Pritesh Panchal - My LinkedIn || My GitHub

Hotel Bookings Analysis using Python

Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading the dataset

In [3]:
```
df = pd.read_csv(r'C:\Users\DC\Desktop\HotelBookingsAnalysis_Python\Hotel Bookings_Data
```

EDA and Data Cleaning

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

Out[4]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	
2	Resort Hotel	0	7	2015	July	27	
3	Resort Hotel	0	13	2015	July	27	
4	Resort Hotel	0	14	2015	July	27	

5 rows × 36 columns

In [5]:
```
columns_to_drop = ['name', 'email', 'phone-number', 'credit_card']
df = df.drop(columns=columns_to_drop)
```

In [6]:
```
df.head()
```

Out[6]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	
2	Resort Hotel	0	7	2015	July	27	
3	Resort Hotel	0	13	2015	July	27	
4	Resort Hotel	0	14	2015	July	27	

5 rows × 32 columns

In [7]:
```
df.shape
```

Out[7]:
```
(119390, 32)
```

In [8]:
```
df.columns
```

Out[8]:
```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [9]:
```
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 [10]:

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

In [11]:

```
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 [12]:

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

Out[12]:

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

In [13]:

```
for col in df.describe(include = 'object').columns:
  print("COL_NAME -> ",col)
  print("UNIQUE_VALUES -> ",df[col].unique())
  print('-----------------------------------------------------------------------------------------')
```

```
COL_NAME ->  hotel
UNIQUE_VALUES ->  ['Resort Hotel' 'City Hotel']
-----------------------------------------------------------------------------------------
-----
COL_NAME ->  arrival_date_month
UNIQUE_VALUES ->  ['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
-----------------------------------------------------------------------------------------
-----
COL_NAME ->  meal
UNIQUE_VALUES ->  ['BB' 'FB' 'HB' 'SC' 'Undefined']
-----------------------------------------------------------------------------------------
-----
COL_NAME ->  country
UNIQUE_VALUES ->  ['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']
--------------------------------------------------------------------------------
-----
COL_NAME ->  market_segment
UNIQUE_VALUES ->  ['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Gro
ups'
 'Undefined' 'Aviation']
--------------------------------------------------------------------------------
-----
COL_NAME ->  distribution_channel
UNIQUE_VALUES ->  ['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
--------------------------------------------------------------------------------
-----
COL_NAME ->  reserved_room_type
UNIQUE_VALUES ->  ['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
--------------------------------------------------------------------------------
-----
COL_NAME ->  assigned_room_type
UNIQUE_VALUES ->  ['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K']
--------------------------------------------------------------------------------
-----
COL_NAME ->  deposit_type
UNIQUE_VALUES ->  ['No Deposit' 'Refundable' 'Non Refund']
--------------------------------------------------------------------------------
-----
COL_NAME ->  customer_type
UNIQUE_VALUES ->  ['Transient' 'Contract' 'Transient-Party' 'Group']
--------------------------------------------------------------------------------
-----
COL_NAME ->  reservation_status
UNIQUE_VALUES ->  ['Check-Out' 'Canceled' 'No-Show']
--------------------------------------------------------------------------------
-----
```

In [14]:

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

Out[14]:

```
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 [15]:
```
columns_to_drop = ['agent', 'company']
df = df.drop(columns=columns_to_drop)
```

In [16]:
```
df.isnull().sum()
```

Out[16]:
```
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 [17]:
```
df.dropna(inplace = True)
```

In [18]:
```
df.isnull().sum()
```

Out[18]:
```
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
```

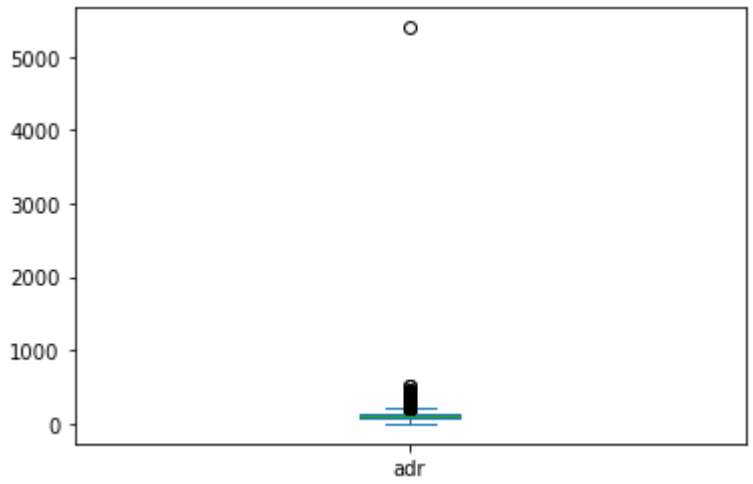
In [19]:
```
df.describe()
```

Out[19]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_mo
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000
mean	0.371352	104.311435	2016.157656	27.166555	15.800
std	0.483168	106.903309	0.707459	13.589971	8.780
min	0.000000	0.000000	2015.000000	1.000000	1.000
25%	0.000000	18.000000	2016.000000	16.000000	8.000
50%	0.000000	69.000000	2016.000000	28.000000	16.000
75%	1.000000	161.000000	2017.000000	38.000000	23.000
max	1.000000	737.000000	2017.000000	53.000000	31.000

In [20]:
```
df['adr'].plot(kind ='box')
```

Out[20]:
```
<AxesSubplot:>
```



In [21]:
```
df = df[df['adr']<5000]
```

In [22]:
```
df.describe()
```

Out[22]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_mo
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.000
mean	0.371347	104.312018	2016.157657	27.166674	15.800
std	0.483167	106.903570	0.707462	13.589966	8.780
min	0.000000	0.000000	2015.000000	1.000000	1.000
25%	0.000000	18.000000	2016.000000	16.000000	8.000
50%	0.000000	69.000000	2016.000000	28.000000	16.000
75%	1.000000	161.000000	2017.000000	38.000000	23.000
max	1.000000	737.000000	2017.000000	53.000000	31.000

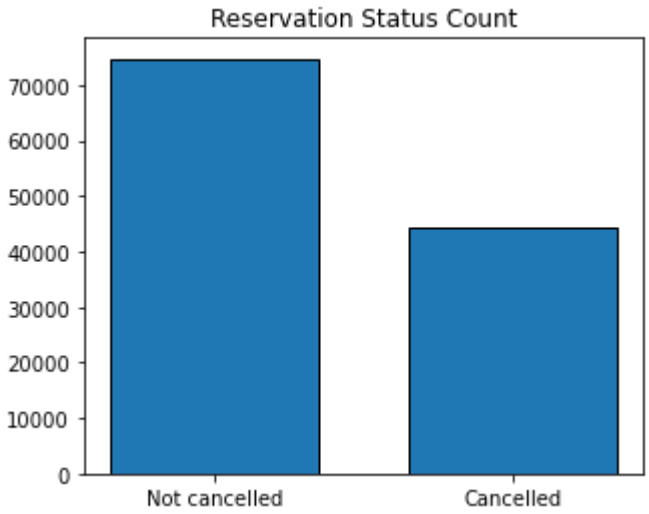
Data Analysis and Visualizations

In [23]:
```
cancelled_perc = df['is_canceled'].value_counts(normalize = 'True')
```

In [24]:
```
cancelled_perc
```

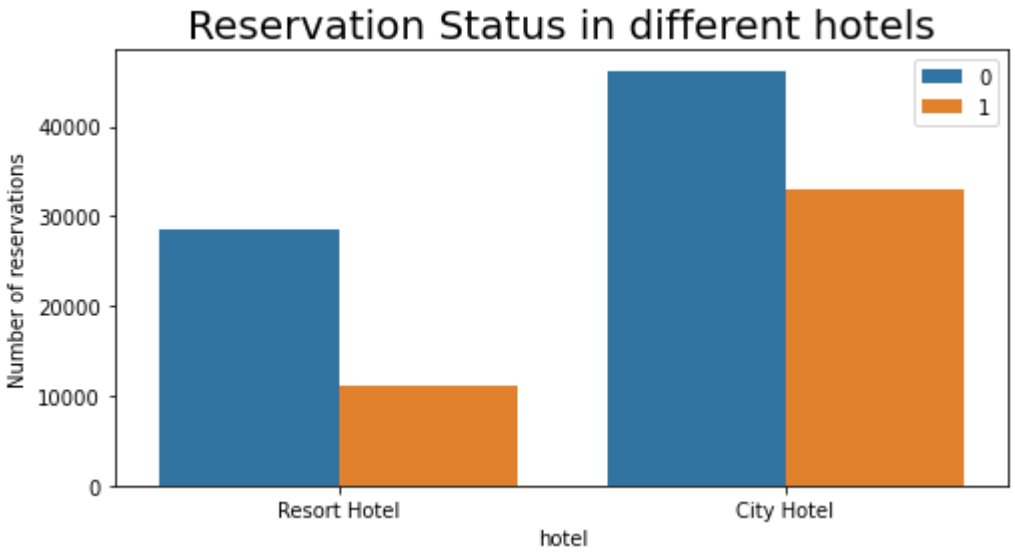
Out[24]:
```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

In [25]:
```
plt.figure(figsize = (5,4))
plt.title("Reservation Status Count")
plt.bar(['Not cancelled','Cancelled'],df['is_canceled'].value_counts(),edgecolor = 'k',
plt.show()
```



In [26]:
```
plt.figure(figsize = (8,4))
ax1 = sns.countplot(x = 'hotel',hue = 'is_canceled',data = df)
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')
```

Out[26]:
```
Text(0, 0.5, 'Number of reservations')
```



Findings :

City hotels has more bookings and cancellations than resort type, reason may be resort hotels are more expensive.

City hotels needs to be focused more seeing the cancellation ratio, may be in maintenance,facilities,etc

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

Out[27]:
```
0    0.72025
1    0.27975
```

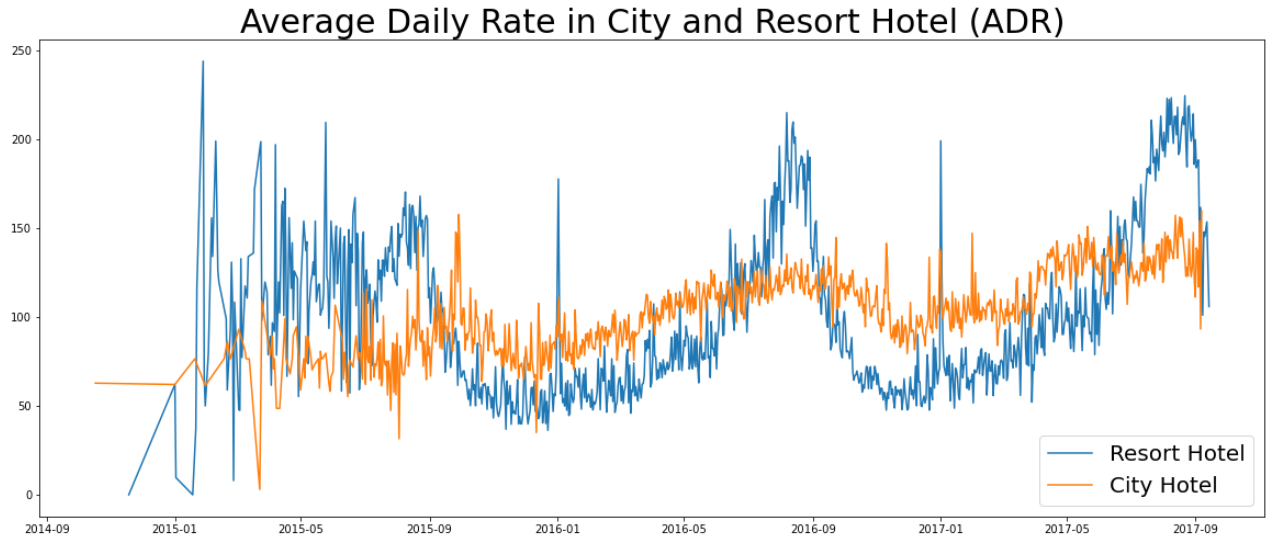
```
Name: is_canceled, dtype: float64
```

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

Out[28]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [29]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

In [30]:
```
plt.figure(figsize = (20,8))
plt.title('Average Daily Rate in City and Resort Hotel (ADR)',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()
```

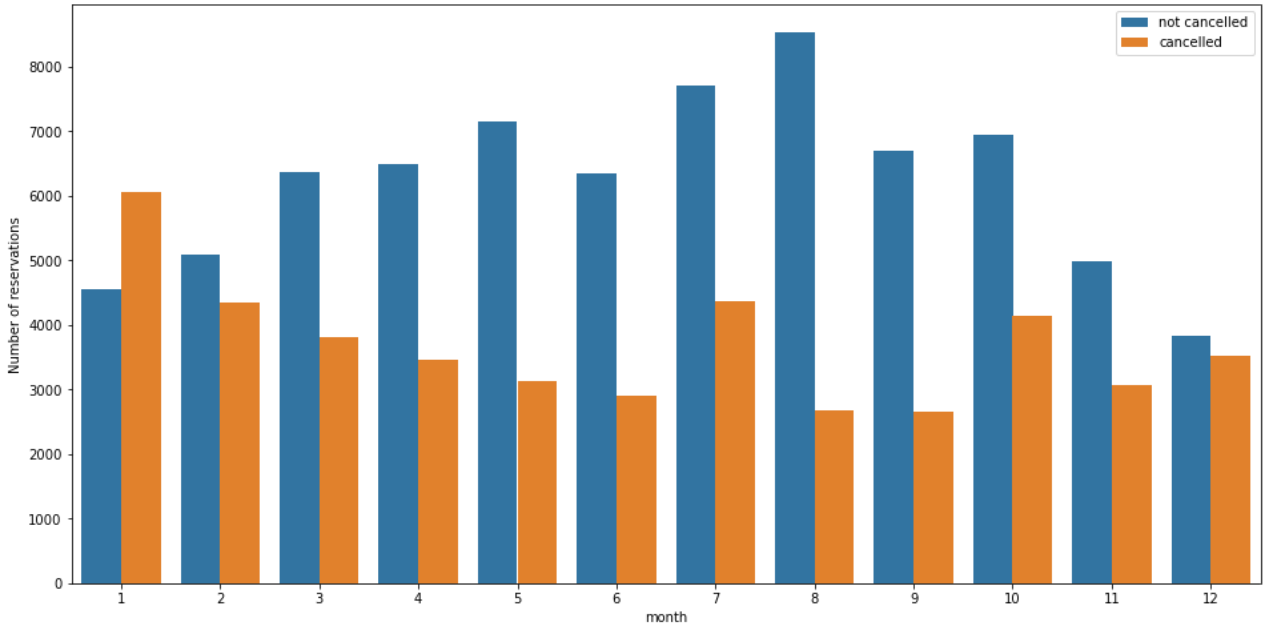


Findings :

1. City hotel is in mid (ADR), some spikes may be due to weekends,holidays
2. City hotel price < resort hotel price

Hence hypothesis proved

In [31]:
```
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 cancelled','cancelled'])
plt.show()
```



Findings :

1. Jan has more cancellations
2. Lowest cancellations in aug
3. Aug has most reservations
4. Dec jan has least reservations

Bit confusing!

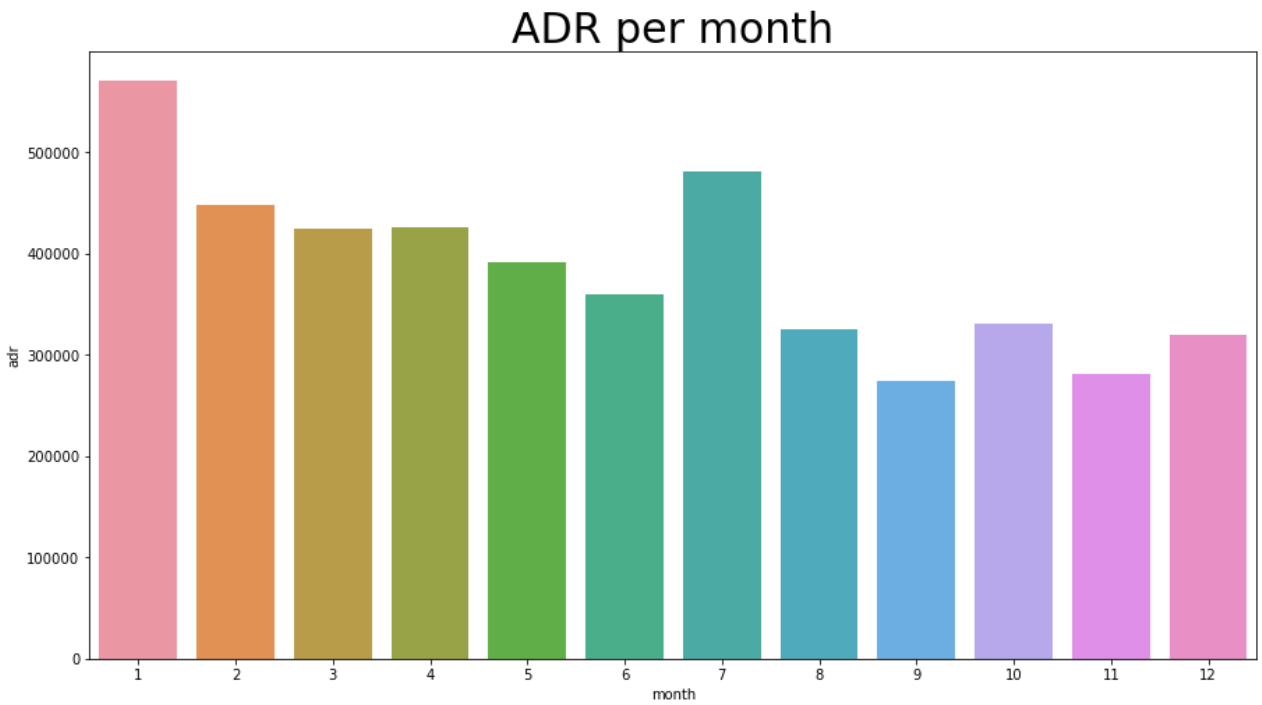
Aug has least cancellations and most reservations, may be the prices are low in aug so more reservations, or maybe the prices were too high, that people preferred cancellations

In [32]:

```
plt.figure(figsize=(15, 8))
plt.title('ADR per month', fontsize=30)

df_filtered = df[df['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_index()
sns.barplot(x='month', y='adr', data=df_filtered)

plt.show()
```



Findings :

ADR in aug is lowest

ADR in jan is highest

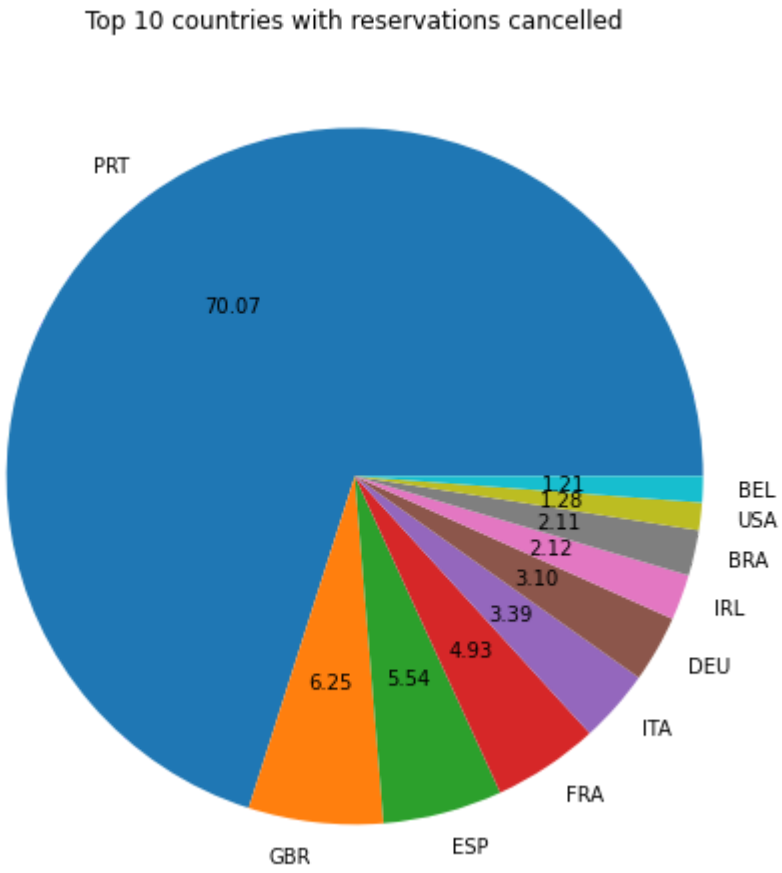
Hence proved our hypothesis that, when higher prices the cancellations will be more

In [33]:
```
cancelled_data = df[df['is_canceled']==1]
top_10_country = cancelled_data['country'].value_counts()[:10]
```

In [34]:
```
top_10_country
```

Out[34]:
```
PRT    27514
GBR     2453
ESP     2177
FRA     1934
ITA     1333
DEU     1218
IRL      832
BRA      830
USA      501
BEL      474
Name: country, dtype: int64
```

In [35]:
```
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()
```



Findings :

In portugal country, increase the facilities and decrease the prices, hold promotional campaigns, give discounts, advertisements, marketing

Now lets assume that the most of the customers are coming through offline agents, like see is it true?

In [36]:
```
df['market_segment'].value_counts()
```

Out[36]:
```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

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

Out[37]:
```
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
```

So we see that almost half of the customers are coming through online agents, thus proving our hypothesis wrong

In [38]:
```
cancelled_data['market_segment'].value_counts(normalize = 'True')
```

Out[38]:
```
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
```

Almost 47% customers coming through online agents cancel their booking, this may be due to reasons like, the online picture was very good and when the customer visited it may not be up to the mark due to lack of facilities, spaces, etc etc

So may be the hotels can portray the real picture in online ads to avoid this

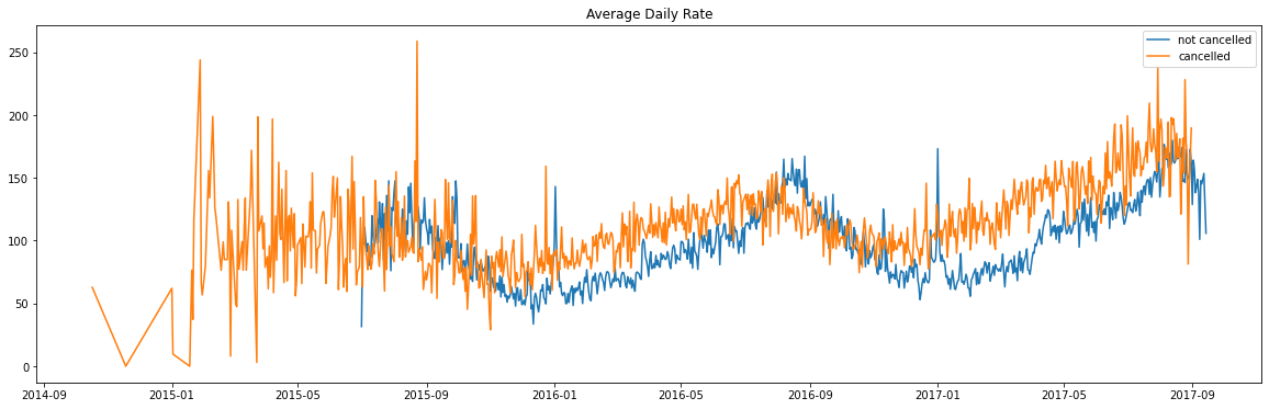
Now lets check the prices of cancelled bookings, are they high ?

In [39]:
```
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']].m
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'],
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label= '
plt.legend()
```

Out[39]:
```
<matplotlib.legend.Legend at 0x13cc7cbfcd0>
```

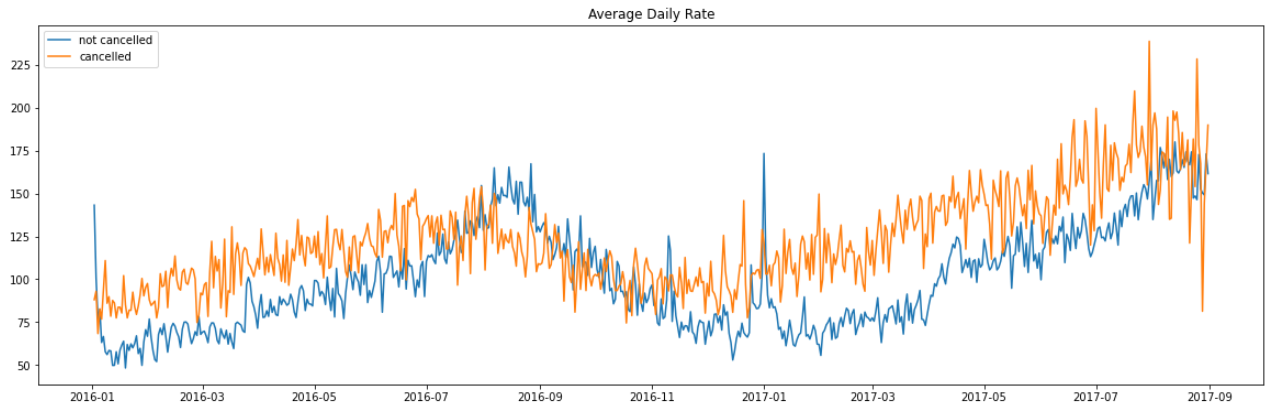


Now this may be due to inconsistent data

So lets take only data of 2016 and 17 till sept

In [40]:
```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] > '201
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_d
```

In [41]:
```
plt.figure(figsize =(20,6))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancelled_df_adr['adr'],
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label= '
plt.legend()
plt.show()
```



We see that the cancelled line(orange) is more than the not cancelled one

It proves that the prices have effect on cancellation

Therefore, ADR is the most influencing factor on the cancellation rate

In [ ]: