

Business Problem

In recent years, City Hotel and Resort Hotel have seen high cancellation rates. Each hotel is now dealing with a number of issues as a result, including fewer revenues and less than ideal hotel room use. Consequently, lowering cancellation rates is both hotels' primary goal in order to increase their efficiency in generating revenue, and for us to offer thorough business advice to address this problem.

The analysis of hotel booking cancellations as well as other factors that have no bearing on their business and yearly revenue generation are the main topics of this report.



Assumptions

1. No unusual occurrences between 2015 and 2017 will have a substantial impact on the data used.
2. The information is still current and can be used to analyze a hotel's possible plans in an efficient manner.
3. There are no unanticipated negatives to the hotel employing any advised technique.
4. The hotels are not currently using any of the suggested solutions.

5. The biggest factor affecting the effectiveness of earning income is booking cancellations.
6. Cancellations result in vacant rooms for the booked length of time.
7. Clients make hotel reservations the same year they make cancellations.

Research Question

1. What are the variables that affect hotel reservation cancellations?
2. How can we make hotel reservations cancellations better?
3. How will hotels be assisted in making pricing and promotional decisions?

Loading The DataSet

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

Exploratory Data Analysis and Cleaning

```
df.head()
```

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

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

	stays_in_weekend_nights	stays_in_week_nights	adults	...	
deposit_type \					
0	0	0	2	...	No
Deposit					
1	0	0	2	...	No
Deposit					
2	0	1	1	...	No
Deposit					
3	0	1	1	...	No
Deposit					
4	0	2	2	...	No
Deposit					

	agent	company	days_in_waiting_list	customer_type	adr	\
0	NaN	NaN	0	Transient	0.0	
1	NaN	NaN	0	Transient	0.0	
2	NaN	NaN	0	Transient	75.0	

3	304.0	NaN	0	Transient	75.0
4	240.0	NaN	0	Transient	98.0

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

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

[5 rows x 32 columns]

df.shape

(118897, 30)

Data Tranformation

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

```

```

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

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']
-----

```

Data Cleaning (Removing nulls)

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

hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	4
babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0
reservation_status	0
reservation_status_date	0
dtype: int64	

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

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

hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	0

```

babies                                0
meal                                  0
country                              0
market_segment                       0
distribution_channel                  0
is_repeated_guest                    0
previous_cancellations               0
previous_bookings_not_canceled       0
reserved_room_type                   0
assigned_room_type                   0
booking_changes                      0
deposit_type                         0
days_in_waiting_list                0
customer_type                        0
adr                                  0
required_car_parking_spaces          0
total_of_special_requests            0
reservation_status                   0
reservation_status_date              0
dtype: int64

```

Removing Outlier

```
df.describe()
```

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

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

	stays_in_weekend_nights	stays_in_week_nights	adults \
count	118898.000000	118898.000000	118898.000000
mean	0.928897	2.502145	1.858391
min	0.000000	0.000000	0.000000

25%	0.000000	1.000000	2.000000
50%	1.000000	2.000000	2.000000
75%	2.000000	3.000000	2.000000
max	16.000000	41.000000	55.000000
std	0.996216	1.900168	0.578576

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

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

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

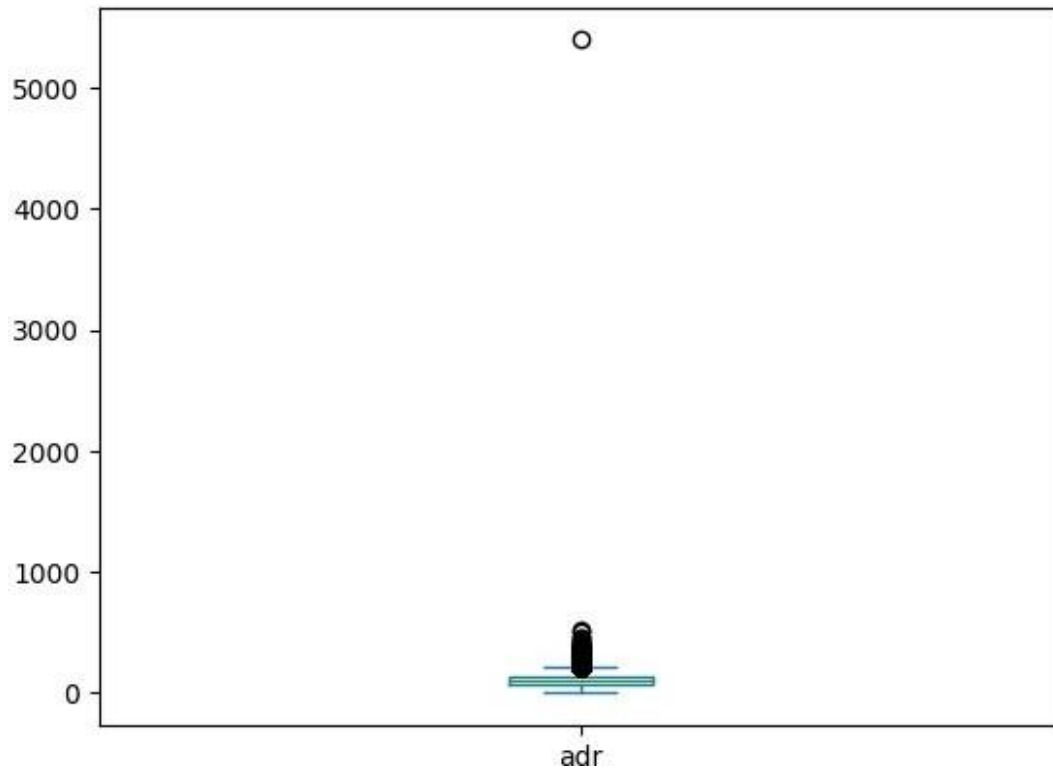
	required_car_parking_spaces	total_of_special_requests \
count	118898.000000	118898.000000
mean	0.061885	0.571683
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	1.000000
max	8.000000	5.000000
std	0.244172	0.792678

	reservation_status_date
count	118898
mean	2016-07-30 07:37:53.336809984
min	2014-10-17 00:00:00

```
25%          2016-02-02 00:00:00
50%          2016-08-08 00:00:00
75%          2017-02-09 00:00:00
max          2017-09-14 00:00:00
std          NaN
```

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

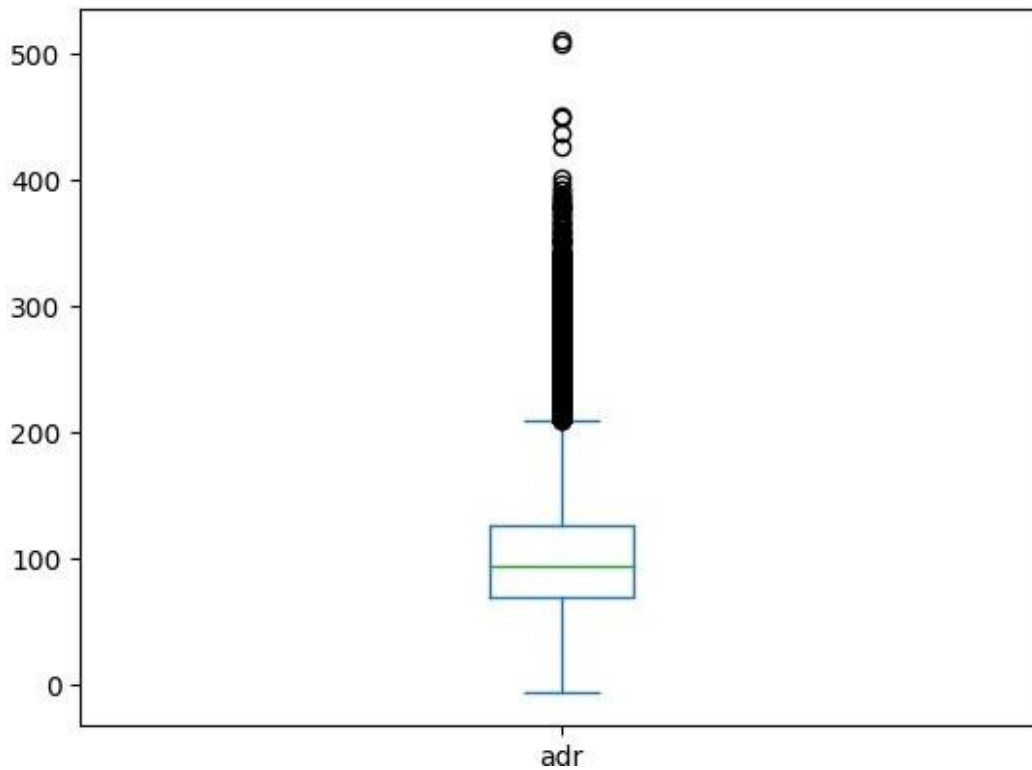
```
<Axes: >
```



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

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

```
<Axes: >
```

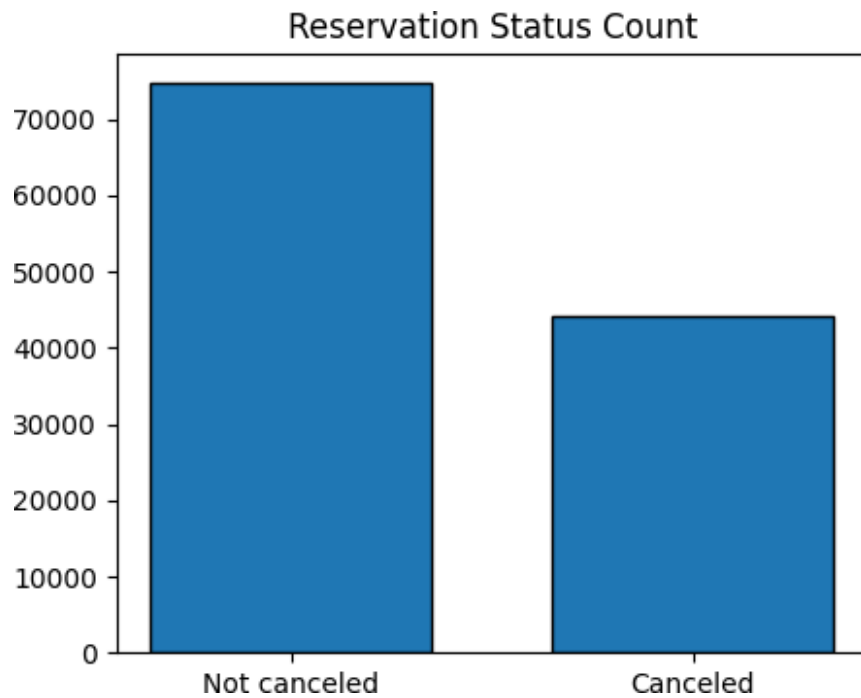


Data Analysis and Visualization

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

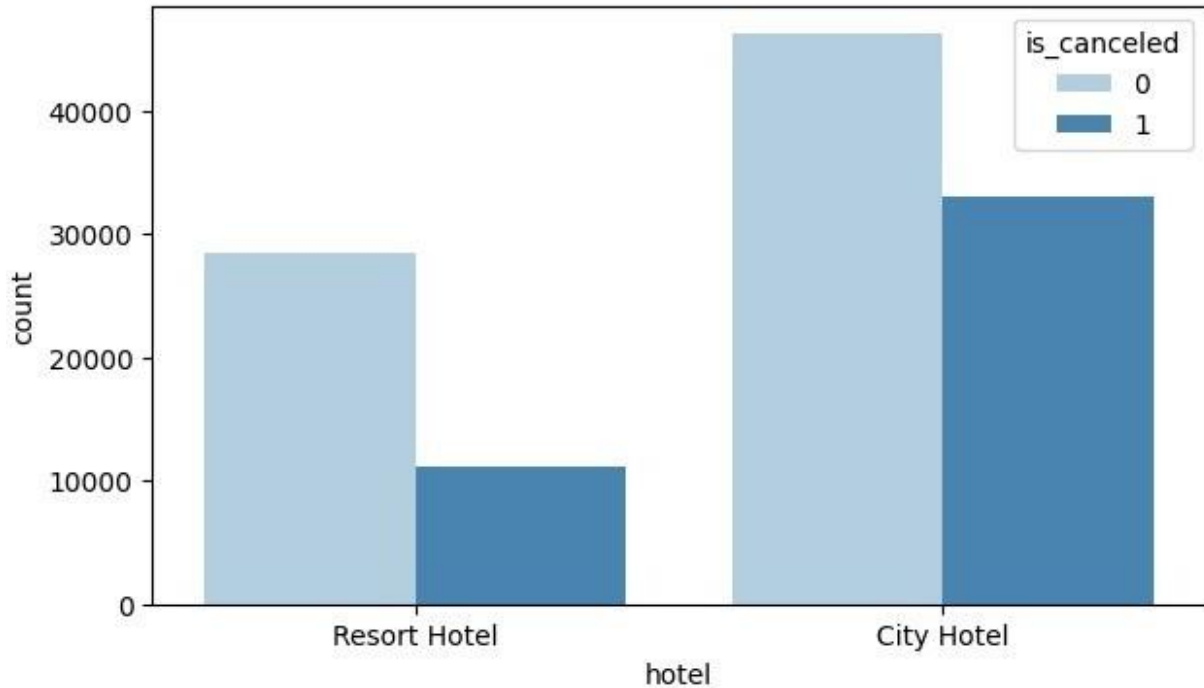
plt.figure(figsize=(5,4))
plt.title('Reservation Status Count')
plt.bar(['Not canceled', 'Canceled'],df['is_canceled'].value_counts(),
edgecolor = 'k' , width = 0.7)

<BarContainer object of 2 artists>
```



```
plt.figure(figsize=(7,4))
ax1= sns.countplot(x = 'hotel', hue = 'is_canceled', data = df,
palette='Blues')
legend_labels,_ = ax1. get_legend_handles_labels()
plt.title('Reservation status in different hotels', size = 20)
plt.xlabel = ('Hotel')
plt.ylabel = ('Number of reservations')
```

Reservation status in different hotels



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

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

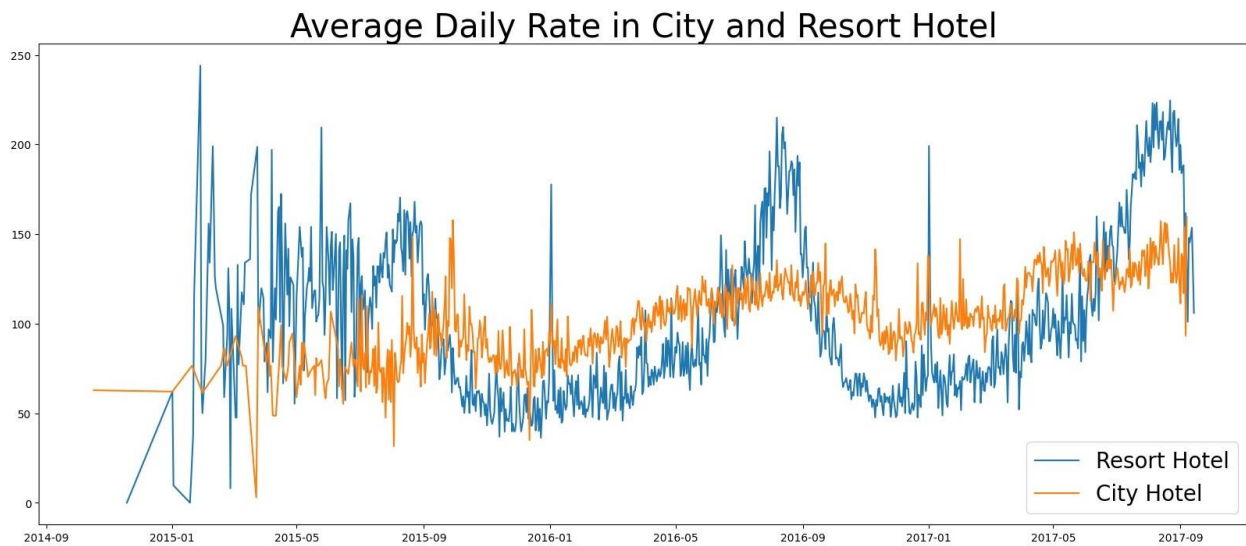
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)

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

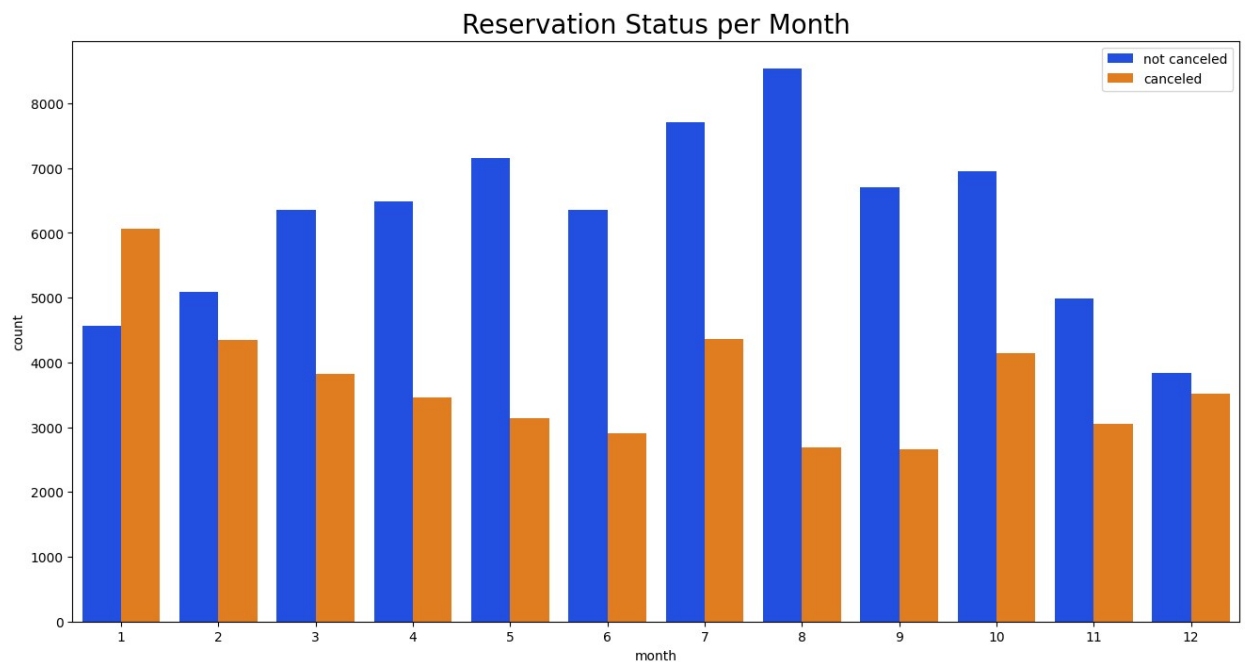
city_hotel = city_hotel.groupby('reservation_status_date')
[['adr']].mean()
resort_hotel = resort_hotel.groupby('reservation_status_date')
[['adr']].mean()

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')
```

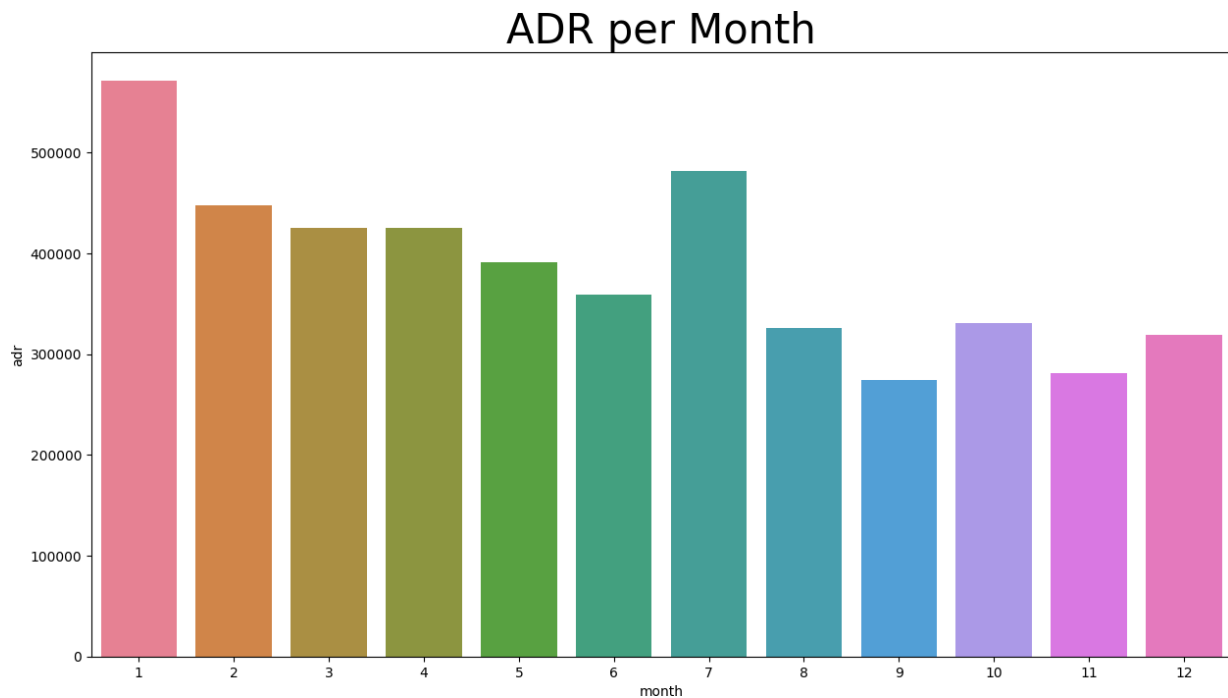
```
plt.legend(fontsize = 20)
plt.show()
```



```
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()
plt.title('Reservation Status per Month' , size = 20)
plt.legend(['not canceled', 'canceled'])
plt.show()
```



```
plt.figure(figsize=(15,8))
plt.title('ADR per Month' , size = 30)
palette = sns.color_palette("husl", len(df[df['is_canceled']==1]
['month'].unique()))
sns.barplot(x='month', y='adr',hue='month',
data=df[df['is_canceled']==1].groupby('month')
[['adr']].sum().reset_index(), palette=palette, legend=False)
plt.show()
```



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

Top 10 countries with reservation cancelled

