Insight into hotel dataset

April 27, 2023

0.0.1 Problem Statement and possible measures to be taken

Recently, hotels and resorts are facing difficult time and high cancellation rates that results into fewer revenues and less then ideal hotel room use expected. In this case, the primary goal is to lowering cancellation rates in order to increase revenue. *** Now, as an analyst our responsibility is to break down the problems, provide a thorugh analysis of the data to find out the meaningful reasons behind high cancellation rate and suggest necessary solutions and advice to address this problem as well essential meassures to deal with this issue.

0.0.2 Assumptions:

```
- The provided info is current and can be used to analyze the possible plans in an efficient ma
- No unanticipated negatives to the hotel employing any advise technique.
- The hotels are neither have unusual occurrences (outliers) between 2015 and 2017 that might l
- The biggest factors affecting the effectiveness of earning income is booking cencellations.
- Booking and cancellation are assumed to be made in the same year.
```

0.0.3 Research Queries:

```
* Which variables are affecting cancellations of hotel reservation ?
* Which KPIs can help to improve the current situation?
* What about pricing and promotional decisions?
```

0.0.4 Hypothesis

- Higher Price = More cancellations
- Longer waiting list = Frequent Cancellation
- The majority of clients use off line travel agents for hotel reservations.

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

[33]:
```
# DataFrame import
df = pd.read_csv('hotel_bookings 2.csv')
```

[3]:
```
# Let's have a look at the Number of Rows and Columns
df.shape
```

[3]:
```
(119390, 32)
```

We have got 119390 Rows and 32 different columns

[35]:
```
# A detailed overview of the dataFrame
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
```

[38]:
```
# The first 10 rows
df.head(10)
```

[38]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	\
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	
5	Resort Hotel	0	14	2015	July	
6	Resort Hotel	0	0	2015	July	
7	Resort Hotel	0	9	2015	July	
8	Resort Hotel	1	85	2015	July	
9	Resort Hotel	1	75	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	
5	27	1	
6	27	1	
7	27	1	
8	27	1	
9	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	
5	0	2	2	…	No Deposit	
6	0	2	2	…	No Deposit	
7	0	2	2	…	No Deposit	
8	0	3	2	…	No Deposit	
9	0	3	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	

```
5  240.0     NaN                    0     Transient   98.0
6    NaN     NaN                    0     Transient  107.0
7  303.0     NaN                    0     Transient  103.0
8  240.0     NaN                    0     Transient   82.0
9   15.0     NaN                    0     Transient  105.5

   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
5                            0                          1           Check-Out
6                            0                          0           Check-Out
7                            0                          1           Check-Out
8                            0                          1            Canceled
9                            0                          0            Canceled

  reservation_status_date
0                1/7/2015
1                1/7/2015
2                2/7/2015
3                2/7/2015
4                3/7/2015
5                3/7/2015
6                3/7/2015
7                3/7/2015
8                6/5/2015
9               22/4/2015

[10 rows x 32 columns]
```

[39]:
```
# Last Five rows
df.tail()
```

[39]:
```
             hotel  is_canceled  lead_time  arrival_date_year  \
119385  City Hotel            0         23               2017
119386  City Hotel            0        102               2017
119387  City Hotel            0         34               2017
119388  City Hotel            0        109               2017
119389  City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
```

```
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
119385                         30                        2
119386                         31                        2
119387                         31                        2
119388                         31                        2
119389                         29                        2

        stays_in_week_nights  adults  …  deposit_type  agent company  \
119385                     5       2  …    No Deposit  394.0     NaN
119386                     5       3  …    No Deposit    9.0     NaN
119387                     5       2  …    No Deposit    9.0     NaN
119388                     5       2  …    No Deposit   89.0     NaN
119389                     7       2  …    No Deposit    9.0     NaN

       days_in_waiting_list customer_type     adr  \
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     Transient  151.20

        required_car_parking_spaces  total_of_special_requests  \
119385                            0                          0
119386                            0                          2
119387                            0                          4
119388                            0                          0
119389                            0                          2

        reservation_status reservation_status_date
119385           Check-Out                6/9/2017
119386           Check-Out                7/9/2017
119387           Check-Out                7/9/2017
119388           Check-Out                7/9/2017
119389           Check-Out                7/9/2017

[5 rows x 32 columns]
```

[40]:
```
# Let's have a statistical (descriptive) view of the dataset
df.describe()
```

[40]:
```
         is_canceled      lead_time  arrival_date_year  \
count  119390.000000  119390.000000      119390.000000
mean        0.370416     104.011416        2016.156554
std         0.482918     106.863097           0.707476
min         0.000000       0.000000        2015.000000
```

```
25%         0.000000      18.000000        2016.000000
50%         0.000000      69.000000        2016.000000
75%         1.000000     160.000000        2017.000000
max         1.000000     737.000000        2017.000000

       arrival_date_week_number  arrival_date_day_of_month  \
count             119390.000000              119390.000000
mean                  27.165173                  15.798241
std                   13.605138                   8.780829
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

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            119390.000000         119390.000000  119390.000000
mean                  0.927599              2.500302       1.856403
std                   0.998613              1.908286       0.579261
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                  19.000000             50.000000      55.000000

            children         babies  is_repeated_guest  \
count  119386.000000  119390.000000      119390.000000
mean        0.103890       0.007949           0.031912
std         0.398561       0.097436           0.175767
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

       previous_cancellations  previous_bookings_not_canceled  \
count           119390.000000                   119390.000000
mean                 0.087118                        0.137097
std                  0.844336                        1.497437
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

       booking_changes          agent      company  days_in_waiting_list  \
count    119390.000000  103050.000000  6797.000000         119390.000000
```

```
mean          0.221124      86.693382   189.266735              2.321149
std           0.652306     110.774548   131.655015             17.594721
min           0.000000       1.000000     6.000000              0.000000
25%           0.000000       9.000000    62.000000              0.000000
50%           0.000000      14.000000   179.000000              0.000000
75%           0.000000     229.000000   270.000000              0.000000
max          21.000000     535.000000   543.000000            391.000000

                adr  required_car_parking_spaces  total_of_special_requests
count  119390.000000                119390.000000              119390.000000
mean      101.831122                     0.062518                   0.571363
std        50.535790                     0.245291                   0.792798
min        -6.380000                     0.000000                   0.000000
25%        69.290000                     0.000000                   0.000000
50%        94.575000                     0.000000                   0.000000
75%       126.000000                     0.000000                   1.000000
max      5400.000000                     8.000000                   5.000000
```

[7]:
```
df.describe(include = 'object') # to examine the categorical column
```

[7]:
```
             hotel arrival_date_month    meal country market_segment  \
count       119390             119390  119390  118902         119390
unique           2                 12       5     177              8
top     City Hotel             August      BB     PRT      Online TA
freq         79330              13877   92310   48590          56477

       distribution_channel reserved_room_type assigned_room_type  \
count                119390             119390             119390
unique                    5                 10                 12
top                   TA/TO                  A                  A
freq                  97870              85994              74053

       deposit_type customer_type reservation_status reservation_status_date
count        119390        119390             119390                  119390
unique            3             4                  3                     926
top      No Deposit     Transient          Check-Out             21/10/2015
freq         104641         89613              75166                    1461
```

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

[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')
```

[9]:
```
# Converting reservation date to datetime format
df['reservation_status_date']=pd.to_datetime(df['reservation_status_date'])
```

[10]:
```
# Type of variables in each column
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']
--------------------------------------------------
```

[11]:
```
# Is there any null value?
df.isnull().sum() # Finding the missing value
```

[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
dtype: int64
```

'Yes. But, for now We are not considering company and agent column because it is difficult or time cosuming to clean and such a large ammount of data.'

[13]:
```
df.drop(['agent', 'company'], axis =1, inplace = True) # to remove these two columns
df.dropna(inplace = True) # this will remove missing values from all the rows
```

[14]:
```
df.isnull().sum() # so there is no missing values
```

[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
dtype: int64
```

Now we have got a clean dataset

[44]:
```
# So, lets have a look at the statitical overview of our new dataset
df.describe()
```

[44]:

```
         is_canceled      lead_time  arrival_date_year  \
count  119390.000000  119390.000000      119390.000000
mean        0.370416     104.011416        2016.156554
std         0.482918     106.863097           0.707476
min         0.000000       0.000000        2015.000000
25%         0.000000      18.000000        2016.000000
50%         0.000000      69.000000        2016.000000
75%         1.000000     160.000000        2017.000000
max         1.000000     737.000000        2017.000000

       arrival_date_week_number  arrival_date_day_of_month  \
count             119390.000000              119390.000000
mean                  27.165173                  15.798241
std                   13.605138                   8.780829
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

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            119390.000000         119390.000000  119390.000000
mean                  0.927599              2.500302       1.856403
std                   0.998613              1.908286       0.579261
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                  19.000000             50.000000      55.000000

            children         babies  is_repeated_guest  \
count  119386.000000  119390.000000      119390.000000
mean        0.103890       0.007949           0.031912
std         0.398561       0.097436           0.175767
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

       previous_cancellations  previous_bookings_not_canceled  \
count           119390.000000                   119390.000000
```

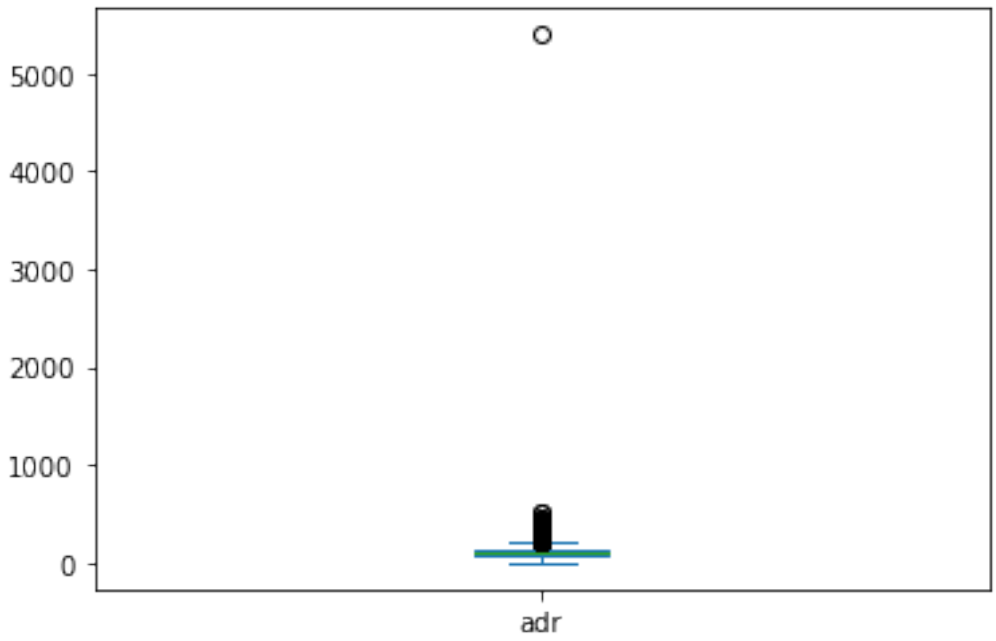
```
mean                 0.087118                      0.137097
std                  0.844336                      1.497437
min                  0.000000                      0.000000
25%                  0.000000                      0.000000
50%                  0.000000                      0.000000
75%                  0.000000                      0.000000
max                 26.000000                     72.000000

       booking_changes          agent      company  days_in_waiting_list  \
count    119390.000000  103050.000000  6797.000000         119390.000000
mean          0.221124      86.693382   189.266735              2.321149
std           0.652306     110.774548   131.655015             17.594721
min           0.000000       1.000000     6.000000              0.000000
25%           0.000000       9.000000    62.000000              0.000000
50%           0.000000      14.000000   179.000000              0.000000
75%           0.000000     229.000000   270.000000              0.000000
max          21.000000     535.000000   543.000000            391.000000

                adr  required_car_parking_spaces  total_of_special_requests
count  119390.000000                119390.000000              119390.000000
mean      101.831122                     0.062518                   0.571363
std        50.535790                     0.245291                   0.792798
min        -6.380000                     0.000000                   0.000000
25%        69.290000                     0.000000                   0.000000
50%        94.575000                     0.000000                   0.000000
75%       126.000000                     0.000000                   1.000000
max      5400.000000                     8.000000                   5.000000
```

[16]:
```
df['adr'].plot(kind = 'box')
```

[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a57826e2e0>
```



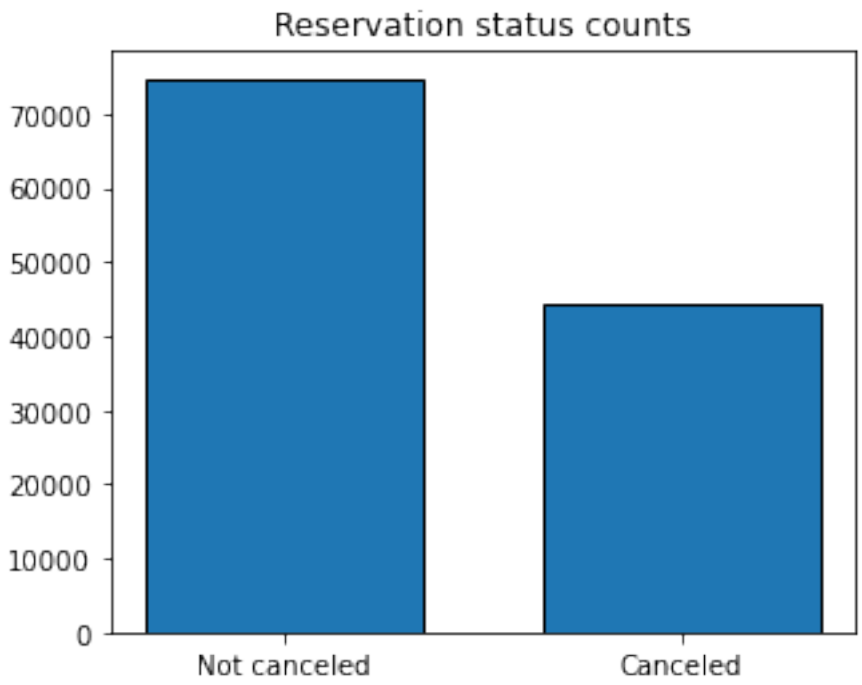
[45]:
```
df = df[df['adr']<5000]
```

0.1 Data Analysis and Visualization

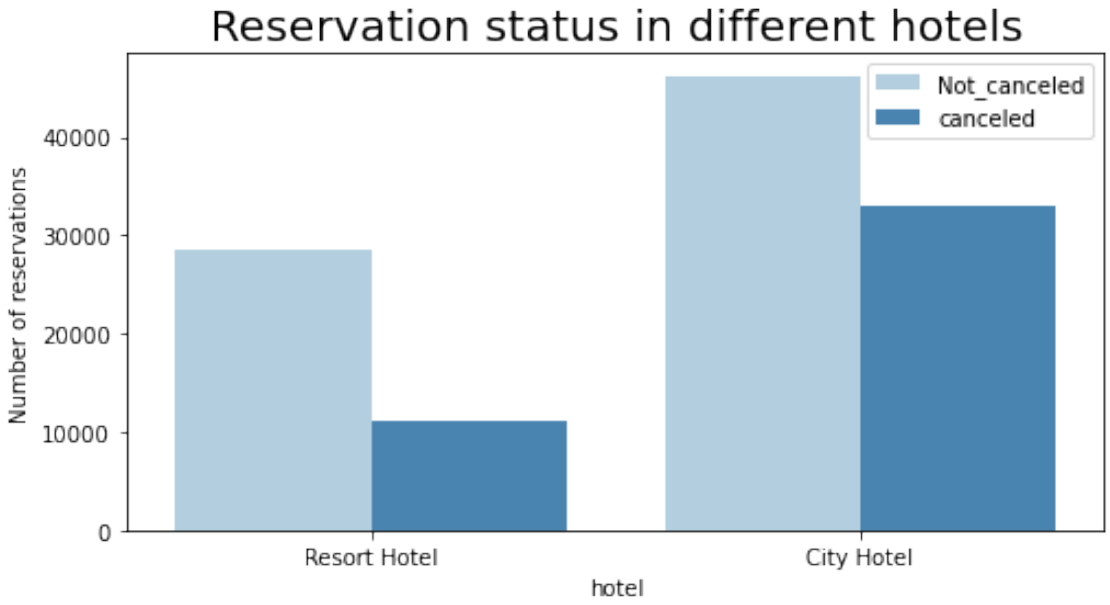
[18]:
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
print(cancelled_perc)

plt.figure(figsize = (5,4))
plt.title('Reservation status counts')
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
```



[19]:
```
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(box_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()
```



0.1.1 As we can see that around 37% of client cancelled their reservations that might affect the revenue significantly.

[20]:
```
resort_hotel = df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize = True)
```

[20]:
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

[21]:
```
city_hotel = df[df['hotel'] == 'City Hotel']
city_hotel['is_canceled'].value_counts(normalize = True)
```

[21]:
```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

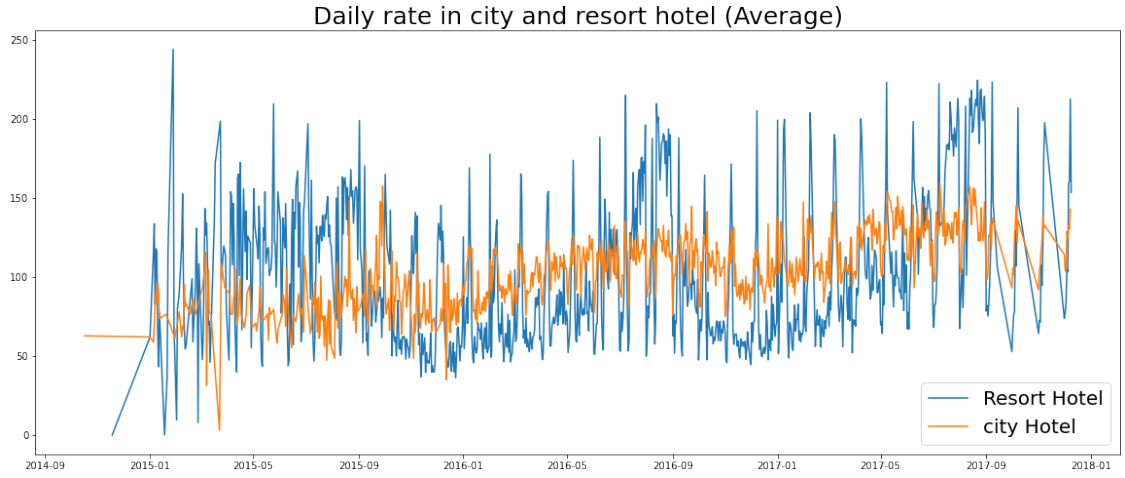
[22]:
```
resort_hotel = resort_hotel.groupby('reservation_status_date')[['adr']].mean()
print('mean_resot_hotel',resort_hotel.head())
city_hotel = city_hotel.groupby('reservation_status_date')[['adr']].mean()
print('mean_city_hotel',city_hotel.head())
```

```
mean_resot_hotel                             adr
reservation_status_date
2014-11-18                  0.000000
2015-01-01                 61.966667
2015-01-05                115.363333
```

```
2015-01-06                133.677143
2015-01-07                 82.485455
mean_city_hotel                                adr
reservation_status_date
2014-10-17                62.800000
2015-01-01                62.063158
2015-01-05                58.900000
2015-01-06                69.216667
2015-01-07                82.877500
```

Compared to resort hotels, city hotels have more bookings. It is possible that resort hotels are moe expensive than city hotels

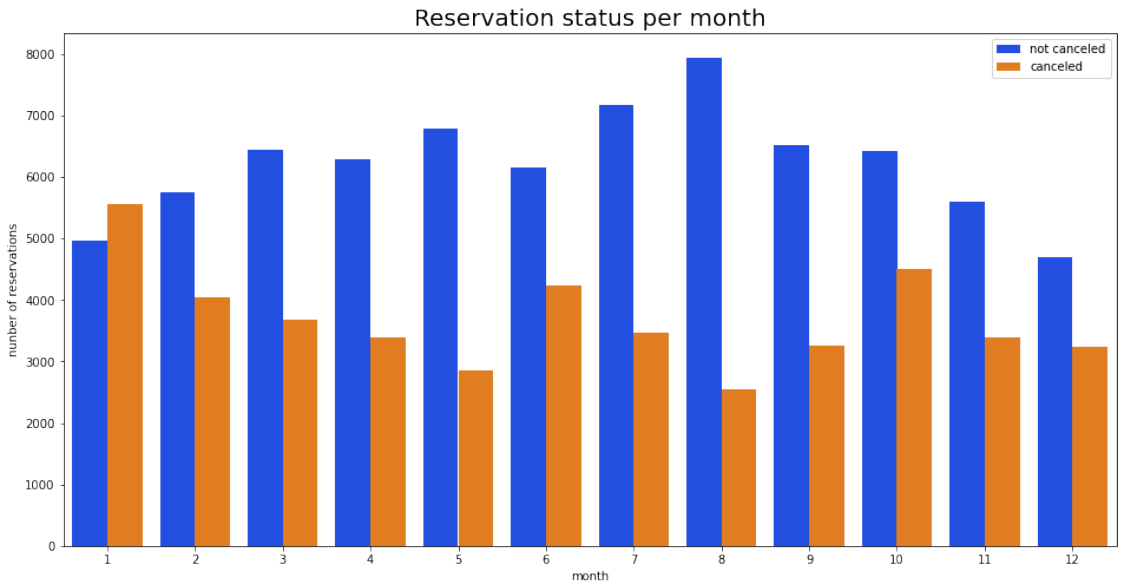
[47]:
```
plt.figure(figsize =(20,8))
plt.title('Daily rate in city and resort hotel (Average)', fontsize = 25)
plt.plot(resort_hotel.index, resort_hotel['adr'], label = 'Resort Hotel') #␣
 ↪index is the reservation date
plt.plot(city_hotel.index, city_hotel['adr'], label = 'city Hotel')
plt.legend(fontsize = 20)
plt.show()
```



It is clear from the chart that on certain days the average daily rate for a city hotel is less than resort hotel, and on other days, it is even less. It can be realized that the weekends and holidays may see a rise in resort hotel rates.

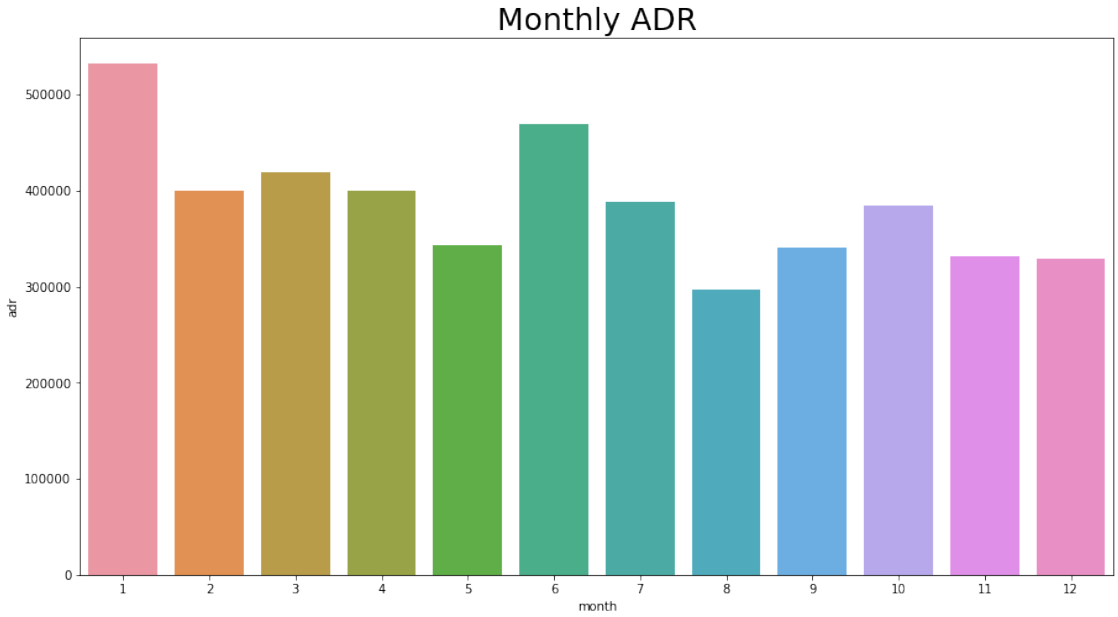
[24]:
```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x='month', hue = 'is_canceled', data = df, palette =␣
 ↪'bright')
legend_labels,_ = ax1. get_legend_handles_labels()
ax1.legend(bbox_to_anchor =(1,1))
plt.title('Reservation status per month', size = 20)
```

```
plt.xlabel('month')
plt.ylabel('nunber of reservations')
plt.legend(['not canceled', 'canceled'])
plt.show()
```



Both the group of number of confirmed reservations and the number of cancelled reservations are largest in the month of August. While in january has highest canceled reservations.

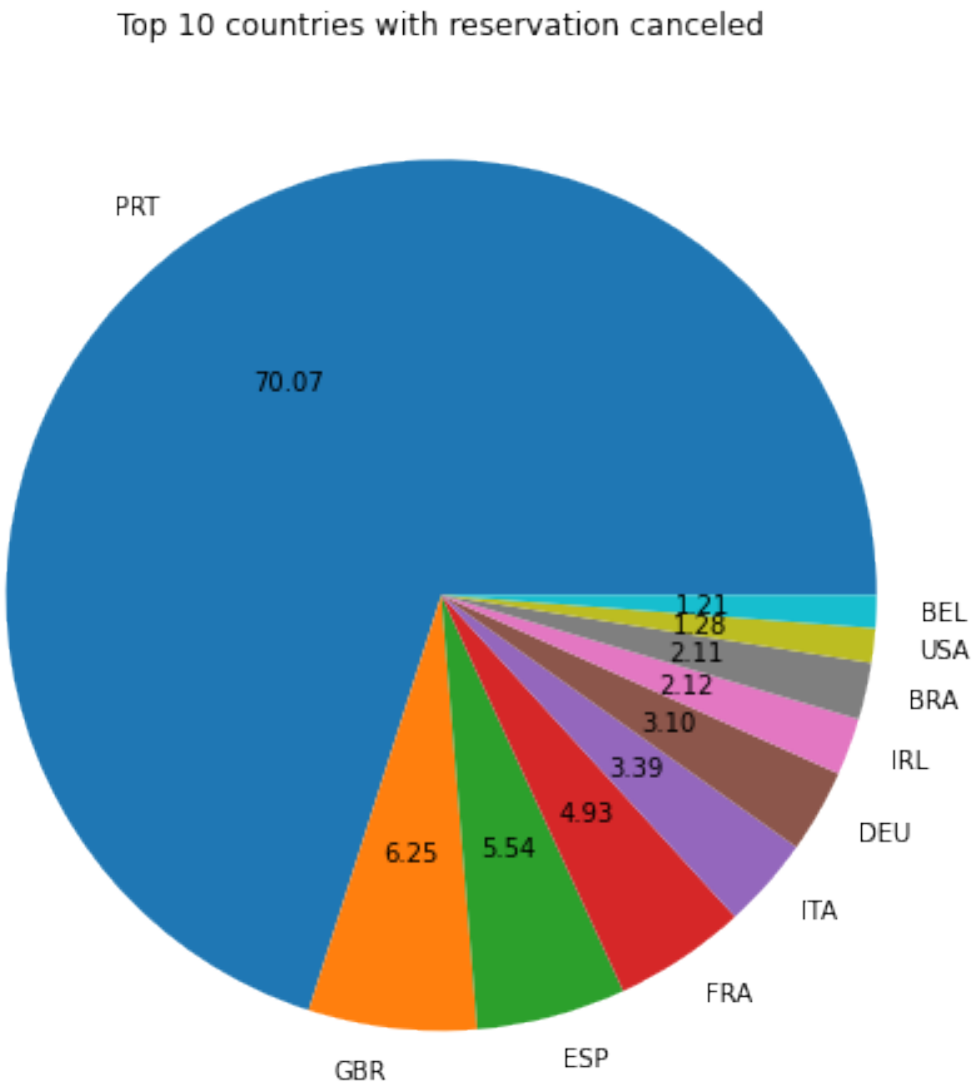
[25]:
```
plt.figure(figsize = (15,8))
plt.title('Monthly ADR', fontsize = 25)
sns.barplot('month', 'adr', data = df[df['is_canceled'] == 1].
 ↪groupby('month')[['adr']].sum().reset_index())
plt.show()
```



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

0.1.2 Countrywise cancellation

[26]:
```
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()
```



Portugal has the highest cancellation record.

Now, Lets check out the following questions - The customers belongs to which areas are visiting the hotels and making reservations. - Are they comming from direct, Online or Offline travel agency?

[27]:
```
df['market_segment'].value_counts() # Most of the customer comes from online␣
↪reservation
```

[27]:
```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
```

```
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

[28]:
```
df['market_segment'].value_counts(normalize = True) # most of the customers␣
 ↪coming from online reservation as %.
```

[28]:
```
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
```

Around 46% of the clients come from online travel agencies, whereas 27% come form groups. Only 4% of the clients book hotels directly by visiting them and making reservaitons.

[29]:
```
cancelled_data['market_segment'].value_counts(normalize=True) # Online Travel␣
 ↪Agency (TA) has the largest portion of cancellation.
```

[29]:
```
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
```

Online Travel Agency also has the highest cancellation record

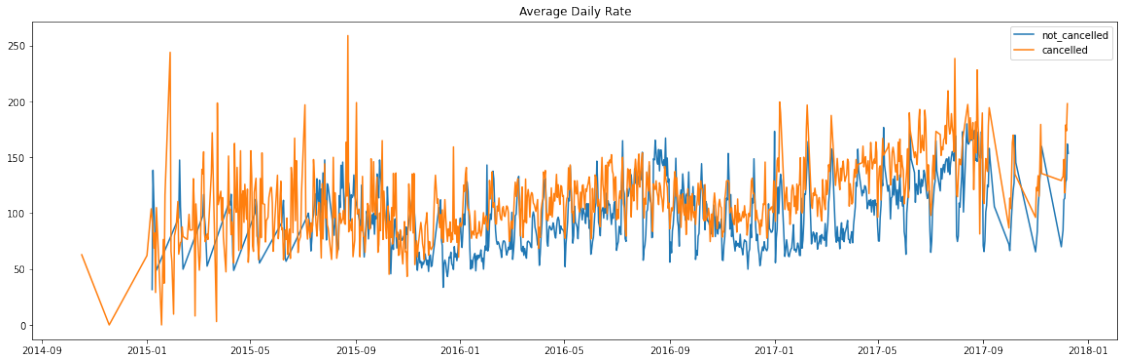
[30]:
```
# lets see the average daily price is higer in cancelled or non cancelled␣
 ↪reservation
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[['adr']].
 ↪mean()
cancelled_df_adr.reset_index(inplace = True)
cancelled_df_adr.sort_values('reservation_status_date', inplace = True)

not_cancelled_data = df[df['is_canceled']==0]
not_cancelled_df_adr = not_cancelled_data.
 ↪groupby('reservation_status_date')[['adr']].mean()
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')
```

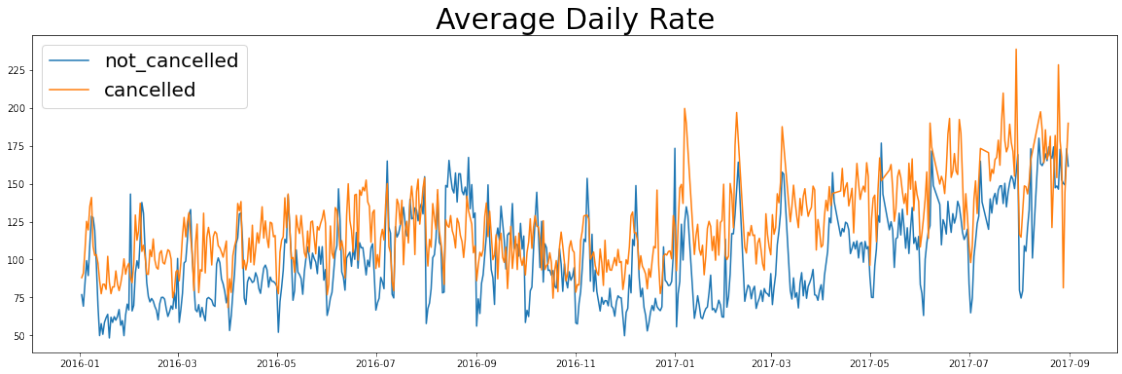
```
plt.
 ↪plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr['adr'],␣
 ↪label = 'not_cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label = 'cancelled')
plt.legend()
```

[30]: <matplotlib.legend.Legend at 0x1a5000fbbe0>



[31]:
```
cancelled_df_adr =␣
 ↪cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') &␣
 ↪(cancelled_df_adr['reservation_status_date']<'2017-09')]
not_cancelled_df_adr =␣
 ↪not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016')␣
 ↪& (not_cancelled_df_adr['reservation_status_date']<'2017-09')]
```

[32]:
```
plt.figure(figsize = (20,6))
plt.title('Average Daily Rate', fontsize = 30)
plt.plot(not_cancelled_df_adr['reservation_status_date'],␣
 ↪not_cancelled_df_adr['adr'], label = 'not_cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'],␣
 ↪label = 'cancelled')
plt.legend(fontsize = 20)
plt.show()
```



So, the cancellation is highly related to the higher price. It clearly proves all the above analysis, that the higher price leads to higher cancellation.

0.2 Suggested Advice from the above analysis

- Hotels should work on their pricing strategies and try to lower the rate for specific hotels based on locations.
- Providing discounts and offers based on rooms, holidays or weekends can also help to reduce the cancellation rate.
- Proper marketing strategies should be applied in the month of January based on customers demand.
- They can also focus on the quality of hotels and services mainly in Portugal to reduce the cancellation rate.