Project Name - Hotel Booking Analysis

```
Project Name - Hotel Booking Analysis
Project Type - EDA
Contribution - Individual
Team Member 1 -Jyoti Ghaytadak
```

Project Summary -

Hotel industry is very volatile industry and the booking depends on variety of factors such as types of hotels. days of week, months and many more in this project I will be analysizing the patterns available in given data set to help the hotel plan better. The given data set contains booking information for a city hotel and a resort hotel includes information such as when the booking was made, length of stay, the number of adults children and babies and the number of available parking spaces, among other things, the aim is to creat meaningful estimators from the dataset i have and to perform explatory dat analysis but before doing that i have done some manipulation with a dataset. I have cleared the duplicate values changed the information types. replaced null values, checked unique values and also created some additional column as per the requirment.The tools for data analysis used in this project are the package numpy and pandas. and to visualize and explore the data matplotlib and seaborn.

importing Labraries

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

Loading the dataset

In [2]:
```
df=pd.read_csv(r"C:\Users\C ZONE\Downloads\hotel_booking.csv")
```

Exploratory Data Analysis and Data Cleaning

In [3]:
```
df.head()
```

Out[3]:

	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 × 33 columns

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

Out[4]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_nu
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 × 33 columns

Variables Description

The name of individual vaariables mentioned in the column of dataset and the discription of them are listed below:

```
hotel: type of hotel(city hotel or resort hotel)
is_cancelled: 1 indicates as the booking is cancelled and 0 indicates no
cancellatio
lead_time : number of dates before arrival from booking date.
arrival_date_year: year of arrival date
arrival_date_month : year of arrival month
arrival_date_week_number : week number of that perticular arrival year
arrival_date_day_by_month : month number of that perticular arrival year
stay_in_weekend_night : number of saturday and sunday night spend in the hotel
by tourist or guests
```

```
stay_in_week_night : number of week night(monday to friday) spend in the hotel
by the tourist/guestes
adults: number of adults among the guests
children : number of children among the guests
Babies : number of babies among the guests
meal : type of meal booked
country: country from where the guest belong
market_segment: Designation of market segment
Distribution_channel: name of the booking distribution channel
is_reapeated_guests : 1 indicates the guest is repeated guest and 0 indicates
new guests
previous_cancellation: number of cancellation prior to current booking
previous_booking_not_cancelled: number of not cancelled booking prior to the
current booking
reserved_room_type: code of room reserved
assigned_room_type: code of room assigned
booking_changes : number of changes made by the booking
deposite_type: type of deposite made by the customer/guest
agent: travel agent ID who made the booking
company: compant ID who made the booking
days_in_waiting_list: number of days in the wainting list
cyustomer_type: type of customer assuming one of four categories
adr: average daily rates defined by dividing the sum of all lodging
transaction by the hotel number of staying night
required_car_parking_spaces: number of car parking spaces required by the
customer/guests
total_of_special_request: number of special request made by the customer
reservation_status: reservation status( cancelled, check out and no show)
reservation_status_Date: date on which the last reservation status was updated
```

In [5]:
```
df.shape
```

Out[5]:
```
(119390, 33)
```

In [6]:
```
df.columns
```

Out[6]:
```
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'],
      dtype='object')
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 33 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 
dtypes: float64(4), int64(16), object(13)
memory usage: 30.1+ MB
```

In [8]:
```
df['reservation_status_date']=
pd.to_datetime(df['reservation_status_date'])
```

In [9]:
```
df.describe(include='object')
```

Out[9]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reser
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 [10]:
```
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
['name' 'Samuel Zavala' 'Dr. Victor Martin' ... 'Wesley Aguilar'
 'Caroline Conley MD' 'Ariana Michael']
--------------------------------------------------
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
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
dtype: int64
```

In [12]:
```
df.drop(['company','agent'],
        axis=1, inplace=True)
```

In [13]:
```
df.dropna(inplace=True)
```

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                        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
name                            0
dtype: int64
```

In [15]:
```
df.describe()
```

Out[15]:

	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	
std	0.483168	106.903309	0.707459	13.589971	
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	

In [16]:
```
df=df[df['adr']<5000]
```

In [17]:
```
# lets find out total cancelatin percentage
total_cancellation = df['is_canceled'].sum()
total_cancellation
```

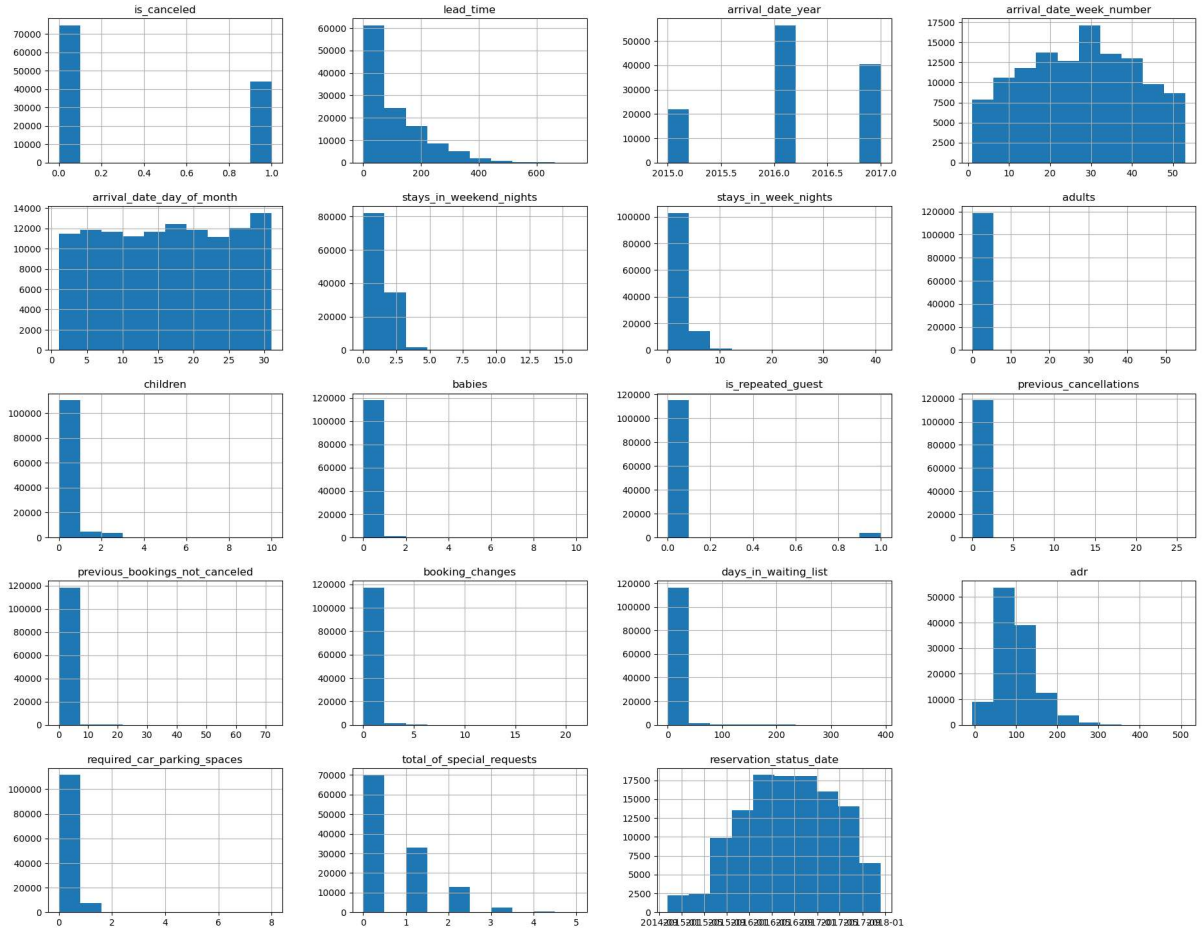
Out[17]: 44152

In [18]:
```
number_of_rows = df.shape[0]
number_of_rows
```

Out[18]: 118897

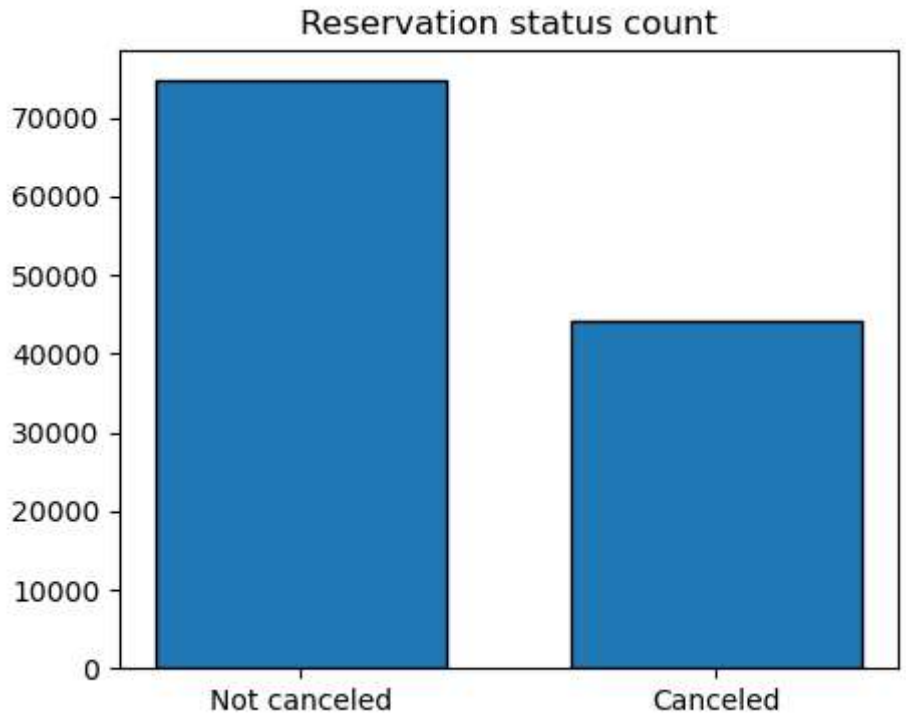
Data Analysis and Visulization

In [19]:
```
# Chart - 1 visualization code
# lets first look at the overview of the dataset through histogram plotting
df.hist(figsize=(23,18))
plt.show()
```

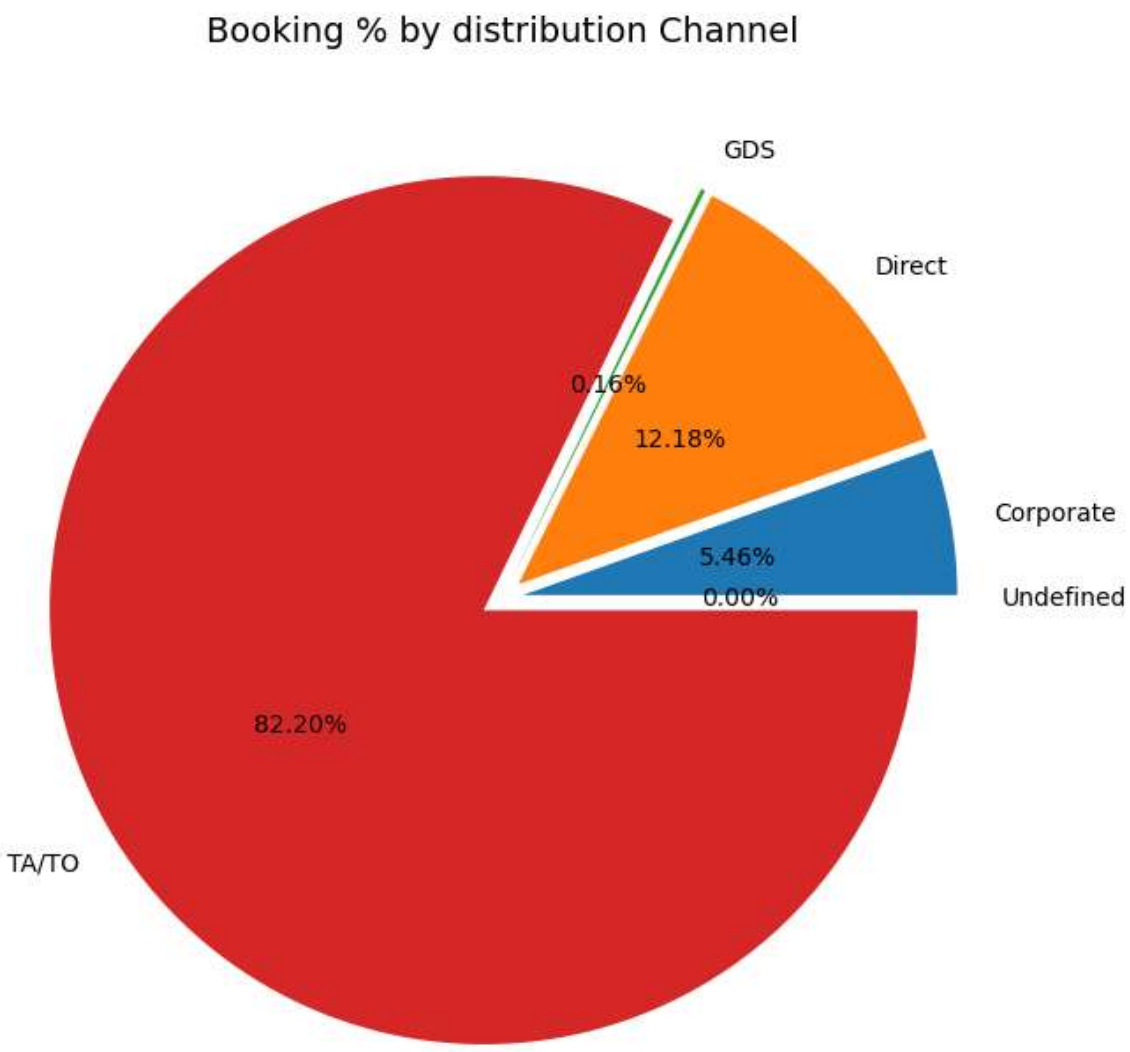


In [20]:
```
cancelled_perc=df['is_canceled'].
value_counts(normalize=True)
print(cancelled_perc)
plt.figure(figsize=(5,4))
plt.title('Reservation status count')
plt.bar(['Not canceled','Canceled'],
        df['is_canceled'].value_counts(),edgecolor='k',width=0.7)
plt.show()
```

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



In [22]:
```
# Chart - 2 Visualization code
# now we  pie chart is showing the booking ratio of distribution channel.
group_by_dc = df.groupby('distribution_channel')
d1 = pd.DataFrame(round((group_by_dc.size()/df.shape[0])*100,2)).
reset_index().rename(columns = {0:'Booking_%'})
plt.figure(figsize = (8,8))
data = d1['Booking_%']
labels = d1['distribution_channel']
plt.pie(x = data, autopct = "%.2f%%",
        explode =[0.05]*5, labels = labels, pctdistance=0.5)
plt.title("Booking % by distribution Channel",
          fontsize = 14);
```



In [23]:
```
resort_hotel=df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)
```

Out[23]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

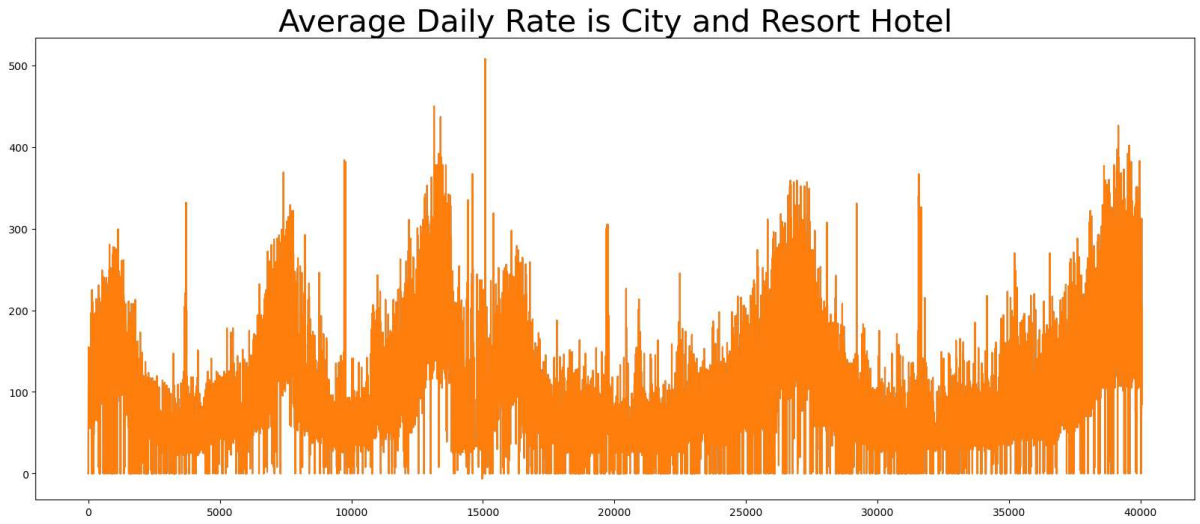
In [24]:
```
city_hotel=df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)
```

Out[24]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [25]:
```
resort_hotrl=resort_hotel.groupby('reservation_status_date')
[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')
[['adr']].mean()
```

In [26]:
```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate is City and Resort Hotel',fontsize=30)
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort Hotel')
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort Hotel')
```

Out[26]:
```
[<matplotlib.lines.Line2D at 0x1d88f62b4c0>]
```



conclusion

```
Summarise your finding and any insights gained from the analysis
```