Business Problem

Over the past few years, both City Hotel and Resort Hotel have faced a surge in booking cancellations.This trend has led to several challenges for each hotel, including reduced revenue and underutillized rooms. As a result, both hotels are now focused on bringing down cancellation rates to improve their revenue efficiency. Our goal is to provide well-rounded business recommendations to help tackle this issue.

This project explores the patterns behind hotel booking cancellations and other unrelated factors, aiming to understand their impact on business performance and yearly revenue.

Assumptions

- No major unexpected events between 2015 and 2017 have significantly influenced the data being used.
- The data remains relevant and can still effectively support the analysis of potential strategies for the hotels.
- It's assumed that any recommended approach will not bring unforeseen negative consequences for the hotels.
- The hotels have not yet implemented any of the strategies being proposed in this project.
- The most significant challenge to maintaining steady income is the high rate of booking cancellations.
- When bookings are canceled, the reserved rooms typically remain unoccupied for the entire duration they were initially booked for.
- Guests usually make and cancel their reservations withing the same calender year.

Research Questions

1. What factors contribute to hotel reservation cancellations?
2. What strategies can be implemented to reduce the rate of cancellations?
3. How can this analysis support hotels in making smarter pricing and promotional decisions?

Hypothesis

1. Guests are more likely to cancel their reservations when room prices are higher.
2. A longer waiting list often leads to an increase in customer cancellations.
3. Most hotel bookings are being made through offline travel agents.

Importing Libraries

In [445…
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading Dataset

In [447…
```
df = pd.read_csv('hotel_booking.csv')
df.head()
```

Out[447…

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
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 [448…
```
df. tail()
```

Out [448…

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_nur
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

Data Cleaning

In [450…

```
df.shape
```

Out[450…

```
(119390, 36)
```

In [451…

```
df.columns
```

Out[451…

```
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', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

In [452…

```
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
```

Changing the reservation_status_date into datetime.

In [454…
```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
```

In [455…
```
df['reservation_status_date'].dtype
```

Out[455…
```
dtype('<M8[ns]')
```

In [456…
```
df.describe(include = 'object')
```

Out[456…

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserv
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 [457…

```
# Printing the unique values in categorical columns
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' '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']
--------------------------------------------------
name
['Ernest Barnes' 'Andrea Baker' 'Rebecca Parker' ... 'Wesley Aguilar'
 'Caroline Conley MD' 'Ariana Michael']
--------------------------------------------------
email
['Ernest.Barnes31@outlook.com' 'Andrea_Baker94@aol.com'
 'Rebecca_Parker@comcast.net' ... 'Mary_Morales@hotmail.com'
```

```
 'MD_Caroline@comcast.net' 'Ariana_M@xfinity.com']
--------------------------------------------------
phone-number
['669-792-1661' '858-637-6955' '652-885-2745' ... '395-518-4100'
 '531-528-1017' '422-804-6403']
--------------------------------------------------
credit_card
['************4322' '************9157' '************3734' ...
 '************9170' '************6349' '************7959']
--------------------------------------------------
```

In [458…

```
# Checking missing values
df.isnull().sum()
```

Out[458…

```
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
name                                   0
email                                  0
phone-number                           0
credit_card                            0
dtype: int64
```

Now here we drop the columns which we don't need for our analysis, since we don't have anything to do with name, email, phone-number and credit-card columsn we simply drop them. column 'company' and 'agent' have null values we cannot handle so we drop them as well. Also in column 'country' we have only 488 null values which are not even 1% of dataset so we simply drop the null values.

In [460…
```
df.drop(['name', 'email', 'phone-number', 'credit_card', 'company', 'agent'], axis = 1,
df.dropna(inplace = True)
```

In [461…
```
df.isnull().sum()
```

Out[461…
```
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 [462…
```
df.describe()
```

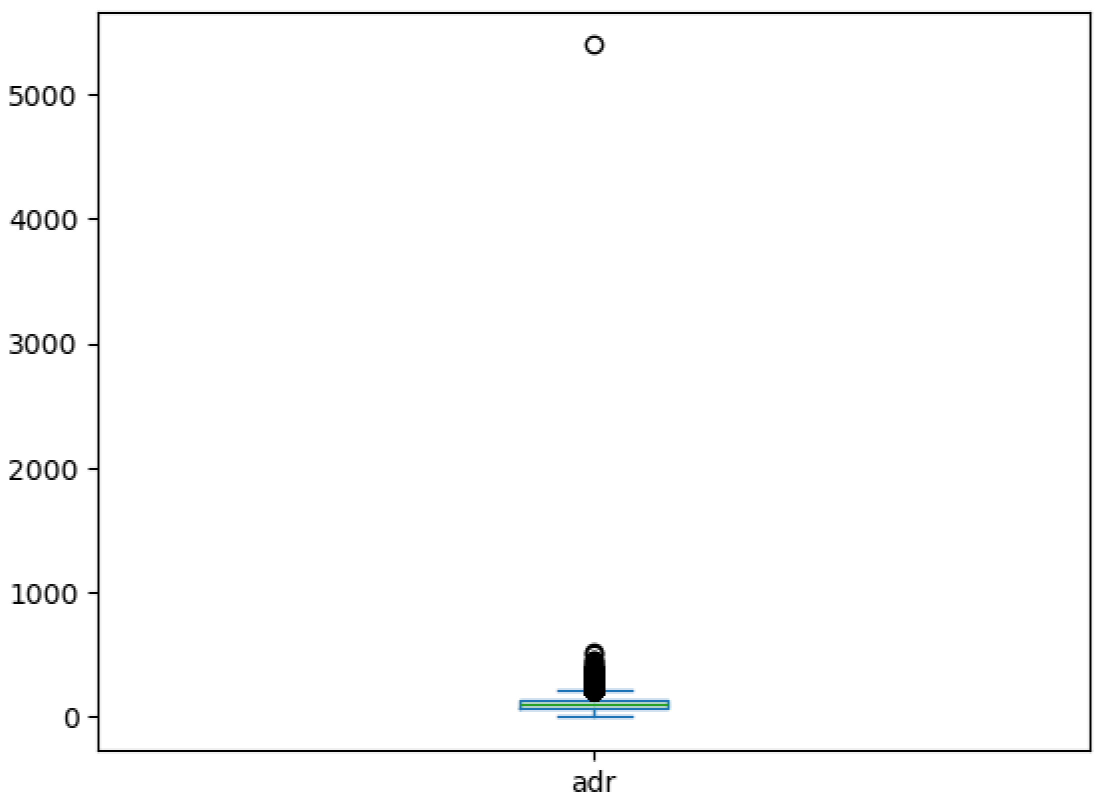
Out [462…

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_da
count	118898.000000	118898.000000	118898.000000	118898.000000	118
mean	0.371352	104.311435	2016.157656	27.166555	
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	
std	0.483168	106.903309	0.707459	13.589971	

In [463…

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

Out [463… <Axes: >



In [464…

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

Exploratory Data Analysis

In [466…

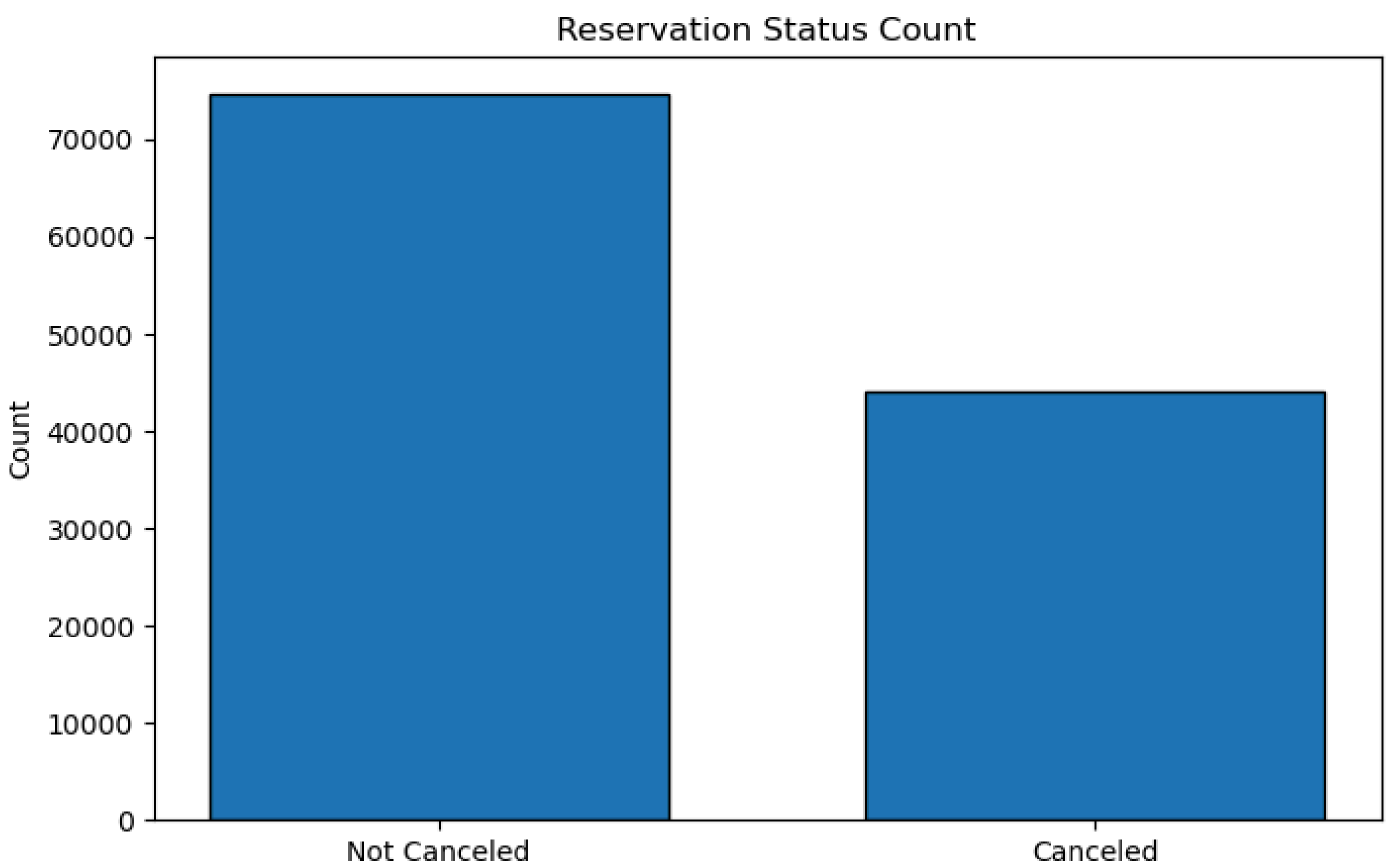
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
cancelled_perc
```

Out[466…
```
is_canceled
0    0.628653
1    0.371347
Name: proportion, dtype: float64
```

In [467…
```
print("Total value in percentage of no hotel cancellation is around: 62.86%")
print("Total value in percentage of hotel cancellation is around: 37.13%")
```

```
Total value in percentage of no hotel cancellation is around: 62.86%
Total value in percentage of hotel cancellation is around: 37.13%
```

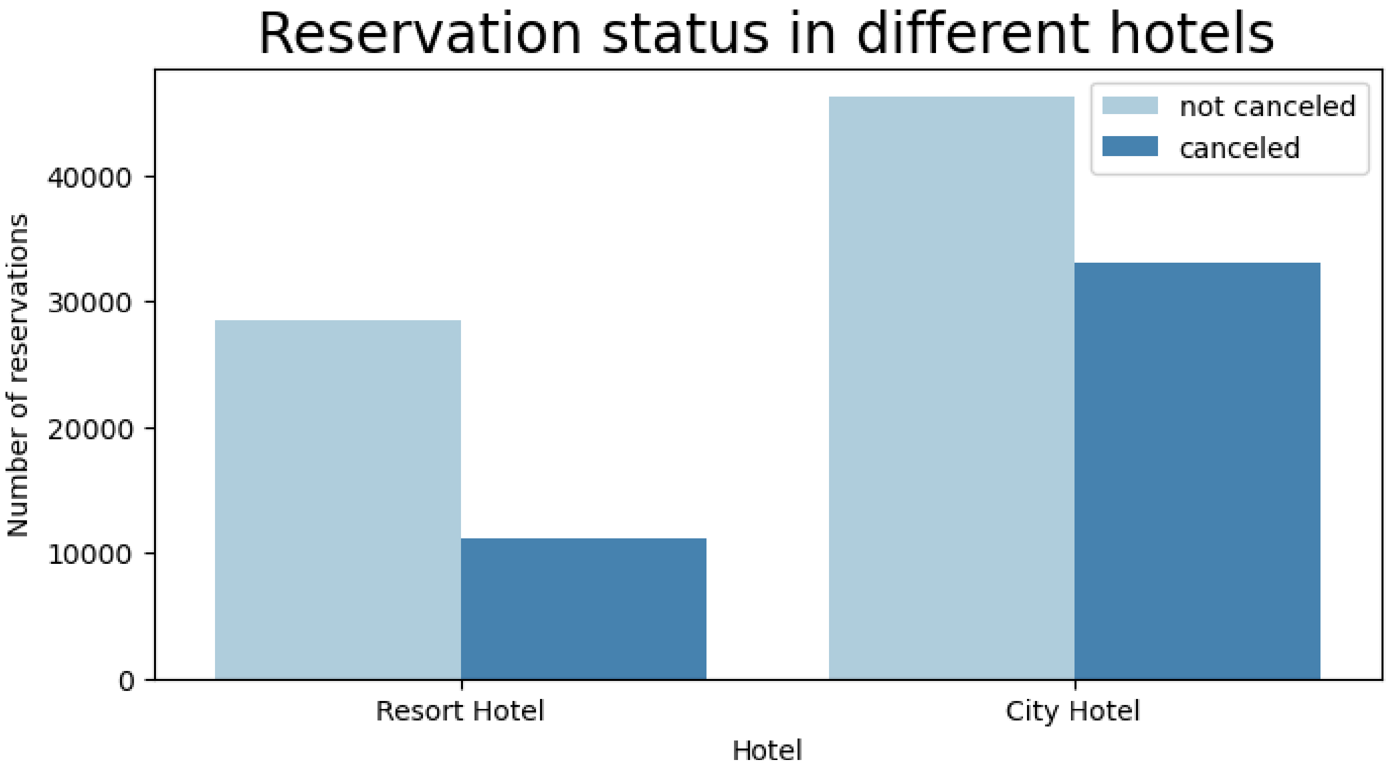
In [468…
```
plt.figure(figsize= (8,5))
plt.bar(['Not Canceled', 'Canceled'], df['is_canceled'].value_counts(), edgecolor = 'k',
plt.title('Reservation Status Count')
plt.ylabel('Count')
plt.show()
```



The bar graph illustrates the proportion of reservations that were canceled versus those that were not. It's clear that a large portion of bookings remained active. However, 37% of clients still ended up canceling their reservations, which has a noticeable impact on the hotel's overall revenue.

In [470…
```
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")
```

```
plt.legend(['not canceled', 'canceled'])
plt.show()
```



Compared to resort hotels, city hotels tend to receive more bookings. This could be because resort hotels are generally priced higher than those located in urban areas.

In [472…
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

Out[472…
```
is_canceled
0    0.72025
1    0.27975
Name: proportion, dtype: float64
```

In [473…
```
print("So out of total hotel cancellation resort hotels cancellation percentage is : 27.
```

```
So out of total hotel cancellation resort hotels cancellation percentage is : 27.97%
```

In [474…
```
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

Out[474…
```
is_canceled
0    0.582918
1    0.417082
Name: proportion, dtype: float64
```

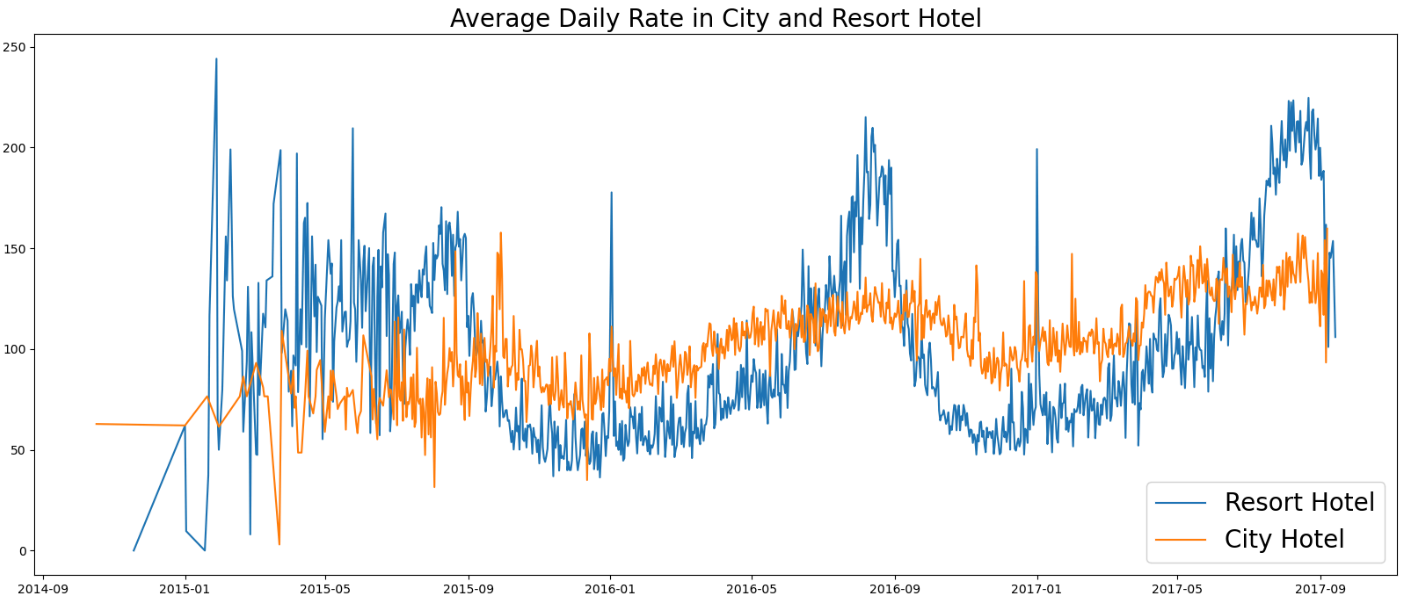
In [475…
```
print("The city hotel cancellation out of total cancellation is : 41.70%")
```

```
The city hotel cancellation out of total cancellation is : 41.70%
```

In [476…
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

In [477…

```
plt.figure(figsize = (20,8))
plt.title('Average Daily Rate in City and Resort Hotel', fontsize = 20)
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()
```



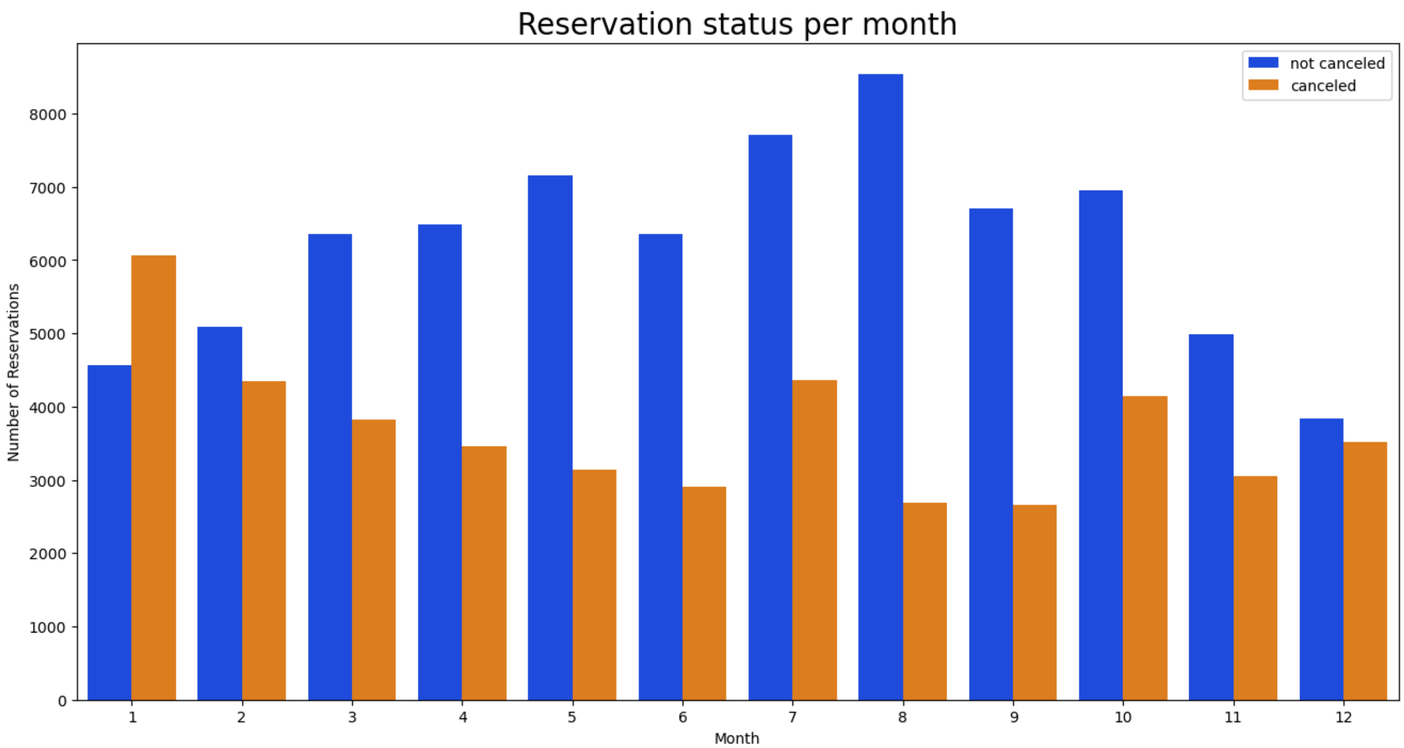
The line graph illustrates that, on some days, city hotels have a lower average daily rate compared to resort hotels—and on other days, the difference is even more noticeable. Naturally, weekends and holidays tend to drive up prices at resort hotels.

In [479…

```
df['month'] = df['reservation_status_date'].dt.month
```

In [480…

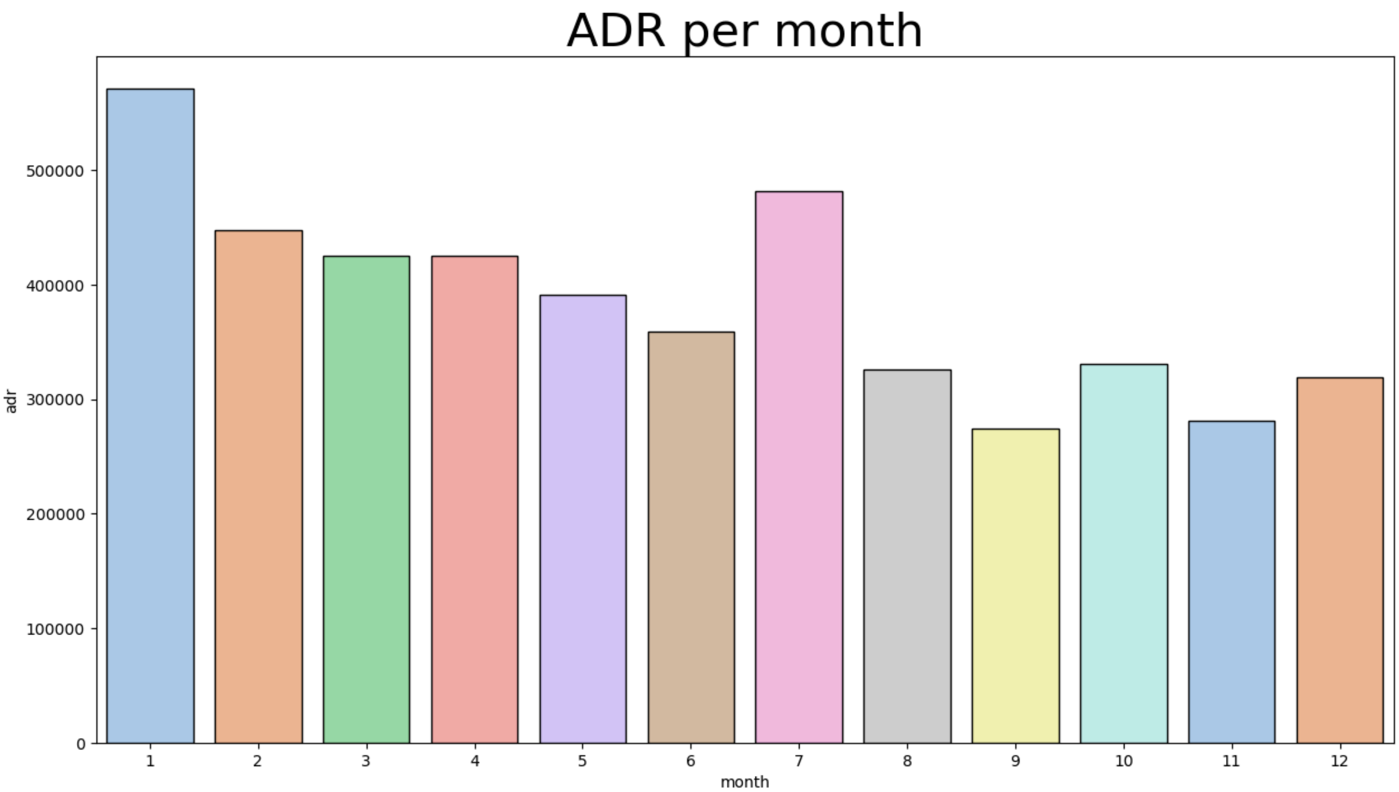
```
plt.figure(figsize = (16,8))
ax= sns.countplot(x = 'month', hue = 'is_canceled', data = df, palette = 'bright')
legend_labels,_ = ax.get_legend_handles_labels()
ax.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per month', size = 20)
plt.xlabel('Month')
plt.ylabel('Number of Reservations')
plt.legend(['not canceled', 'canceled'])
plt.show()
```



We created a grouped bar chart to explore which months had the highest and lowest reservation activity based on booking status. The data reveals that August had the most confirmed and canceled reservations overall, while January recoreded the highest number of cancellations specifically.

In [482…

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

plt.figure(figsize = (15,8))
plt.title('ADR per month', fontsize = 30)
sns.barplot(x = 'month', y = 'adr', data = dff, palette = 'pastel', edgecolor = 'k')
plt.show()
```

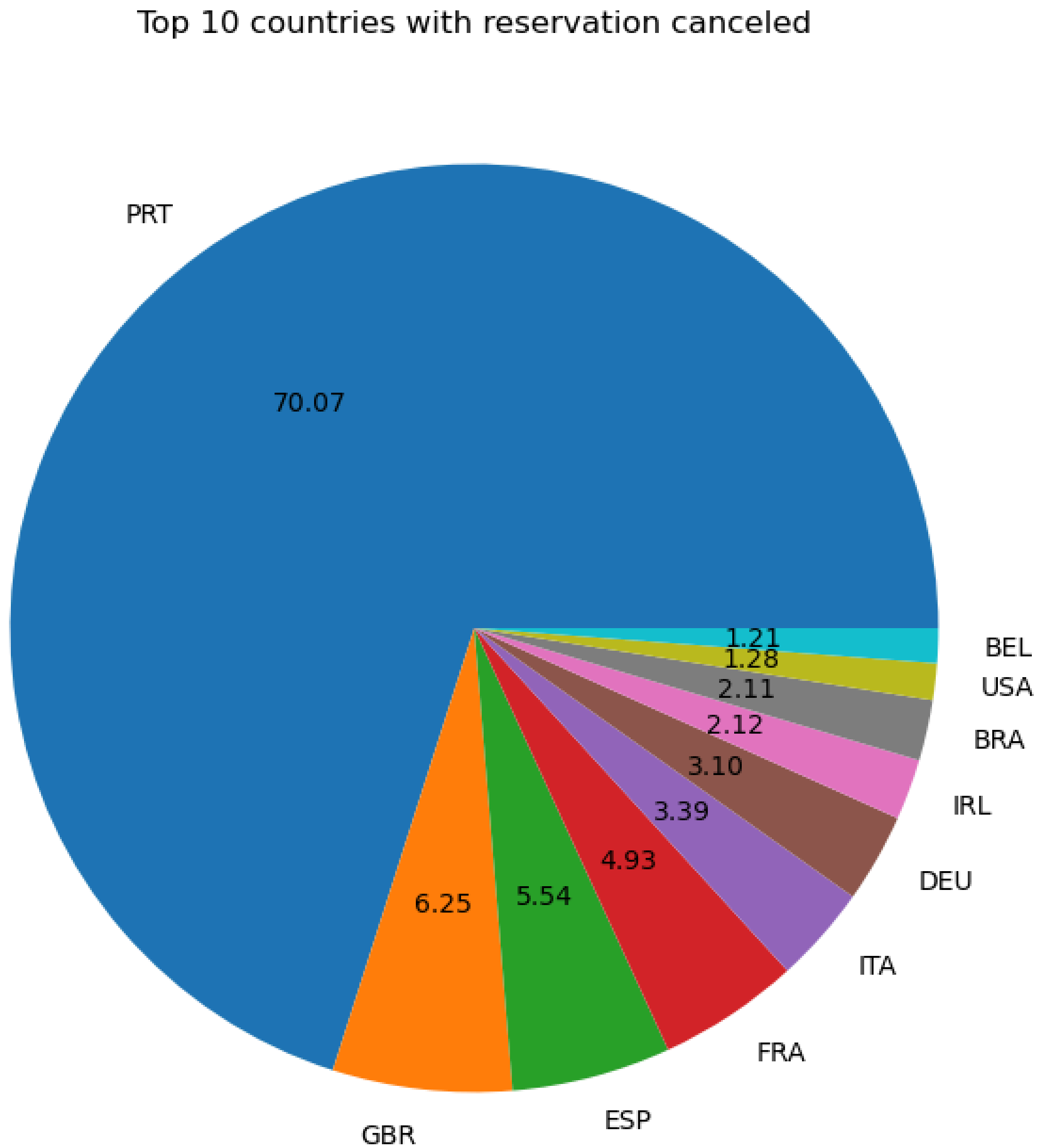


The bar chart indicates that cancellations are most frequent when prices are at their highest and drop significantly when prices are lower. This suggests that accommodation cost plays a major role in guests' decisions to cancel.

Now, turning to geographic trends—Portugal stands out as the country with the highest number of canceled reservations.

In [484…

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



In [485…

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

Out[485…

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

In [486…

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

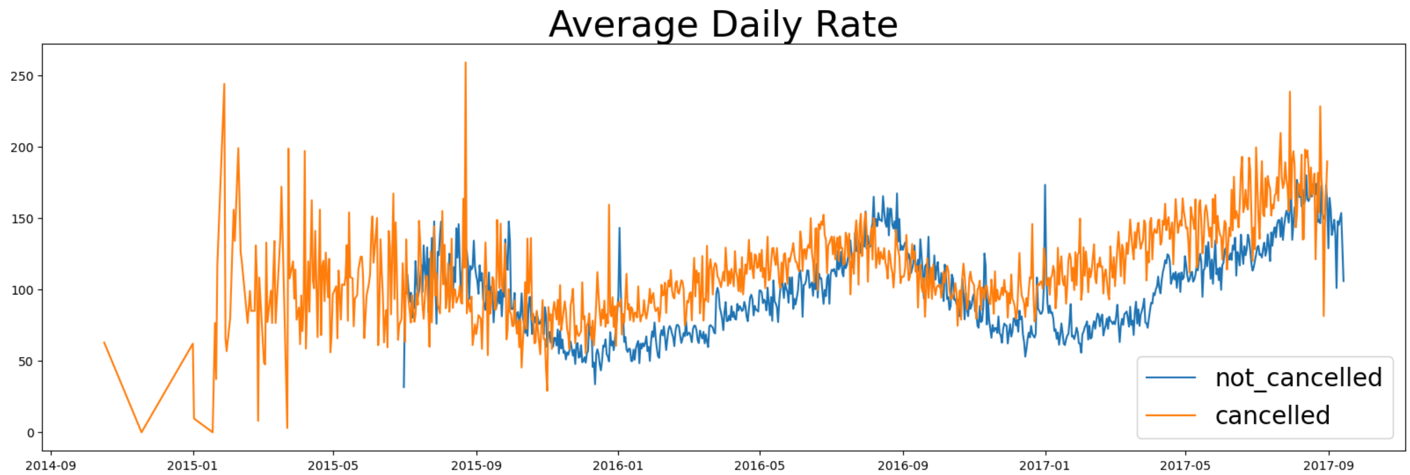
Out[486…
```
market_segment
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: proportion, dtype: float64
```

In [487…
```
cancelled_data['market_segment'].value_counts(normalize = True)
```

Out[487…
```
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
```

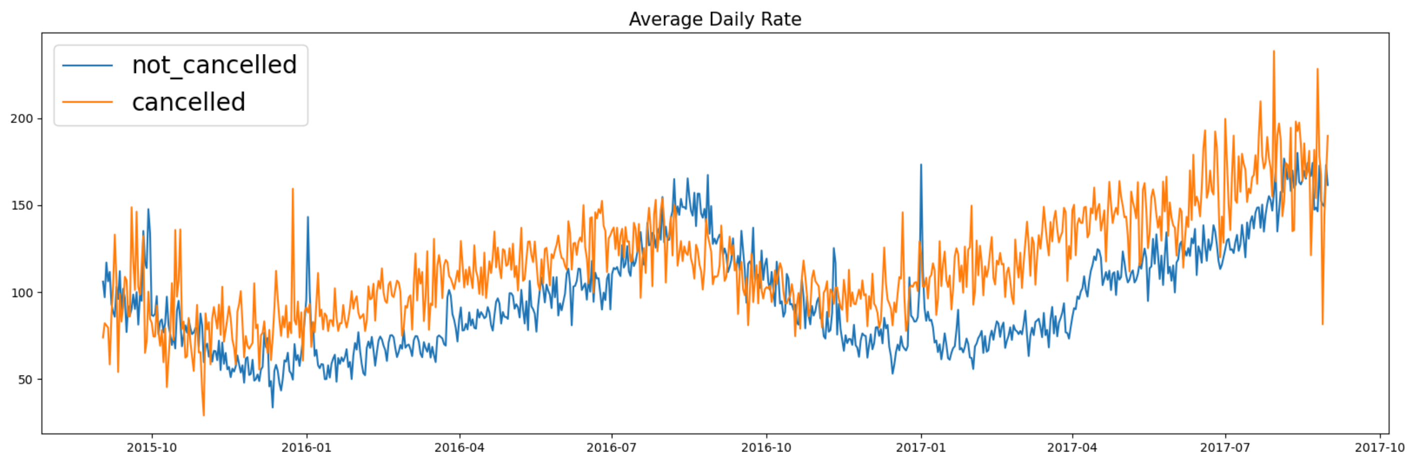
Let's take a look at how guests are making their hotel bookings—whether through direct channels, group bookings, or travel agencies, both online and offline. About 46% of clients book through online travel agencies, while 27% make reservations as part of a group. Interestingly, only 4% of guests make direct bookings by physically visiting the hotel.

In [489…
```
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']].me
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', size = 30)
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancelled_df_adr['adr'], l
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label = '
plt.legend(fontsize = 20)
plt.show()
```




```
In [490…
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] >'2015-
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_da
```

```
In [491…
plt.figure(figsize = (20,6))
plt.title('Average Daily Rate', size = 15)
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancelled_df_adr['adr'], l
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label = '
plt.legend(fontsize = 20)
plt.show()
```



As shown in the graph, cancellations tend to happen more often when the average daily rate is higher compared to times when bookings are not canceled. This clearly supports the earlier analysis that higher prices are closely linked to increased cancellation rates.

Suggestions

- Since cancellation rates tend to rise with price increases, hotels should revisit their pricing strategies. Adjusting rates based on location and offering targeted discounts could help reduce cancellations and attract more bookings.
- City hotels have a higher cancellation rate compared to Resort hotels. To address this, offering special discounts on weekends or holidays may encourage guests to follow through with their bookings.

- Given that January sees the higher number of cancellations, hotels could launch promotional campaigns or marketing efforts during this month to boost reservations and offset potential losses.
- Hotels—especially those in Portugal—should focus on enhancing service quality and overall guest experience to help lower the cancellation rate.