Project Name - Hotel Booking EDA

Project Type - EDA

Contribution - Individual

Project Summary -

Hotel booking EDA, or exploratory data analysis, involves analyzing a large dataset of hotel bookings to identify trends, patterns, and insights. The dataset includes information on city hotels, resort hotels, and various other features related to bookings.

The dataset contains nearly 32 columns, including information such as hotel type, booking date, lead time, number of adults and children, meal plan, room type, and booking status. With such a large amount of data, it is important to perform thorough exploratory data analysis to gain a better understanding of the dataset and identify potential issues or insights.

One key trend that emerges from the data is the seasonal nature of hotel bookings. The dataset includes bookings made between July 2015 and August 2017, and it is clear that there are peaks and valleys in hotel bookings throughout the year. For example, bookings tend to be highest in the summer months, particularly in August, while they are lowest in the winter months, particularly in December and January.

Another important trend to consider is the difference in booking patterns between city hotels and resort hotels. The dataset includes information on both types of hotels, and it is clear that there are significant differences in the way bookings are made. For example, resort hotels tend to have longer lead times than city hotels, and they are more likely to offer all-inclusive meal plans.

One interesting finding from the data is the relationship between lead time and cancellation rates. The data shows that bookings made further in advance are more likely to be cancelled, with cancellation rates declining as the booking date approaches. This may be due to the fact that people are more likely to change their plans the further in advance they make a booking, whereas last-minute bookings are more likely to be committed.

Another important insight from the data is the relationship between customer demographics and booking patterns. The data includes information on the number of adults and children in each booking, as well as the country of origin for each booking. By analyzing this information, it is

possible to identify trends and patterns in the types of customers who book different types of hotels.

Overall, exploratory data analysis of the hotel booking dataset provides valuable insights into trends, patterns, and potential issues in the data. By understanding these trends and patterns, hotel operators can better understand their customers and make more informed decisions about pricing, marketing, and other business strategies.

GitHub Link -

https://github.com/RiyazAhammad555/EDA-Project

Problem Statement

The problem statement could be to identify trends, patterns, and potential issues in a large dataset of hotel bookings, in order to gain insights into customer behavior and inform business strategies for hotel operators. Specifically, the project aims to answer questions such as: What are the seasonal patterns in hotel bookings? How do booking patterns differ between city hotels and resort hotels? What is the relationship between lead time and cancellation rates? What are the demographics of customers who book different types of hotels? By answering these questions, the project seeks to provide valuable insights that can help hotel operators optimize their pricing, marketing, and other business strategies to better serve their customers and improve profitability.

Define Your Business Objective?

The business objective for the hotel booking EDA project could be to use data-driven insights to optimize pricing, marketing, and other business strategies for hotel operators, with the ultimate goal of improving customer satisfaction and profitability. Specifically, the project aims to help hotel operators:

1.Identify seasonal patterns in hotel bookings and adjust pricing and marketing strategies accordingly to maximize occupancy and revenue.

2.Understand the differences in booking patterns between city hotels and resort hotels, and tailor marketing and service offerings to better serve each customer segment.

3.Analyze the relationship between lead time and cancellation rates to optimize revenue management and minimize lost revenue due to cancellations.

4.Identify customer demographics and preferences for different types of hotels to tailor marketing and service offerings to specific customer segments.

5.By achieving these business objectives, hotel operators can improve customer satisfaction by offering personalized and tailored services, while also increasing revenue and profitability by optimizing pricing and marketing strategies.Answer Here.

General Guidelines : -

1. Well-structured, formatted, and commented code is required.
2. Exception Handling, Production Grade Code & Deployment Ready Code will be a plus. Those students will be awarded some additional credits.

   The additional credits will have advantages over other students during Star Student selection.

   ```
   [ Note: - Deployment Ready Code is defined as, the whole .ipynb notebook should
             without a single error logged. ]
   ```

3. Each and every logic should have proper comments.
4. You may add as many number of charts you want. Make Sure for each and every chart the following format should be answered.

```
# Chart visualization code
```

- Why did you pick the specific chart?
- What is/are the insight(s) found from the chart?
- Will the gained insights help creating a positive business impact? Are there any insights that lead to negative growth? Justify with specific reason.

5. You have to create at least 20 logical & meaningful charts having important insights.

[Hints : - Do the Vizualization in a structured way while following "UBM" Rule.

U - Univariate Analysis,

B - Bivariate Analysis (Numerical - Categorical, Numerical - Numerical, Categorical - Categorical)

M - Multivariate Analysis]

Let's Begin !

1. Know Your Data

```
#mounting drive to access csv file
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Import Libraries

```
# Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from datetime import datetime
import seaborn as sns
%matplotlib inline
```

Dataset Loading

```
# Load Dataset
path='/content/drive/MyDrive/Hotel Dataset/'
df=pd.read_csv(path+'Hotel Bookings.csv')
```

Dataset First View

```
# Dataset First Look
df
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_d
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	
...	...	...	...	...	...	
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	

Dataset Rows & Columns count

```
# Dataset Rows & Columns count
print('Number of rows are',len(df.index))
print('Number of Columns are',len(df.columns))
```

```
Number of rows are 119390
Number of Columns are 32
```

Dataset Information

```
# Dataset Info
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
```

Duplicate Values

```
# Dataset Duplicate Value Count
df.duplicated().value_counts() # true means number of duplicate rows presented

#here we have to drop the duplicate values from the dataset

df.drop_duplicates(inplace=True)
```

Missing Values/Null Values

```
# Missing Values/Null Values Count
df.isna().sum().sort_values(ascending=False) #to find the null values
```

```
company                         82137
agent                           12193
country                           452
children                            4
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
hotel                               0
previous_cancellations              0
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
previous_bookings_not_canceled      0
is_repeated_guest                   0
is_canceled                         0
distribution_channel                0
market_segment                      0
meal                                0
babies                              0
adults                              0
stays_in_week_nights                0
stays_in_weekend_nights             0
arrival_date_day_of_month           0
arrival_date_week_number            0
arrival_date_month                  0
arrival_date_year                   0
lead_time                           0
reservation_status_date             0
dtype: int64
```

```
# Visualizing the missing values

#there are number of null values were found in company,agent,country and children columns and
df['company'].fillna(0,inplace=True)
df['agent'].fillna(0,inplace=True)
df['country'].fillna('Others',inplace=True)
df['children'].fillna(df['children'].mode()[0],inplace=True)

#checking if there are any null values left or not
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
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
agent                           0
company                         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
```

What did you know about your dataset?

Hotel bookings dataset contains a wide range of information related to hotel bookings, such as:

Hotel information: This may include details about the hotel, such as its name, location, number of rooms, and amenities.

1.Booking information: This may include details about the booking itself, such as the booking date, check-in and check-out dates, and length of stay.

2.Customer information: This may include details about the customer making the booking, such as their name, age, country of origin, and contact information.

3.Room information: This may include details about the type of room booked, such as the room type, number of adults and children, and meal plan.

4.Pricing information: This may include details about the price of the booking, such as the total price, taxes and fees, and payment method.

5.Cancellation information: This may include details about whether the booking was cancelled, and if so, the reason for the cancellation.

Overall, the hotel bookings dataset can provide valuable insights into customer behavior, pricing and revenue management, and other key aspects of hotel operations. By analyzing this data, hotel

operators can make more informed decisions to optimize their business strategies and improve customer satisfaction.

2. Understanding Your Variables

```
# Dataset Columns
df.columns  #there are 32 columns in the given data set
```

```
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')
```

```
# Dataset Describe
df.describe()   #this will give the statistical information of different columns
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_
count	87396.000000	87396.000000	87396.000000	87396.000000	
mean	0.274898	79.891368	2016.210296	26.838334	
std	0.446466	86.052325	0.686102	13.674572	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	11.000000	2016.000000	16.000000	
50%	0.000000	49.000000	2016.000000	27.000000	
75%	1.000000	125.000000	2017.000000	37.000000	
max	1.000000	737.000000	2017.000000	53.000000	

Variables Description

This dataset contains booking information for a city hotel and a resort hotel. It contains the following features

- hotel: Name of hotel (City or Resort)
- is_canceled: Whether the booking is canceled or not (0 for no canceled and 1 for canceled)
- lead_time: time (in days) between booking transaction and actual arrival.
- arrival_date_year: Year of arrival
- arrival_date_month: month of arrival
- arrival_date_week_number: week number of arrival date.
- arrival_date_day_of_month: Day of month of arrival date
- stays_in_weekend_nights: No. of weekend nights spent in a hotel
- stays_in_week_nights: No. of weeknights spent in a hotel
- adults: No. of adults in single booking record.
- children: No. of children in single booking record.
- babies: No. of babies in single booking record.
- meal: Type of meal chosen
- country: Country of origin of customers (as mentioned by them)
- market_segment: What segment via booking was made and for what purpose.
- distribution_channel: Via which medium booking was made.
- is_repeated_guest: Whether the customer has made any booking before(0 for No and 1 for Yes)
- previous_cancellations: No. of previous canceled bookings.
- previous_bookings_not_canceled: No. of previous non-canceled bookings.
- reserved_room_type: Room type reserved by a customer.
- assigned_room_type: Room type assigned to the customer.
- booking_changes: No. of booking changes done by customers
- deposit_type: Type of deposit at the time of making a booking (No deposit/ Refundable/ No refund)
- agent: Id of agent for booking
- company: Id of the company making a booking
- days_in_waiting_list: No. of days on waiting list.
- customer_type: Type of customer(Transient, Group, etc.)
- adr: Average Daily rate.
- required_car_parking_spaces: No. of car parking asked in booking
- total_of_special_requests: total no. of special request.
- reservation_status: Whether a customer has checked out or canceled,or not showed
- reservation_status_date: Date of making reservation status.

Check Unique Values for each variable.

```
# Check Unique Values for each variable.

# creating a dictionary to store unique values of each column
unique_value_dict={}

#creating list of columns
list_of_columns=list(df.columns)

for i in list_of_columns:
  unique_value_dict[i]=df[i].unique()

#unique value dict.
unique_value_dict
```

```
{'hotel': array(['Resort Hotel', 'City Hotel'], dtype=object),
 'is_canceled': array([0, 1]),
 'lead_time': array([342, 737,   7,  13,  14,   0,   9,  85,  75,  23,  35,  68,
18,
        37,  12,  72, 127,  78,  48,  60,  77,  99, 118,  95,  96,  69,
        45,  40,  15,  36,  43,  70,  16, 107,  47, 113,  90,  50,  93,
        76,   3,   1,  10,   5,  17,  51,  71,  63,  62, 101,   2,  81,
       368, 364, 324,  79,  21, 109, 102,   4,  98,  92,  26,  73, 115,
        86,  52,  29,  30,  33,  32,   8, 100,  44,  80,  97,  64,  39,
        34,  27,  82,  94, 110, 111,  84,  66, 104,  28, 258, 112,  65,
        67,  55,  88,  54, 292,  83, 105, 280, 394,  24, 103, 366, 249,
        22,  91,  11, 108, 106,  31,  87,  41, 304, 117,  59,  53,  58,
       116,  42, 321,  38,  56,  49, 317,   6,  57,  19,  25, 315, 123,
        46,  89,  61, 312, 299, 130,  74, 298, 119,  20, 286, 136, 129,
       124, 327, 131, 460, 140, 114, 139, 122, 137, 126, 120, 128, 135,
       150, 143, 151, 132, 125, 157, 147, 138, 156, 164, 346, 159, 160,
       161, 333, 381, 149, 154, 297, 163, 314, 155, 323, 340, 356, 142,
       328, 144, 336, 248, 302, 175, 344, 382, 146, 170, 166, 338, 167,
       310, 148, 165, 172, 171, 145, 121, 178, 305, 173, 152, 354, 347,
       158, 185, 349, 183, 352, 177, 200, 192, 361, 207, 174, 330, 134,
       350, 334, 283, 153, 197, 133, 241, 193, 235, 194, 261, 260, 216,
       169, 209, 238, 215, 141, 189, 187, 223, 284, 214, 202, 211, 168,
       230, 203, 188, 232, 709, 219, 162, 196, 190, 259, 228, 176, 250,
       201, 186, 199, 180, 206, 205, 224, 222, 182, 210, 275, 212, 229,
       218, 208, 191, 181, 179, 246, 255, 226, 288, 253, 252, 262, 236,
       256, 234, 254, 468, 213, 237, 198, 195, 239, 263, 265, 274, 217,
       220, 307, 221, 233, 257, 227, 276, 225, 264, 311, 277, 204, 290,
       266, 270, 294, 319, 282, 251, 322, 291, 269, 240, 271, 184, 231,
       268, 247, 273, 300, 301, 267, 244, 306, 293, 309, 272, 242, 295,
       285, 243, 308, 398, 303, 245, 424, 279, 331, 281, 339, 434, 357,
       325, 329, 278, 332, 343, 345, 360, 348, 367, 353, 373, 374, 406,
       400, 326, 379, 399, 316, 341, 320, 385, 355, 363, 358, 296, 422,
       390, 335, 370, 376, 375, 397, 289, 542, 403, 383, 384, 359, 393,
       337, 362, 365, 435, 386, 378, 313, 351, 287, 471, 462, 411, 450,
       318, 372, 371, 454, 532, 445, 389, 388, 407, 443, 437, 451, 391,
       405, 412, 419, 420, 426, 433, 440, 429, 418, 447, 461, 605, 457,
```

```
        475, 464, 482, 626, 489, 496, 503, 510, 517, 524, 531, 538, 545,
        552, 559, 566, 573, 580, 587, 594, 601, 608, 615, 622, 629, 396,
        410, 395, 423, 408, 409, 448, 465, 387, 414, 476, 479, 467, 490,
        493, 478, 504, 507, 458, 518, 521, 377, 444, 380, 463]),
 'arrival_date_year': array([2015, 2016, 2017]),
 'arrival_date_month': array(['July', 'August', 'September', 'October', 'November',
'December',
        'January', 'February', 'March', 'April', 'May', 'June'],
       dtype=object),
 'arrival_date_week_number': array([27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,
39, 40, 41, 42, 43,
        44, 45, 46, 47, 48, 49, 50, 51, 52, 53,  1,  2,  3,  4,  5,  6,  7,
         8,  9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
        25, 26]),
 'arrival_date_day_of_month': array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12,
13, 14, 15, 16, 17,
        18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]),
 'stays_in_weekend_nights': array([ 0,  1,  2,  4,  3,  6, 13,  8,  5,  7, 12,  9,
16, 18, 19, 10, 14]),
 'stays_in_week_nights': array([ 0,  1,  2,  3,  4,  5, 10, 11,  8,  6,  7, 15,  9,
12, 33, 20, 14,
        16, 21, 13, 30, 19, 24, 40, 22, 42, 50, 25, 17, 32, 26, 18, 34, 35
```

3. ***Data Wrangling***

Data Wrangling Code

```
# Write your code to make your dataset analysis ready.
df1=df.copy() #creating a copy of dataframe

# we will remove those rows where adults, children and babies equal to zero
df1.drop(df1[df1['adults']+df1['babies']+df1['children']==0].index,inplace=True)

#coverting data type of adults ,babies,children to int
df1[['adults','babies','children']]=df1[['adults','babies','children']].astype(int)

#converting reservation status date to datetime object
df1['reservation_status_date']=df1['reservation_status_date'].apply(lambda x:datetime.strptim

#addting total stay and total people columns
df1['total_stays']=df1['stays_in_weekend_nights']+df['stays_in_week_nights']
df1['total_people']=df1['adults']+df1['children']+df1['babies']
```

What all manipulations have you done and insights you found?

We are given dataset of 119390 observations which is having 32 columns. Performed feature engineering for the given dataset, while doing feature engineering I have found missing values for

some columns and replaced them with appropriate values.I have found a lot of duplicate data in observations and successfully dropped them from the dataset to make calculations easy.

I have figured out the categorical and non-categorical variables.I have changed the data type of for some columns which were need to be changed.I have created new columns which will be helpful for our calculations while performing Exploratory Data Analysis.

4. Data Vizualization, Storytelling & Experimenting with charts : Understand the relationships between variables

Chart - 1

```
# Chart - 1 visualization code
#Which agent makes the most no. of bookings
plt.rcParams['figure.figsize']=(10,5)
top_agent=df1['agent'].value_counts()
top_agent.iloc[:10].plot(kind='bar')
plt.title('Top agent with most bookings')
plt.xlabel('Agent IDs')
plt.ylabel('No of bookings')
```

```
Text(0, 0.5, 'No of bookings')
```

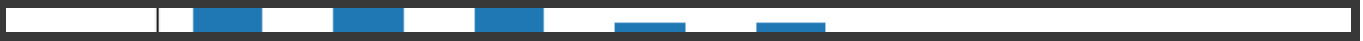
Agent with most number of bookings is "9.0"

1. Why did you pick the specific chart?

Using barplot it's easy to find the max value

2. What is/are the insight(s) found from the chart?

From the above chart it's shown that agent 9 has made most number of bookings ,and he is top agent in suggesting resort to customers



3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

It'll have an positive Impact as if hotels can encourage the top agents they will bring more customers.

Chart - 2

```
# Chart - 2 visualization code
#Which meal type is the  most preffered meal of customers
sns.countplot(x=df1['meal'])
```

```
<Axes: xlabel='meal', ylabel='count'>
```

The most preferred meal is "BB"

1. Why did you pick the specific chart?

Countplot chart will denote count of each observation in different colors and it's easy to understand

2. What is/are the insight(s) found from the chart?

From the chart we found that most preferred meal for customers is "BB" , "HB meal" and "SC meal" are second preference

3. Will the gained insights help creating a positive business impact?

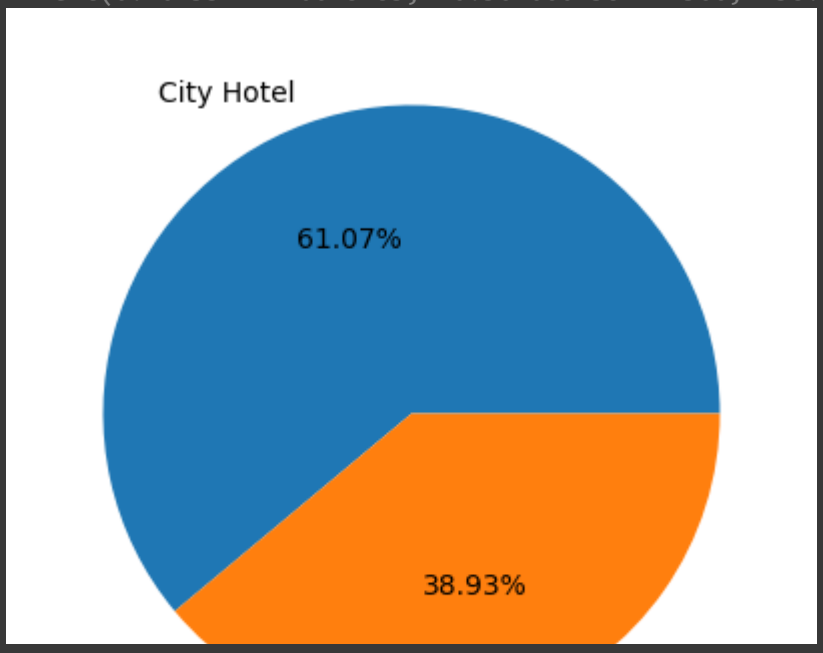
Are there any insights that lead to negative growth? Justify with specific reason.

It'll have an positive impact as hotels can increase the making quantity of 'BB meal' from these insights

Chart - 3

```
# Chart - 3 visualization code
#what is the percentage of bookings in each hotel
hotel_bookings=df1.groupby('hotel').size()
values=hotel_bookings.values
labels=hotel_bookings.index
plt.pie(values,labels=labels,autopct='%1.2f%%')
```

```
([<matplotlib.patches.Wedge at 0x7fc8acbc9ed0>,
  <matplotlib.patches.Wedge at 0x7fc8acbc9ff0>],
 [Text(-0.374985039033291, 1.0341113192017586, 'City Hotel'),
  Text(0.37498494221280004, -1.0341113543103873, 'Resort Hotel')],
 [Text(-0.2045372940181587, 0.5640607195645955, '61.07%'),
  Text(0.20453724120698183, -0.5640607387147566, '38.93%')])
```



1. Why did you pick the specific chart?

While comparing percentage between variables pie chart will provide the best visualization

2. What is/are the insight(s) found from the chart?

From the pie chart I have found that city hotel has booking percentage of 61.13 while resort hotel holds a percentage of 38.87

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Hotels can see the percentage of bookings and get better in their maintenance if needed.

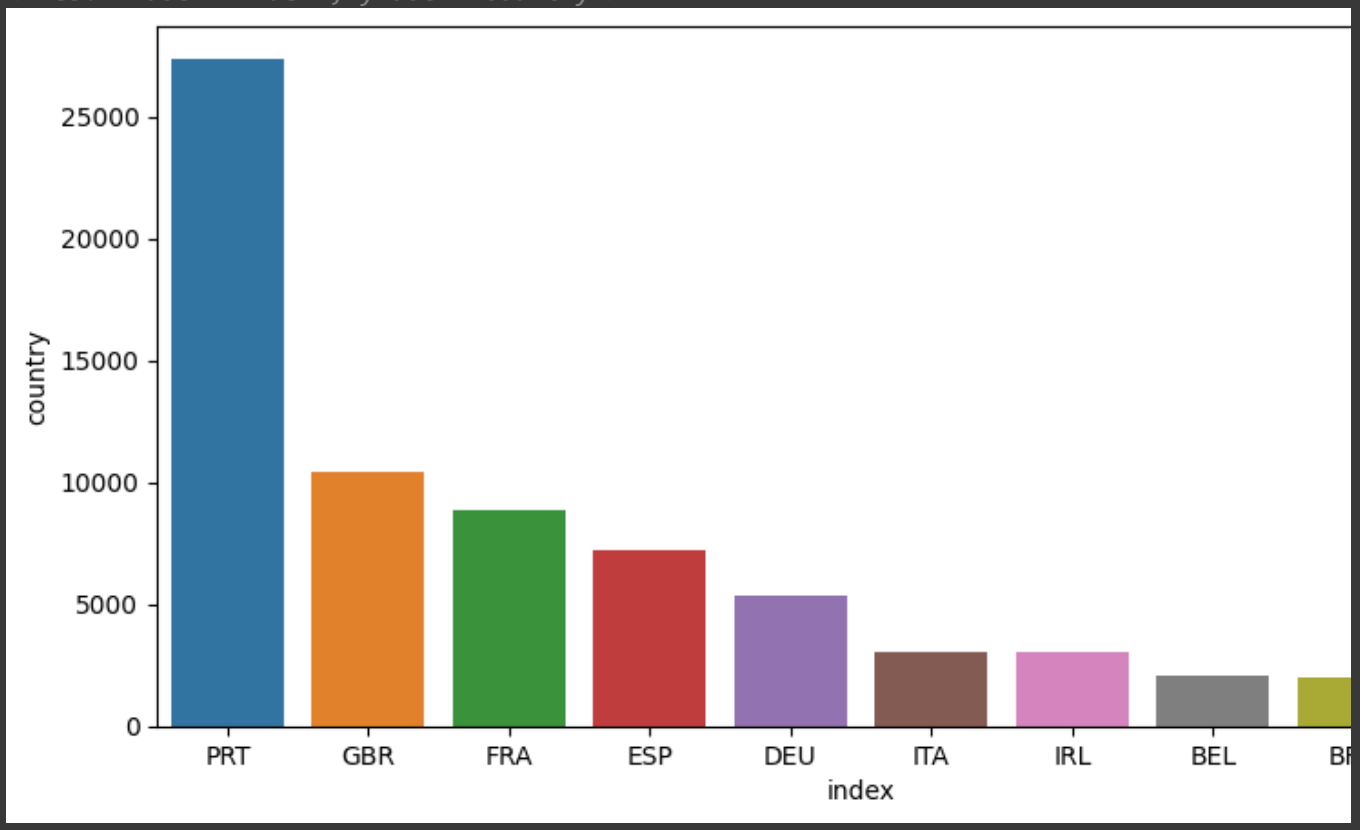
Chart - 4

```
# Chart - 4 visualization code

#from which country most of the guests are coming ?
```

```
country=df1['country'].value_counts().sort_values(ascending=False)[0:10].reset_index()
sns.barplot(x='index',y='country',data=country)
```

```
<Axes: xlabel='index', ylabel='country'>
```



1. Why did you pick the specific chart?

Barplot gives good visualiztions while comparing more than 5 rows

2. What is/are the insight(s) found from the chart?

From the chart we have found that most no of guests are from PRT i.e Portugal

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

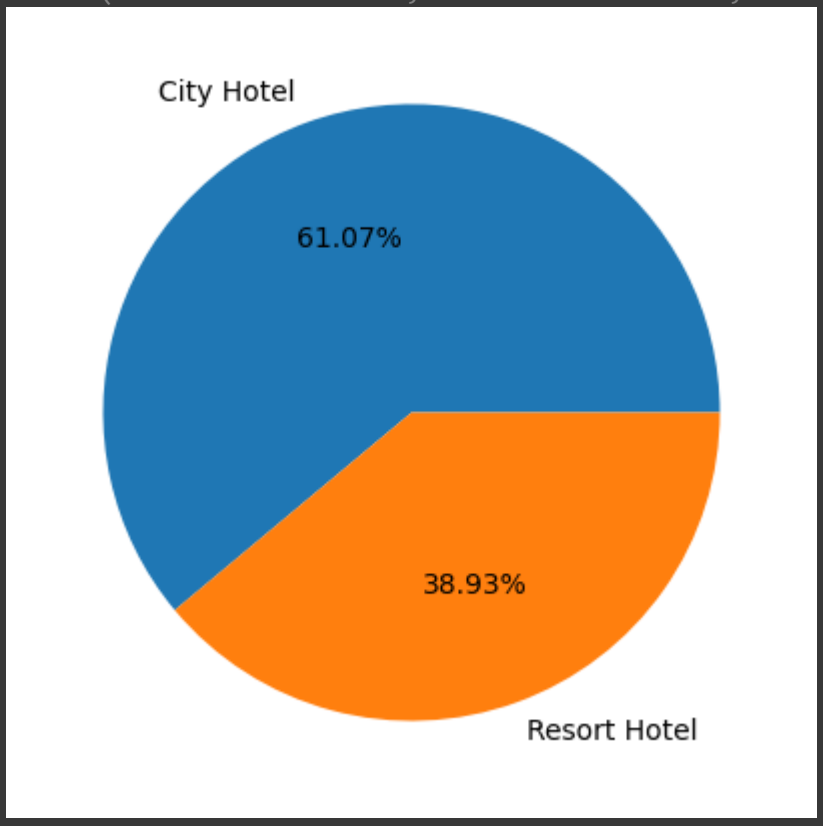
It'll have an positive impact by knowing the country of most no of guests Hotels can prepare special authentic food for those guests

Chart - 5

```
# Chart - 5 visualization code
#Which hotel has higher bookings cancellation rate


cancellations=df1.groupby('hotel')['previous_cancellations'].size()
plt.pie(cancellations.values,labels=cancellations.index,autopct='%1.2f%%')
```

```
([<matplotlib.patches.Wedge at 0x7fc8aca00130>,
  <matplotlib.patches.Wedge at 0x7fc8acaff400>],
 [Text(-0.374985039033291, 1.0341113192017586, 'City Hotel'),
  Text(0.37498494221280004, -1.0341113543103873, 'Resort Hotel')],
 [Text(-0.2045372940181587, 0.5640607195645955, '61.07%'),
  Text(0.20453724120698183, -0.5640607387147566, '38.93%')])
```



1. Why did you pick the specific chart?

as we are comparing the cancellation percentage in between city and resort hotels

2. What is/are the insight(s) found from the chart?

The city hotel has cancellation percentage of 61.13% The resort hotel has cancellation percentage of 38.87%

3. Will the gained insights help creating a positive business impact?

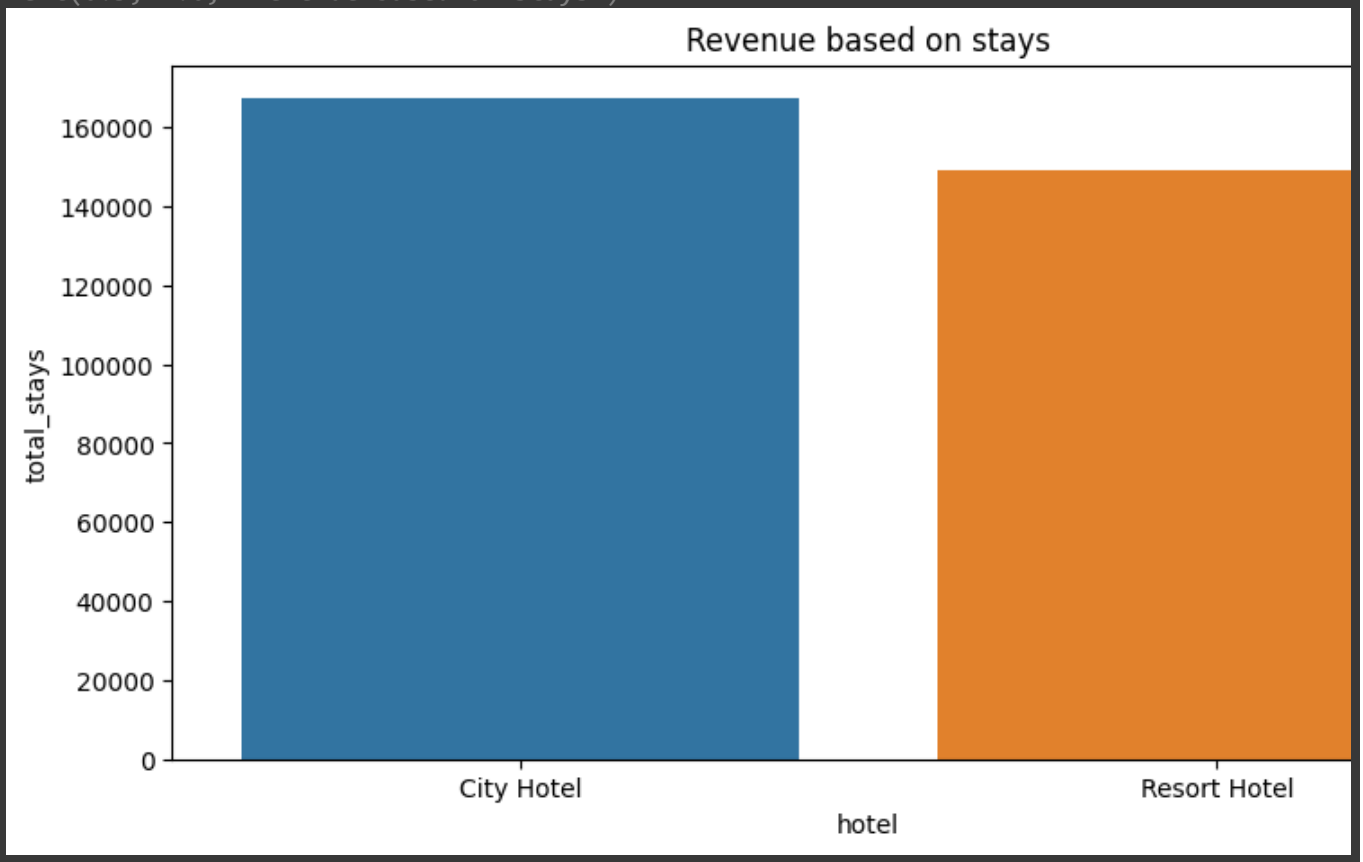
Are there any insights that lead to negative growth? Justify with specific reason.

It'll help the hotels to look into their service and to think about the ways to reduce the percentage of cancellations.

Chart - 6

```
# Chart - 6 visualization code
#Which hotel seems to make more revenue
#stays_in_week_nights #stays_in_week_nights
stays_for_revenue=df1.groupby('hotel')['total_stays'].sum().reset_index()
sns.barplot(x='hotel',y='total_stays',data=stays_for_revenue)
plt.title('Revenue based on stays')
```

```
Text(0.5, 1.0, 'Revenue based on stays')
```



1. Why did you pick the specific chart?

as we are comparing total stays of each hotel

2. What is/are the insight(s) found from the chart?

City hotel has more stays which means it generates more revenue than resort hotel

3. Will the gained insights help creating a positive business impact?

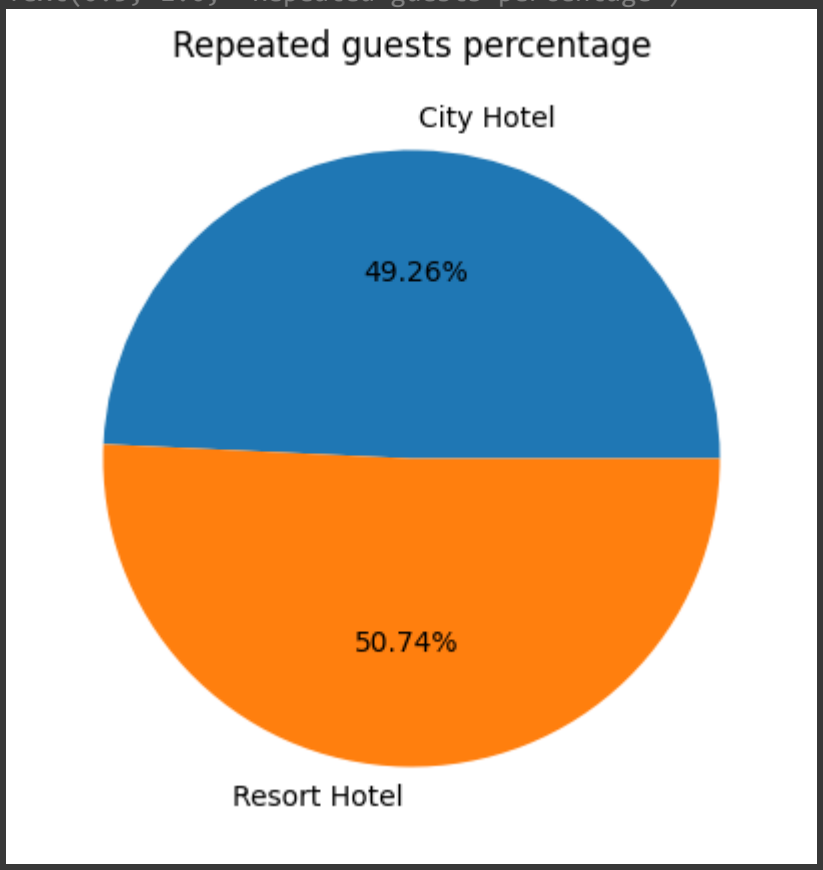
Are there any insights that lead to negative growth? Justify with specific reason.

by knowning each hotel's revenue it'll help hotels to maintain accounts

Chart - 7

```
# Chart - 7 visualization code
#Which hotel has a  high chance that its customer will return for another stay ?
repeated_guests=df1.groupby('hotel')['is_repeated_guest'].value_counts()
plt.pie(repeated_guests.loc[['City Hotel','Resort Hotel'],1].values,labels=['City Hotel','Res
plt.title("Repeated guests percentage")
```

```
Text(0.5, 1.0, 'Repeated guests percentage')
```



1. Why did you pick the specific chart?

pie chart will give better results while performing percentage operations

2. What is/are the insight(s) found from the chart?

Observations we have found from the above chart are that both hotels have almost same percentage of repeated guests yet with a slight difference resort hotel has more percentage of repeated guests

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Hotels can get the feedback from repeated guests as they are preferring to book again and again obviously the feedback is going to be positive and it'll help to boost up the business

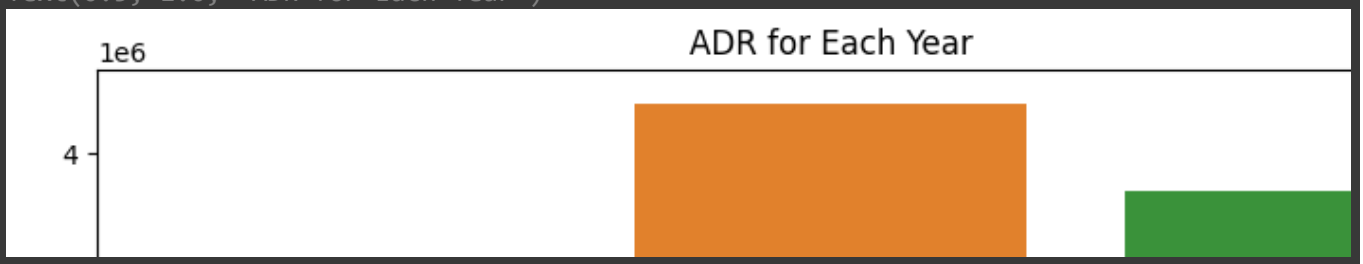
Chart - 8

```
# Chart - 8 visualization code

#which year has generated most ADR

adr_df=df1.groupby('arrival_date_year')['adr'].sum().reset_index()
sns.barplot(x='arrival_date_year',y='adr',data=adr_df)
plt.title('ADR for Each Year')
```

```
Text(0.5, 1.0, 'ADR for Each Year')
```



1. Why did you pick the specific chart?

to visualize the most ADR generated year

2. What is/are the insight(s) found from the chart?

The most ADR was generated in 2016 and the least in in 2015

3. Will the gained insights help creating a positive business impact?

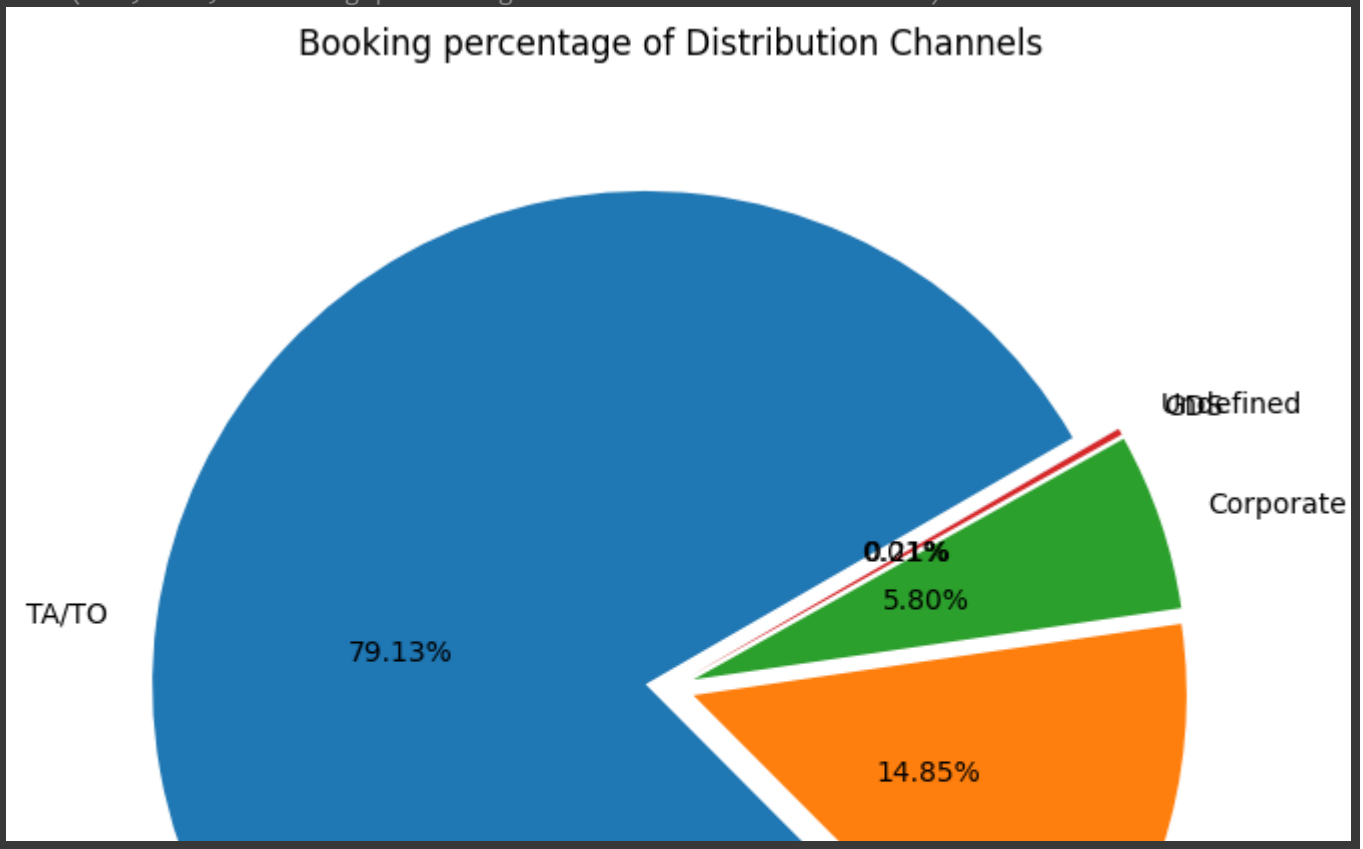
Are there any insights that lead to negative growth? Justify with specific reason.

It'll be helpful for hotels in maintaining their accoutns

Chart - 9

```
# Chart - 9 visualization code
# Which is the most common channel for booking hotels?
channels=df1['distribution_channel'].value_counts()
plt.figure(figsize = (10,8))
plt.pie(channels.values,labels=channels.index,autopct="%0.2f%%",pctdistance=0.5,explode=[0.05
plt.title("Booking percentage of Distribution Channels")
```

```
Text(0.5, 1.0, 'Booking percentage of Distribution Channels')
```



1. Why did you pick the specific chart?

to visualize the Booking percentage of Distribution Channels

2. What is/are the insight(s) found from the chart?

The TA/TO distribution channel has more percentage of bookings

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

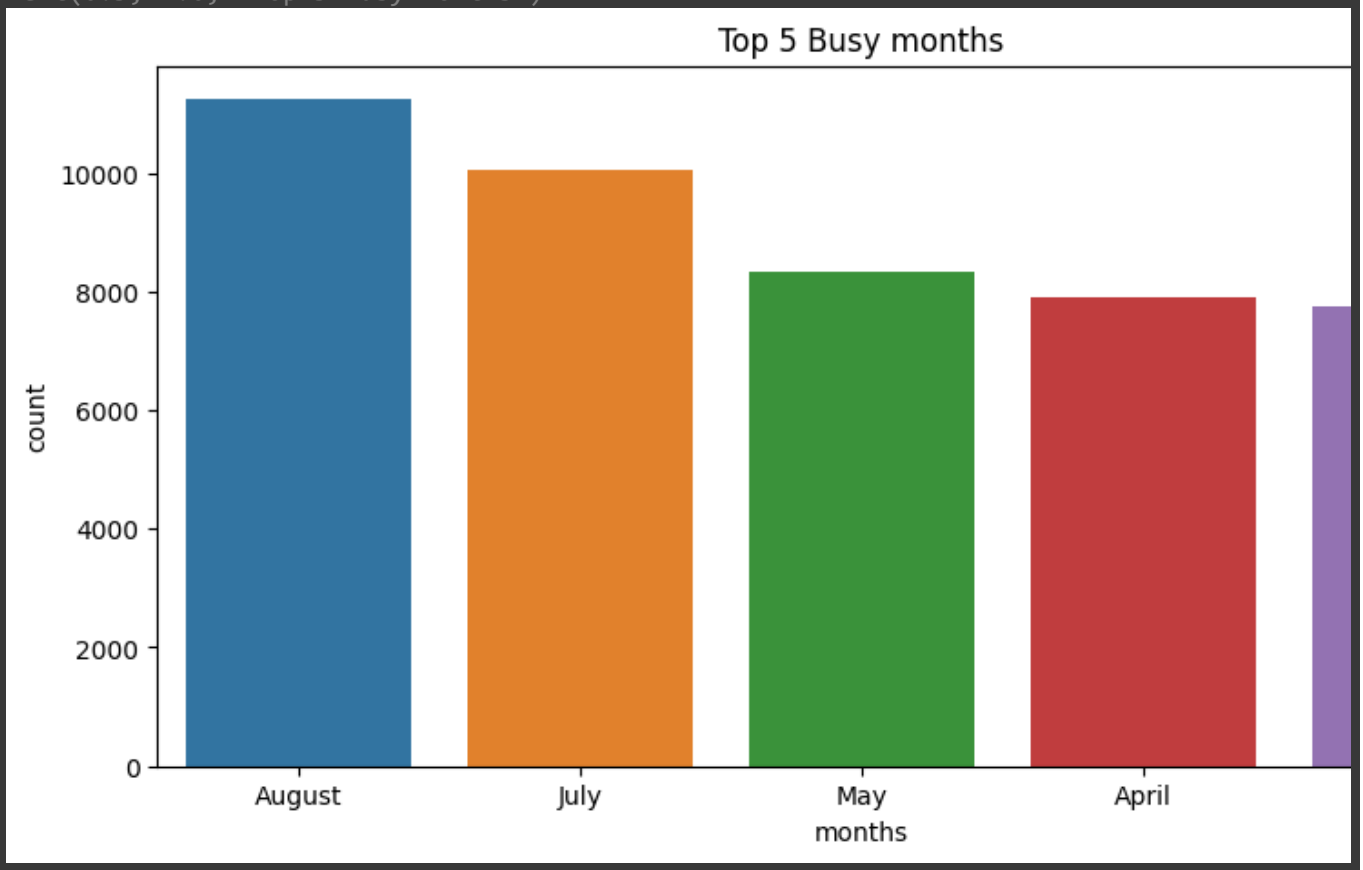
Hotels can attract more guests through others channels as well by providing any kind of discounts

Chart - 10

```
# Chart - 10 visualization code
#Which are the most busy months?
months_df=df1['arrival_date_month'].value_counts().reset_index()[0:5]
months_df.rename(columns={'index':'months','arrival_date_month':'count'},inplace=True)
```

```
sns.barplot(x='months',y='count',data=months_df)
plt.title('Top 5 Busy months')
```

```
Text(0.5, 1.0, 'Top 5 Busy months')
```



1. Why did you pick the specific chart?

To compare month variable values

2. What is/are the insight(s) found from the chart?

August is the most busy month and July as well with a slight difference

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

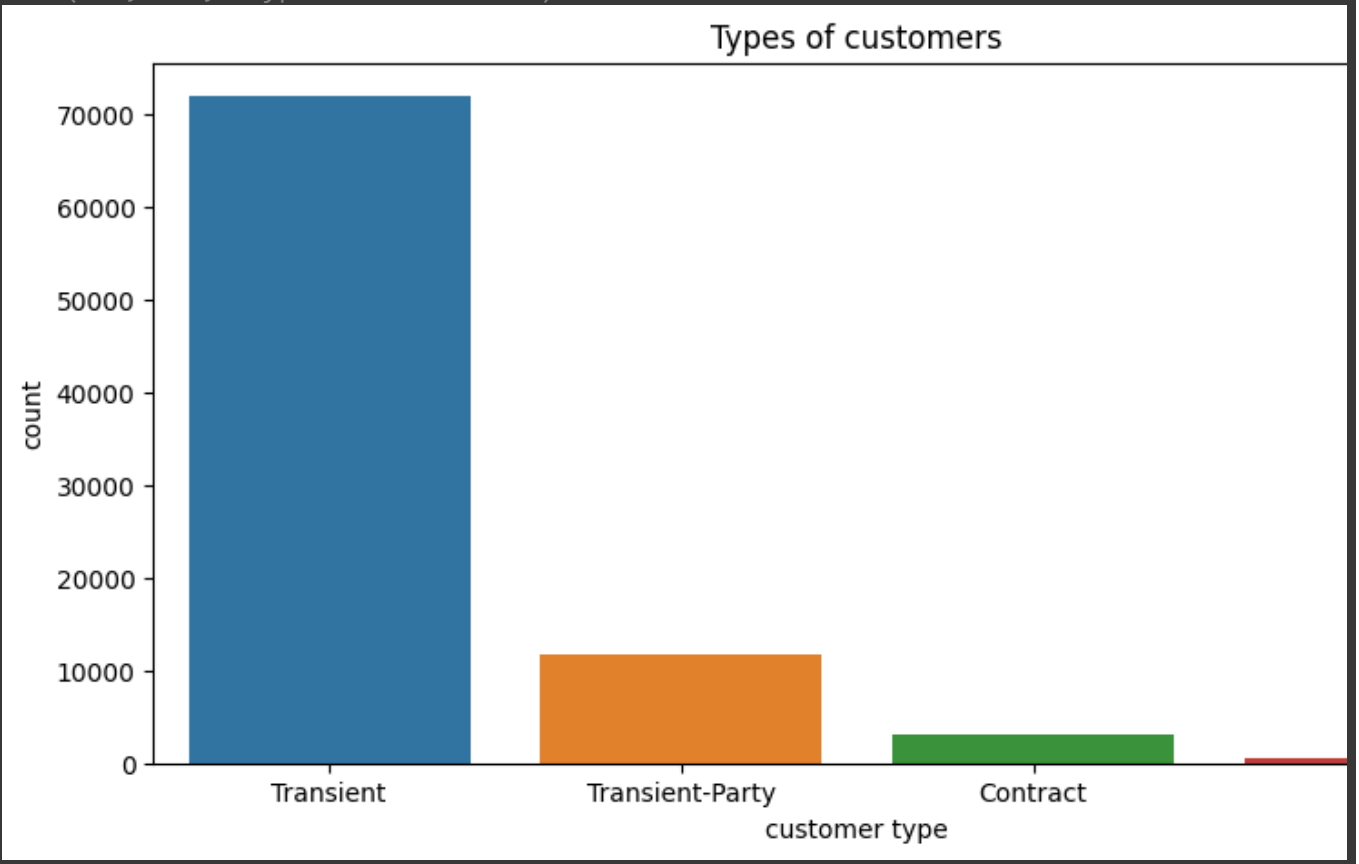
Hotels can prepare the arrangements accordingly prior to the the busy month by observing the data

Chart - 11

```
# Chart - 11 visualization code
#Which types of customers mostly make bookings?

cust_df=df1['customer_type'].value_counts().reset_index()
cust_df.rename(columns={'index':'customer type','customer_type':'count'},inplace=True)
sns.barplot(x='customer type',y='count',data=cust_df)
plt.title('Types of customers')
```

```
Text(0.5, 1.0, 'Types of customers')
```



1. Why did you pick the specific chart?

to visualize the type of customers

2. What is/are the insight(s) found from the chart?

Transient type of customers has made more number of bookings

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

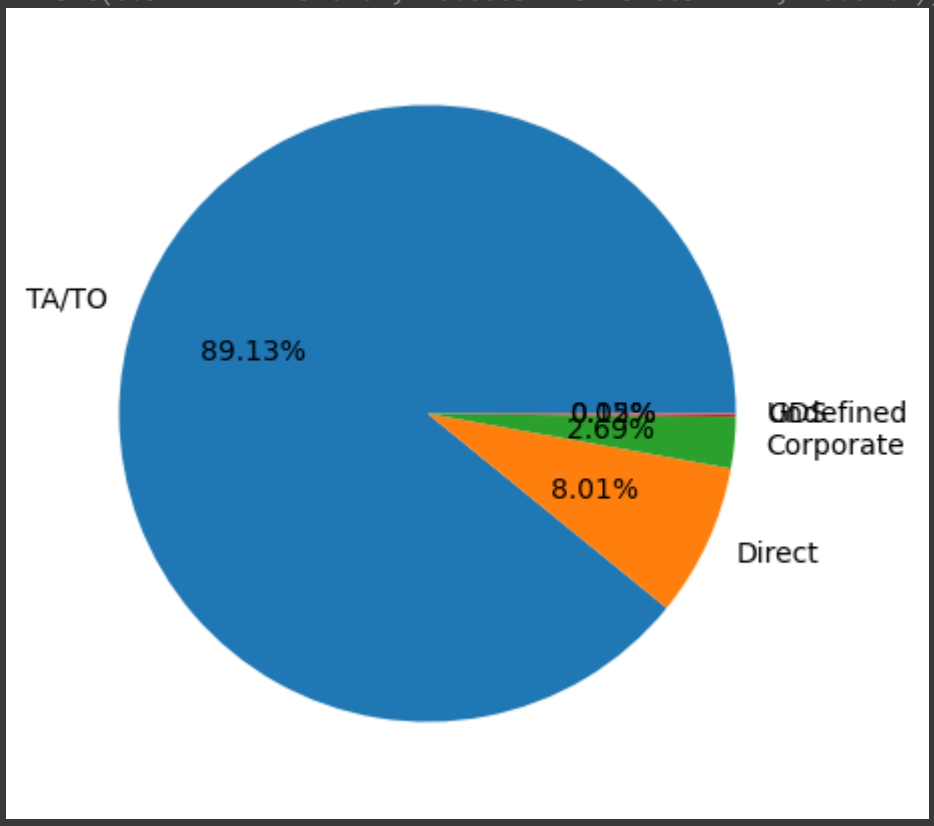
Hotels can offer the service according to the customer type and their majority

Chart - 12

```
# Chart - 12 visualization code
#Which significant distribution channel has the highest cancellation percentage?

dist=df1.groupby('distribution_channel')['is_canceled'].sum().sort_values(ascending=False)
plt.pie(dist.values,labels=dist.index,autopct="%1.2f%%")
```

```
([<matplotlib.patches.Wedge at 0x7fc8aa6d2a70>,
  <matplotlib.patches.Wedge at 0x7fc8aa6d2950>,
  <matplotlib.patches.Wedge at 0x7fc8aa6d3760>,
  <matplotlib.patches.Wedge at 0x7fc8aa6b3c70>,
  <matplotlib.patches.Wedge at 0x7fc8aa6d3e50>],
 [Text(-1.036519460017235, 0.36827626723097423, 'TA/TO'),
  Text(0.9993329451007046, -0.45971041410474084, 'Direct'),
  Text(1.0950402172815155, -0.10434041659899335, 'Corporate'),
  Text(1.099981767438612, -0.006333348453027957, 'GDS'),
  Text(1.099999849215832, -0.0005759558548434116, 'Undefined')],
 [Text(-0.5653742509184917, 0.20087796394416776, '89.13%'),
  Text(0.545090697327657, -0.25075113496622226, '8.01%'),
  Text(0.5972946639717357, -0.05691295450854182, '2.69%'),
  Text(0.5999900549665156, -0.0034545537016516126, '0.15%'),
  Text(0.5999999177540901, -0.00031415773900549717, '0.02%')])
```



1. Why did you pick the specific chart?

to find the highest cancellation percentage among the distribution channels

2. What is/are the insight(s) found from the chart?

TA/TO channel has the highest cancellation percentage

Chart - 13 - Correlation Heatmap

```
# Correlation Heatmap visualization code


list_of_numeric=list(df1.describe())

#we can remove the columns adults,children,babies,stays in week nights and stays in weekend n
#columns like 'is_cancelled', 'arrival_date_year', 'arrival_date_week_number', 'arrival_date_

columns_to_be_removed=['adults','babies','children','stays_in_weekend_nights','stays_in_week_
corr_list=[x for x in list_of_numeric if x not in columns_to_be_removed]
sns.heatmap(df1[corr_list].corr(),annot=True)
plt.title('Correlation Heat Map')
```

```
Text(0.5, 1.0, 'Correlation Heat Map')
```

1. What is/are the insight(s) found from the chart?

Total stay length and lead time have slight correlation. This may mean that for longer hotel stays people generally plan little before the the actual arrival.

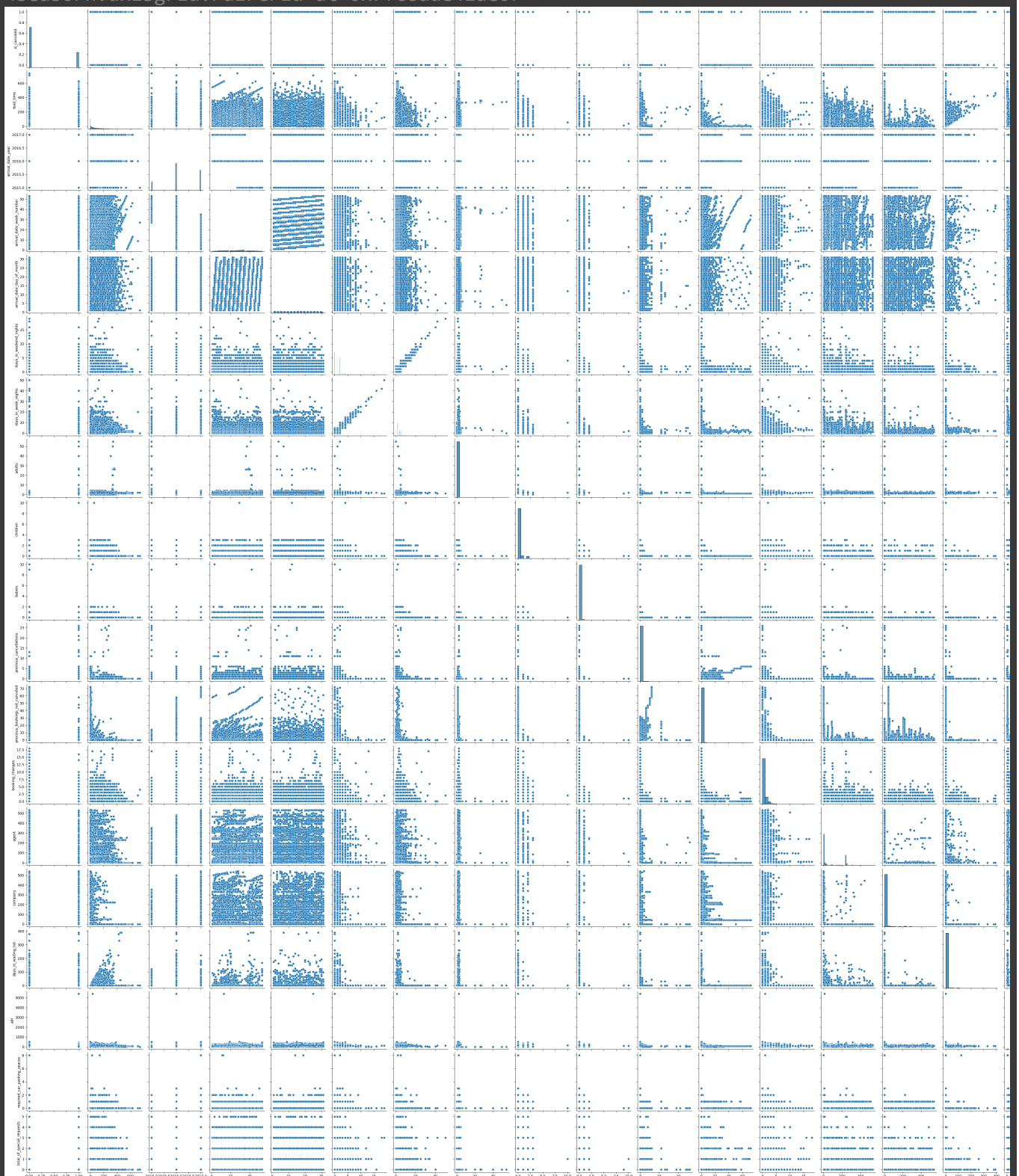
adr is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.

Chart - 15 - Pair Plot

```
# Pair Plot visualization code
#sns.pairplot(data=df1[corr_list])
numerical_columns = ['is_canceled', 'lead_time', 'arrival_date_year', 'arrival_date_week_numb
                     'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_n
                     'adults', 'children', 'babies', 'previous_cancellations', 'previous_book
                     'booking_changes', 'agent', 'company', 'days_in_waiting_list', 'adr', 'r
                     'total_of_special_requests']

# create a pair plot
sns.pairplot(data=df1[numerical_columns])
```

```
<seaborn.axisgrid.PairGrid at 0x7fc8ade41ae0>
```



1. What is/are the insight(s) found from the chart?

The diagonal of the pair plot shows the distribution of each variable, while the off-diagonal plots show the pairwise relationships between variables. For example, the relationship between lead time and stays in week nights. We can see that there is a positive correlation between these variables - as lead time increases, stays in week nights tend to increase. We can also see that there are some interesting patterns in the plots involving the booking status variable, which could be useful for predicting cancellations or no-shows. Overall, the pair plot is a useful visualization for exploring the relationships between different variables in the dataset.

5. Solution to Business Objective

What do you suggest the client to achieve Business Objective ?

Explain Briefly.

1.Maximize revenue: A common business objective for hotels is to maximize revenue. To achieve this objective, the hotel could analyze the data to identify the most profitable room types, rates, and booking channels. The hotel could also analyze seasonal patterns to optimize pricing and promotions during high-demand periods.

2.Reduce cancellations: Another common business objective is to reduce cancellations, which can lead to lost revenue and operational inefficiencies. To achieve this objective, the hotel could analyze the data to identify the most common reasons for cancellations and take steps to address those issues. For example, the hotel could offer flexible cancellation policies or improve the booking experience to reduce the likelihood of cancellations.

3.Improve customer satisfaction: A key business objective for hotels is to provide a positive guest experience that leads to repeat business and positive reviews. To achieve this objective, the hotel could analyze guest feedback data to identify areas for improvement, such as room cleanliness, staff friendliness, or check-in efficiency. The hotel could also use data to personalize the guest experience, such as by offering customized amenities or room preferences.

4.Optimize resource allocation: Hotels have many operational resources to manage, including staff, inventory, and facilities. To achieve this objective, the hotel could use data to optimize resource allocation based on demand patterns. For example, the hotel could adjust staffing levels based on occupancy rates, or optimize inventory levels based on seasonal demand for specific amenities or services.

Overall, the key to achieving business objectives through data analysis is to use the insights gained from the data to make informed decisions and take action. By using data to drive decision-making, hotels can improve efficiency, maximize revenue, and provide a better guest experience.

Conclusion

1)Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. Also the overall adr of City hotel is slightly higher than Resort hotel.

2) Mostly guests stay for less than 5 days in hotel and for longer stays Resort hotel is preferred.

3) Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.

4) Most of the guests came from european countries, with most of guests coming from Portugal.

5) Guests use different channels for making bookings out of which most preferred way is TA/TO.

6) For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.

7) Almost 30% of bookings via TA/TO are cancelled.

8) July- August are the most busier and profitable months for both of hotels.

9) More number of people in guests results in more number of special requests.

10) For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.