Problem-: Hotel Booking End to End Project

In recent years ,City hotels and resorts 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 through business advice to address this problem.

Steps taken to perform data analysis project-:

1. Create a problem statement
2. Identify the data you want to analyse
3. Explore and clean the data
4. Analyse the data to get useful insights(Hypothesis and research analysis can be performed here)
5. Presenting the data in terms of report and dashboard using Visualisation

Note-:
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.

Research Questions-:

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

Hypothesis-:

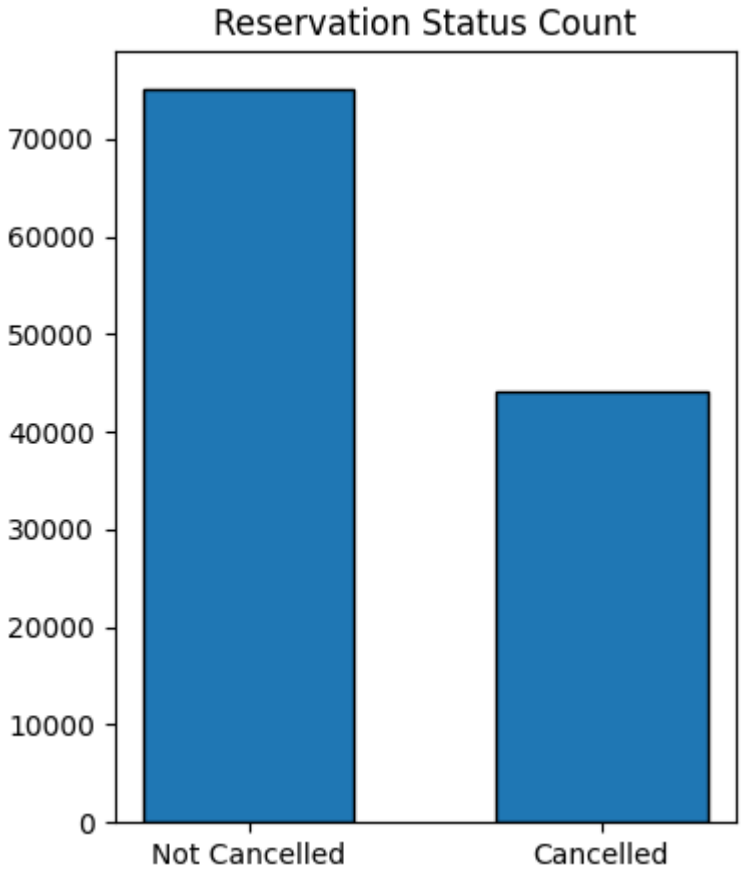
(A hypothesis states what will be the result of your predictions.)

1-More cancellations occur when prices are higher
2-When there is a longer waiting list,customers tend to cancel more frequently
3-The majority of clients are coming from offline travel agents to make their reservations.

Assumptions-:

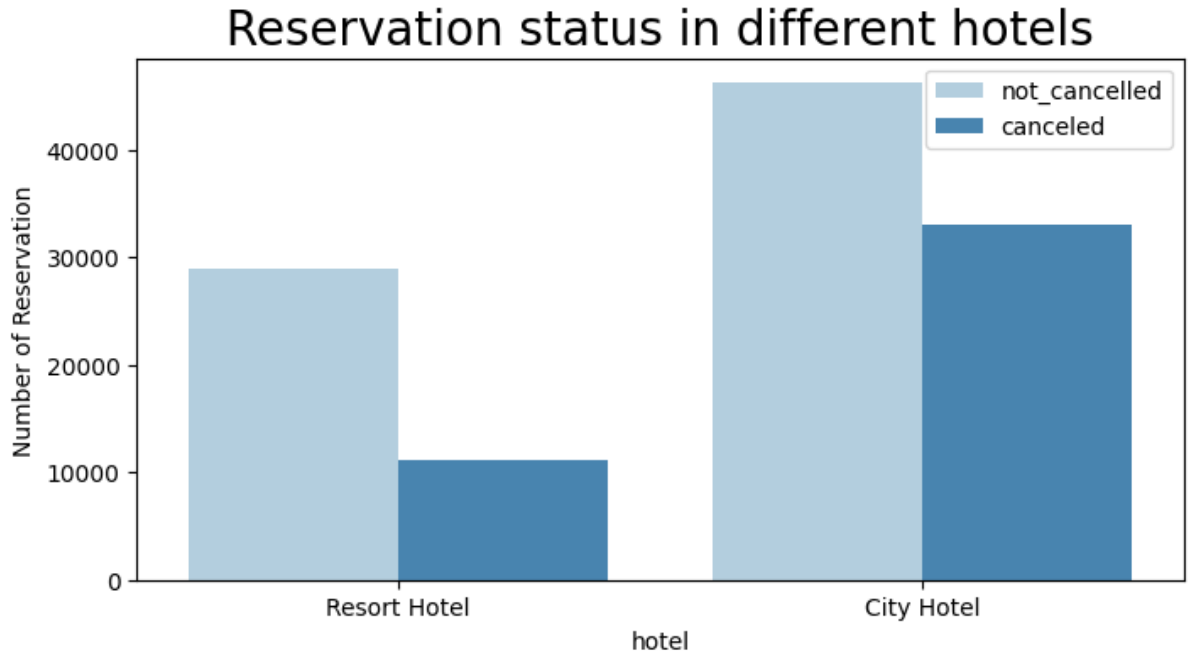
1-No unusual occurrences between 2015 and 2017 will have a substantial impact on the data used.
2-The information is still current and can be analyse a hotels possible plans in an efficient manner
2-There are no unanticipated negatives to the hotel employing any advised techniques. \
4-The hotels are not currently using any of the suggested solutions
5-The biggest factors affecting the effectiveness of earning income is booking,cancellation
6-Cancellations result in vacant rooms for the booked length of time
7-Clients make hotels reservation the same year they make cancellations

Analysis and Findings—:

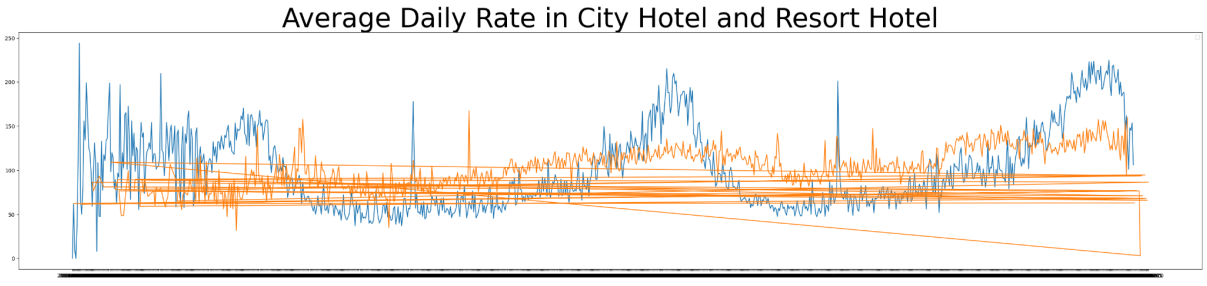


The accompanying bar graph shows the percentage of reservations that are canceled and those that are not. It is obvious that there are still a significant number of reservations that

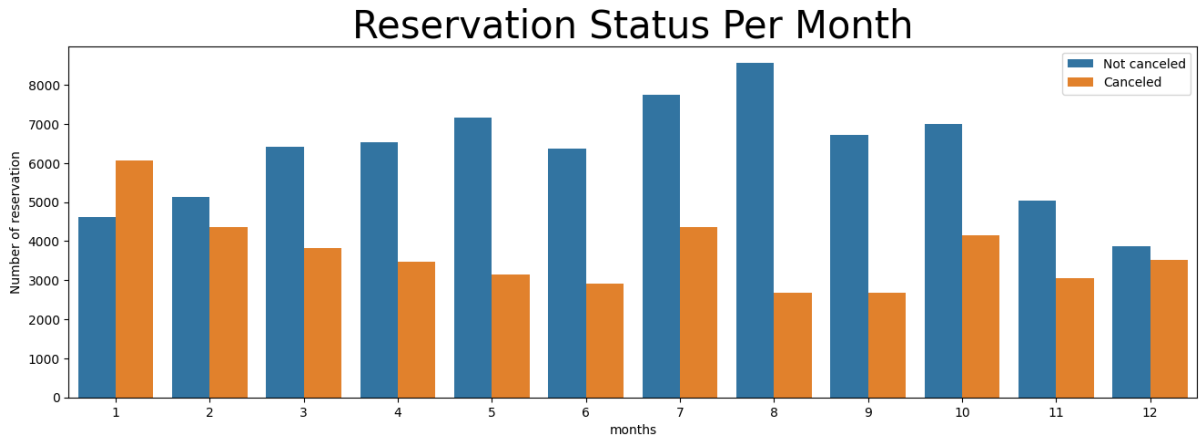
have not been cancelled. There are still a significant number of reservations that have not been cancelled. There are still 37 % of clients who cancelled their reservations, which has a significant impact on the hotels earnings.



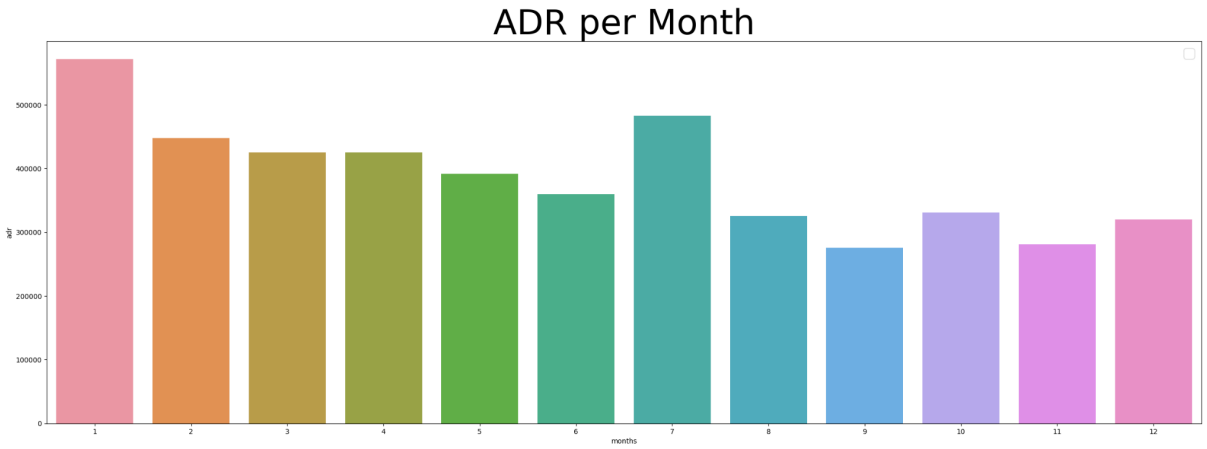
In comparison to resort hotels, city hotels have more bookings. It's possible that resort hotels are more expensive than those in cities.



The above line graph shows that, on certain days the average daily rate (ADR) for a city hotel is less than that of a resort hotel, and on other days, it is even less. It goes without saying that weekends and holidays may see a rise in resort hotel rates.

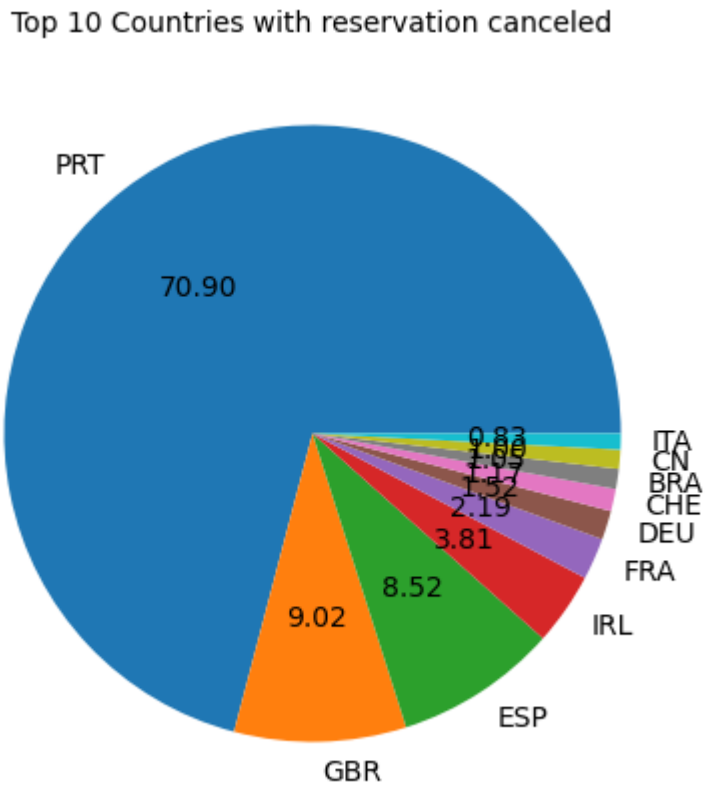


Here developed the grouped bar graph to analyse the months with highest and lowest reservations levels according to reservation status. As can be seen both the number of confirmed reservations and the number of cancelled reservations are largest in the month of august. Whereas January is the month with the most cancelled reservations.



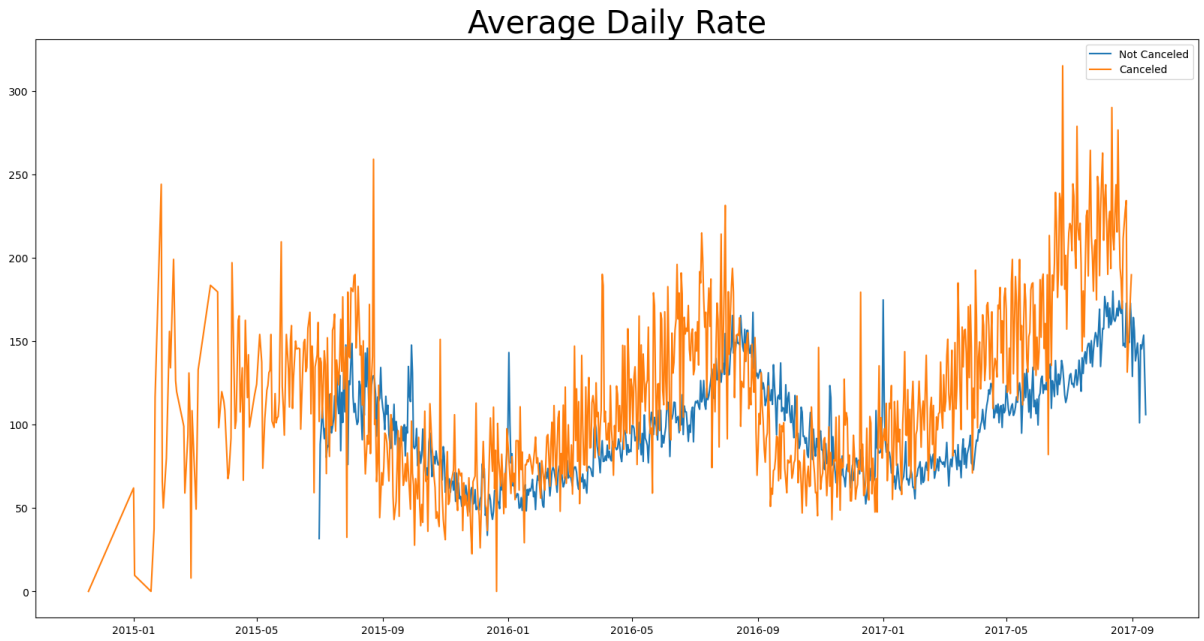
This bar graph demonstrates that cancellations are most common when prices are greatest and least common when they are lowest. Therefore ,the cost of the accommodations is solely responsible for the cancellation.

Now let's see which country has the highest reservations cancelled. Among top 10 countries Portugal with the highest number of cancellations.



Let's check the area from the guests who are visiting the hotels and making reservations. It is coming from Directs or Groups, Online or Offline Travel Agents ? Around 46% of the clients come from online travel agencies, whereas 27 % come from groups. Only 4% of the clients book hotels directly by visiting them and making reservations.

+



As seen in the graph ,reservations are cancelled when the average daily rate is higher than when it is not cancelled. It clearly proves all the above analysis, that the higher price leads to higher cancellation.

Suggestions-:

Cancellation rates rise as the price does. In order to prevent cancellations of reservations hotels could work on their pricing strategies and try to lower the rates for specific hotels based on locations. They can also provide some discounts to the consumers.

As the ratio of cancellations and not cancellations of the resort hotels is higher in the reost hotel than the city hotels . so the hotels should provide reasonable discount on the room prices on weekends or on holidays

In the month of January hotels can start campaigns or marketing with a reasonable amount to increase their revenue as the cancellations are the highest in this month.

They can also increase the quality of their services mainly in Portugal to reduce the cancellation rate.

Link for code editor

Snapshots of code editor-:

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

.oading the data

```
df=pd.read_csv('hotel_booking.csv', on_bad_lines='skip')
df['hotel'].value_counts()
```

```
City Hotel      79330
Resort Hotel    40060
Name: hotel, dtype: int64
```

Exploratory data analysis and data cleaning-:

```
pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns',None)
```

```
#to know how many rows are there in the data set  use shape
df.shape
```


```
#to know how many rows are there in the data set  use shape
df.shape
```


```
(29652, 36)
```

```
df.columns
```

```
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 above data zero represent that cancellation is not happend 1 represent that user has cancelled the booking
df.isna().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
```

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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29652 entries, 0 to 29651
Data columns (total 36 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           29652 non-null  object
 1   is_canceled                     29652 non-null  int64
 2   lead_time                       29652 non-null  int64
 3   arrival_date_year               29652 non-null  int64
 4   arrival_date_month              29652 non-null  object
 5   arrival_date_week_number        29652 non-null  int64
 6   arrival_date_day_of_month       29652 non-null  int64
 7   stays_in_weekend_nights         29652 non-null  int64
 8   stays_in_week_nights            29652 non-null  int64
 9   adults                          29652 non-null  int64
 10  children                        29652 non-null  float64
 11  babies                          29652 non-null  int64
 12  meal                            29652 non-null  object
 13  country                         29191 non-null  object
 14  market_segment                  29652 non-null  object
 15  distribution_channel            29652 non-null  object
 16  is_repeated_guest               29652 non-null  int64
 17  previous_cancellations          29652 non-null  int64
 18  previous_bookings_not_canceled  29652 non-null  int64
 19  reserved_room_type              29652 non-null  object
 20  assigned_room_type              29652 non-null  object
 21  booking_changes                 29652 non-null  int64
 22  deposit_type                    29652 non-null  object
 23  agent                           23488 non-null  float64
```

```
df.describe(include='object') # these are just to check categorical values.
```

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	c
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	
unique	2	12	5	177	8	5	10	12	3	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	

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

```
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'
```

```
5] df.describe(include='object').columns
```

```
Index(['hotel', 'arrival_date_month', 'meal', 'country', 'market_segment',
       'distribution_channel', 'reserved_room_type', 'assigned_room_type',
       'deposit_type', 'customer_type', 'reservation_status', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

```
6] df.describe()
```

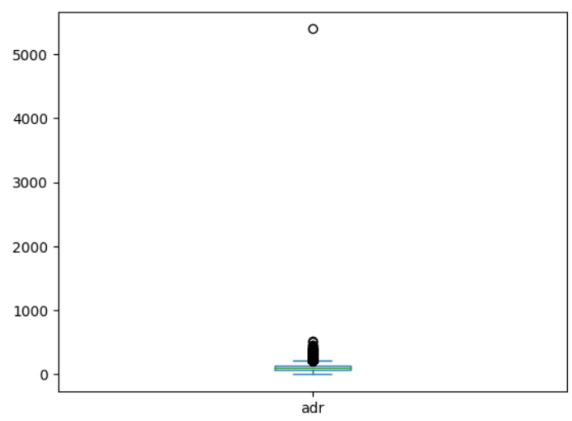
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_wee
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	11939
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	5

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

```
<Axes: >
```

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

```
<Axes: >
```

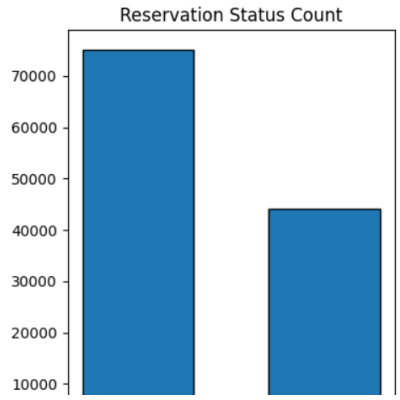


DATA ANALYSIS AND VISUALISATION

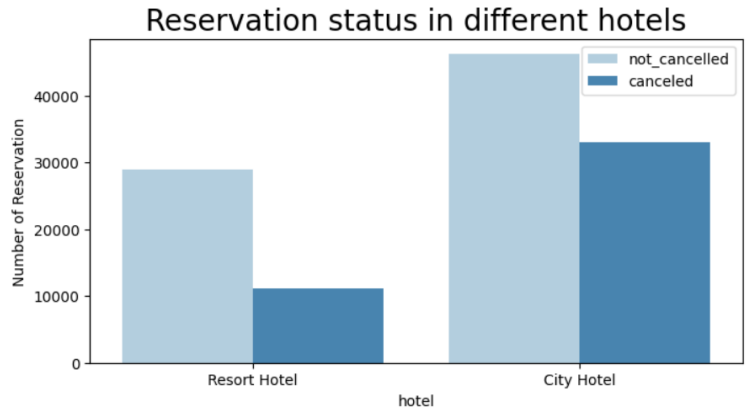
```
[58] df=df[df['adr']<5000]
```

```
cancelled_per=df['is_canceled'].value_counts(normalize=True)
print(cancelled_per)
plt.figure(figsize=(4,5))
plt.title("Reservation Status Count")
plt.bar(['Not Cancelled','Cancelled'],df['is_canceled'].value_counts(),edgecolor='k',width=0.6)
#plt.show()
```

```
0    0.629589
1    0.370411
Name: is_canceled, dtype: float64
<BarContainer object of 2 artists>
```




```
plt.figure(figsize=(8,4))
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,palette='Blues')
legend_lables,_=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.legend(['not_cancelled','canceled'])
plt.ylabel('Number of Reservation')
plt.show()
```




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

```
0    0.722366
1    0.277634
Name: is_canceled, dtype: float64
```

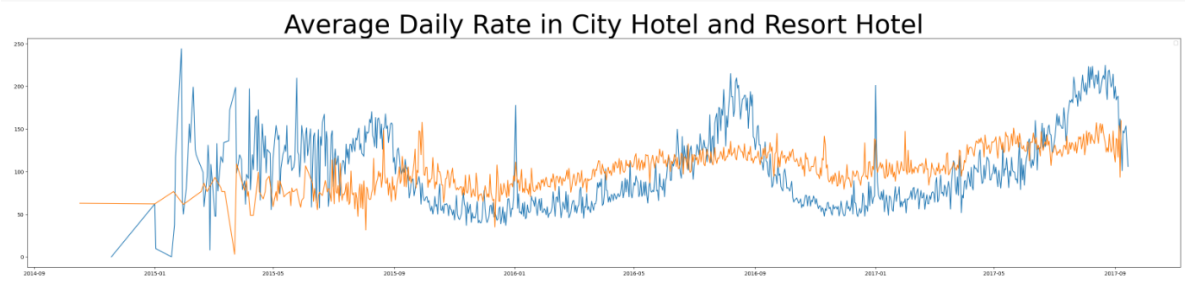
Loading...

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

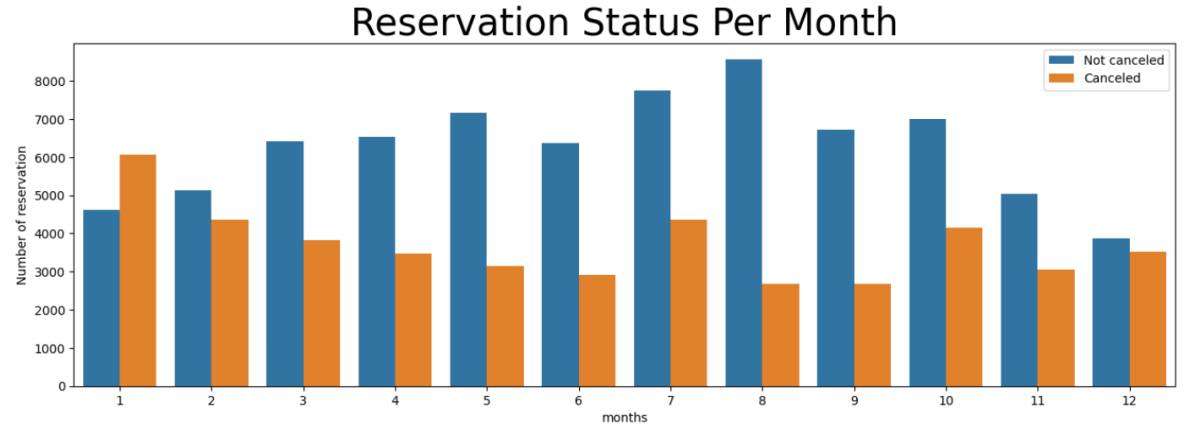
```
0    0.582738
1    0.417262
Name: is_canceled, dtype: float64
```

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

```
[64] plt.figure(figsize=(40,8))
     plt.title("Average Daily Rate in City Hotel and Resort Hotel",fontsize=50)
     plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort Hotel')
     plt.plot(city_hotel.index,city_hotel['adr'],label='City  Hotel')
     plt.legend('Resort Hotel','City Hotel')
     plt.show()
```

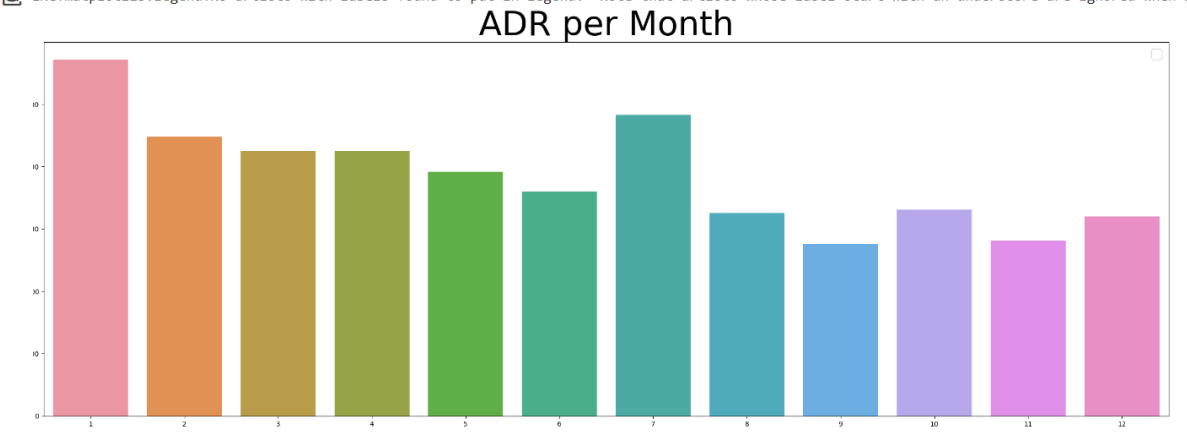



```
df['months']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,5))
ax1=sns.countplot(x='months',hue='is_canceled',data=df)
plt.ylabel('Number of reservation')
plt.legend(['Not canceled','Canceled'])
plt.title("Reservation Status Per Month",fontsize=30)
plt.show()
```

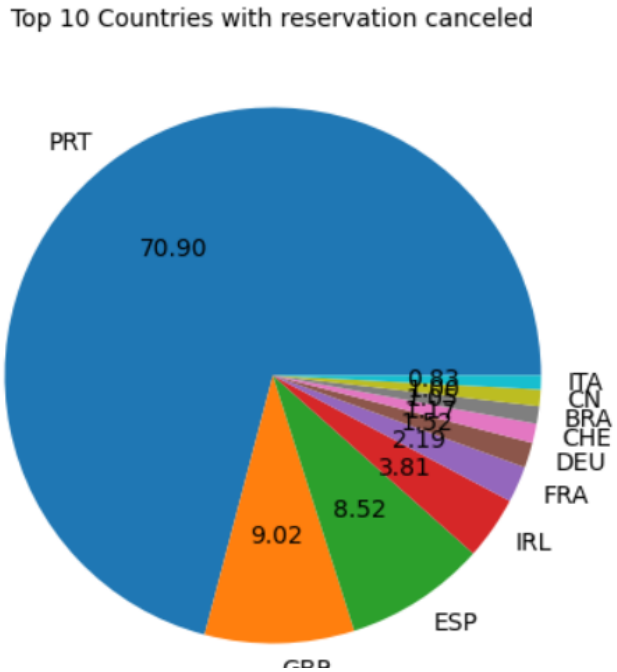



```
plt.figure(figsize=(30,10))
plt.title("ADR per Month",fontsize=50)
grouped_data = df[df['is_canceled'] == 1].groupby('months')[['adr']].sum().reset_index()
sns.barplot(x='months', y='adr', data=grouped_data)
plt.legend(fontsize=20)
plt.show()
```

```
ING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an underscore are ignored when leg
```



```
canceled_data = df[df['is_canceled'] == 1]
top_10_country=canceled_data['country'].value_counts()[:10]
plt.figure(figsize=(20,5))
plt.title("Top 10 Countries with reservation canceled",fontsize=10)
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.show()
```



[21]
```
df['market_segment'].value_counts()
```

```
Online TA        7403
Groups           3015
Offline TA/TO    1973
Direct           1580
Corporate         832
Complementary      45
Name: market_segment, dtype: int64
```

[22]
```
df['market_segment'].value_counts(normalize=True)
```

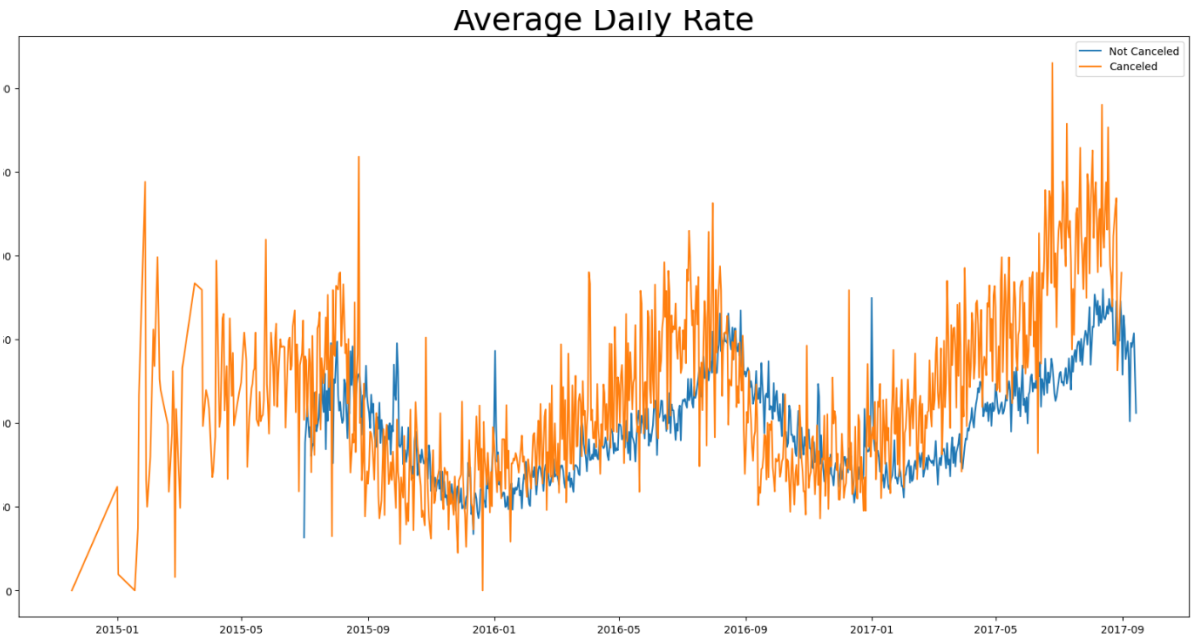
```
Online TA        0.498586
Groups           0.203058
Offline TA/TO    0.132880
Direct           0.106412
Corporate        0.056034
Complementary    0.003031
Name: market_segment, dtype: float64
```

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

```
Online TA        0.589476
Groups           0.206326
Offline TA/TO    0.090346
Direct           0.080383
Corporate        0.030373
Complementary    0.003095
Name: market_segment, dtype: float64
```

```
cancelled_df_adr=canceled_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_canceled_data=df[df['is_canceled']==0]
not_canceled_df_adr=not_canceled_data.groupby('reservation_status_date')[['adr']].mean()
not_canceled_df_adr.reset_index(inplace=True)
not_canceled_df_adr.sort_values('reservation_status_date',inplace=True)
plt.figure(figsize=(20,10))
plt.title("Average Daily Rate",fontsize=30)
plt.plot(not_canceled_df_adr['reservation_status_date'],not_canceled_df_adr['adr'],label='Not Canceled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label='Canceled')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x7803fa3ceb90>
```



DataFrame with shape (850, 2)

```
cancelled_df_adr=cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') &
                                  (cancelled_df_adr['reservation_status_date']<'2017-09')]
not_canceled_df_adr=not_canceled_df_adr[(not_canceled_df_adr['reservation_status_date']>'2016') &
 (not_canceled_df_adr['reservation_status_date']<'2017-09')]
```

[28]
```
plt.figure(figsize=(20,10))
plt.title('Average Daily Rate',fontsize=30)
plt.plot(not_canceled_df_adr['reservation_status_date'],not_canceled_df_adr['adr'],label='not canceled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label='canceled')
plt.legend(fontsize=20)
```

```
<matplotlib.legend.Legend at 0x7804021ef640>
```