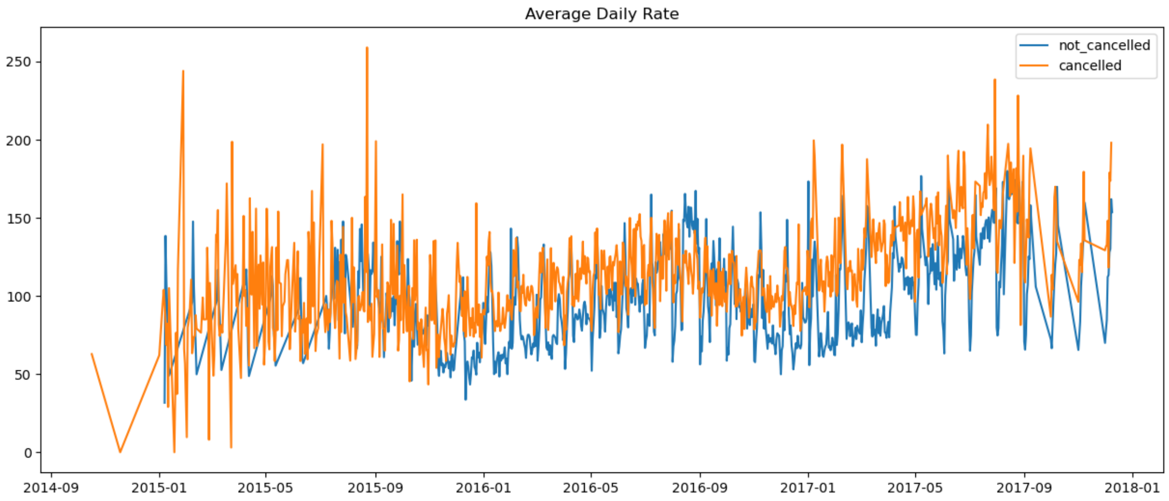
Reason behind this can be that the hotels are not meeting up the demands of the customers in terms of facilities, value for money etc or they are not as promising as they show in pictures during online search.

In [15…

```
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[
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
not_cancelled_df_adr.reset_index(inplace=True)
not_cancelled_df_adr.sort_values('reservation_status_date', inplace=T

plt.figure(figsize=(15,6))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancell
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_ad
plt.legend()
plt.show()
```



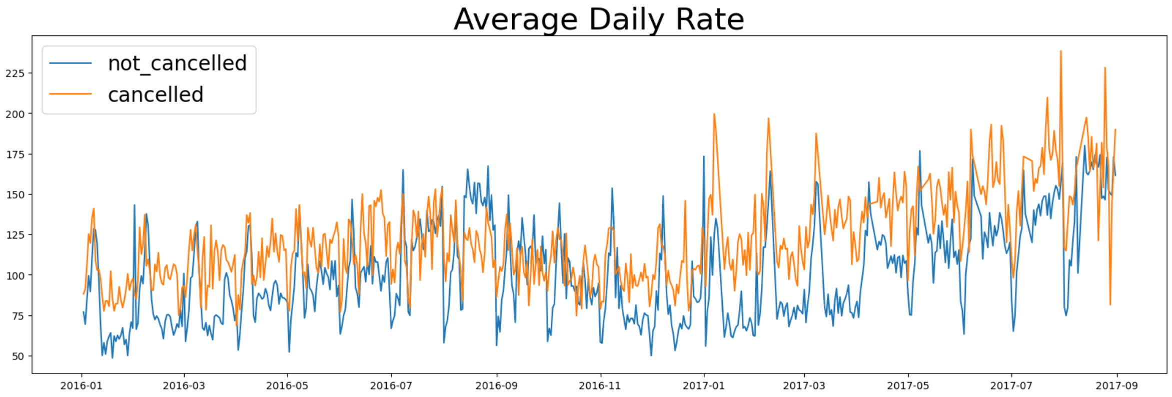
This data is quite messy, so we will take datas from 2016 to 2017-09

In [15…

```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_st
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['re
```

In [15…

```
plt.figure(figsize=(20,6))
plt.title('Average Daily Rate', fontsize=30)
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancell
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_ad
plt.legend(fontsize=20)
plt.show()
```



We can see that ADR is the factor that is mostly influencing the cancellation rates. As ADR is high, Cancellation rates are also high. Spikes is during the weekends.