Project Name - Hotel Booking Analysis

Project Type - EDA

Contribution - Team

Team Member 1 - Omendra Puri

Team Member 2 - Govind Kumar Choudhary

Project Summary -

Study of Hotel bookings related data is vital for any hospitality business as it gives insight into booking behaviour of customers as well as channels through which bookings are made.

This project involves the analysis of the provided data set related to Hotel bookings for City and Resort Hotels.

Project activities have been categorised as under:

1. Defining the problem statement i.e. Business Objective of the study.
2. Collection and preparation of data by data cleaning , treating outliers etc.
3. Perform exploratory data analysis(EDA) through a deep study of relationship between different features, generate new variables based on need inline with related business objectives. Present the data in easily understandable form.
4. Provide observations as well as recommendations based on the EDA.

GitHub Link -

Provide your GitHub Link here.

Problem Statement

** **Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions!**

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that govern the bookings. **

Define Your Business Objective?

The objective of this data set is to gain insights into hotel booking patterns and cancellations, and to identify the factors that influence these patterns. The goal is to develop predictive models that can accurately predict booking cancellations and to identify potential areas for improvement in hotel policies and practices. Ultimately, the objective is to increase revenue and profitability for hotels by reducing booking cancellations and improving overall customer satisfaction.

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 be executable in one go
             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

Import Libraries

```
# Import Libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import folium
```

Dataset Loading

```
# Load Dataset
hotel_booking_df = pd.read_csv('/content/Hotel Bookings.csv')

# copy original dataset to new dataset
df = hotel_booking_df.copy()
```

Dataset First View

```
# Dataset First Look
df.head(5)
```

Dataset Rows & Columns count

```
# Dataset Rows & Columns count
df.shape
```

```
# Dataset Rows & Columns count
print("Dataset_Row_count:    ",df.shape[0])
print("Dataset_Column_count: ",df.shape[1])
```

Dataset Information

```
# Dataset Info
df.info()
```

Duplicate Values

```
# Dataset Duplicate Value Count
df.duplicated().value_counts()           # True means duplicated rows
```

```
# Visualizing the duplicate values
plt.figure(figsize=(5,4))
sns.countplot(x=df.duplicated())
```

So we have 31994 are duplicate row in our dataset

Missing Values/Null Values

```
# Missing Values/Null Values Count
Missing_Values = df.isnull().sum().sort_values(ascending=False)
Missing_Values[:5]
```

```
import missingno as msno

# Visualize missing values as a matrix
msno.bar(df)

# Display the plot
plt.show()
```

What did you know about your dataset?

- This dataset conatins 119390 rows and 32 columns
- All the columns are diveded into three Dtypes(Object, Float64 and Int64)
- This dataset have duplicate as well as missing values. There are 31994 duplicate values and four columns have missing values.
- The missing columns are company, agent, country and children and the maximum missing values are 112593, 16340, 488 & 4 respectively.

2. Understanding Your Variables

```
# Dataset Columns
list(df.columns)
```

```
# Dataset Describe
df.describe()
```

Variables Description

1. **hotel** : *Resort Hotel or City Hotel*

2. **is_canceled** : *Value indicating if the booking was canceled (1) or not (0)*
3. **lead_time** :*The number of days between the booking date and the arrival date*
4. **arrival_date_year** : *Year of arrival date*
5. **arrival_date_month** : *Month of arrival date*
6. **arrival_date_week_number** : *Week number of year for arrival date*
7. **arrival_date_day_of_month** : *Day of arrival date*
8. **stays_in_weekend_nights** : *Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel*
9. **stays_in_week_nights** : *Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel*
10. **adults** : *Number of adults*
11. **children** : *Number of children*
12. **babies** : *Number of babies*
13. **meal** : *The type of meal booked (e.g., Bed & Breakfast, Half board):*
14. **country** : *Country of origin.*
15. **market_segment** : *Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"*
16. **distribution_channel** : *Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"*
17. **is_repeated_guest** : *Value indicating if the booking name was from a repeated guest (1) or not (0)*
18. **previous_cancellations** : *Number of previous bookings that were cancelled by the customer prior to the current booking*
19. **previous_bookings_not_canceled** : *Number of previous bookings not cancelled by the customer prior to the current booking*
20. **reserved_room_type** : *Code of room type reserved. Code is presented instead of designation for anonymity reasons.*
21. **assigned_room_type** : *Code for the type of room assigned to the booking.*
22. **booking_changes** : *Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation*
23. **deposit_type** : *Indication on if the customer made a deposit to guarantee the booking.*
24. **agent** : *ID of the travel agency that made the booking*
25. **company** : *ID of the company/entity that made the booking or responsible for paying the booking.*
26. **days_in_waiting_list** : *Number of days the booking was in the waiting list before it was confirmed to the customer*
27. **customer_type** : *Type of booking, assuming one of four categories*
28. **adr** : *The average daily rate (i.e., the sum of all lodging transactions divided by the total number of staying nights)*
29. **required_car_parking_spaces** : *Number of car parking spaces required by the customer*
30. **total_of_special_requests** :*Number of special requests made by the customer (e.g. twin bed or high floor)*
31. **reservation_status** : *Reservation last status, assuming one of three categories*

- Canceled – booking was canceled by the customer
- Check-Out – customer has checked in but already departed
- No-Show – customer did not check-in and did inform the hotel of the reason why

32. **reservation_status_date** : *Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel*

Check Unique Values for each variable.

```
# Check Unique Values for each variable.
unique_values = df.nunique()
unique_values
```

3. ***Data Wrangling***

Data Wrangling Code

*** Data Cleaning**

```
# To fill the null value in the column, let's check which columns has null value, we have all ready store the same
Missing_Values[:5]
```

```
# let's check, what is the percentage of null value in each column
percent_missing = Missing_Values * 100 / len(df)
percent_missing[:5]
```

```
# It is better to drop company column there are extremely high values are missing compared to the number of rows
df.drop(['company'], axis=1, inplace=True)
```

```
# Replacing null values of agent and children with value 0
df[['agent','children']] = df[['agent', 'children']].fillna(0)
```

```
# Replacing null values of country column with other
df[['country']] = df[['country']].fillna('other')
```

```
#Checking
df.isnull().sum().sort_values(ascending=False)[:4]
```

```
# Drop the duplicate value
df=df.drop_duplicates()
df
```

```
# Checking the shape of the dataset whose combining value of adults, babies and children column is 0
df[df['adults']+df['children']+df['babies']==0].shape
```

```
# Checking the shape of updated data set
df.shape
```

```
# Creating the copy of the dataset for further analysis
hotel_booking_df1 = df.copy()
```

```
# Dropping the row where combining values of adults, babies and children is 0 beacause there is no booking
hotel_booking_df1.drop(hotel_booking_df1[hotel_booking_df1['adults']+hotel_booking_df1['babies']+hotel_booking_df1['children']
```

```
hotel_booking_df1.shape      # checking row is drop
```

```
# Checking total drop row
```

```
87389-87223
```

```
# Checking datetype of column 'reservation_status_date' from object to date_type
hotel_booking_df1['reservation_status_date'] = pd.to_datetime(hotel_booking_df1['reservation_status_date'],format='%Y-%m-%d')
```

```
# Changing agent and children data type float64 into int64
hotel_booking_df1[['agent','children']]=hotel_booking_df1[['agent','children']].astype('int64')
```

```
hotel_booking_df1.info()  # For cheking changes in reservation_status_date,agent and children datatype
```

adding some important columns

```
# Adding total stay day in hotel
hotel_booking_df1['total_stay'] = hotel_booking_df1['stays_in_week_nights'] + hotel_booking_df1['stays_in_weekend_nights']


# Adding total number of people
hotel_booking_df1['total_people']=hotel_booking_df1['adults']+hotel_booking_df1['babies']+hotel_booking_df1['children']


# Checking the shape of the dataset
hotel_booking_df1.shape
```

What all manipulations have you done and insights you found?

- Here, Company, Agent, Country and Children columns have missing values. Company columns have more than 94% missing values. So, We drop the company columns. Agent columns have more than 13% missing values and Country and Children columns have less than 1% missing values. So missing values of Agent and Children columns are replace with zero and Country values replace with other.
- Drop the duplicate values.
- Dropping the row where combining values of adults, babies and children is 0 beacause there is no booking
- Adding new columns total_stay day in hotel (stays_in_week_nights + stays_in_weekend_nights)
- And total_people (adults + children + babies)
- New shape of dataset have 87223 rows and 33 columns

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

Chart - 1 Most Preffered Hotel

```
# Chart - 1 visualization code
# create a pie chart
hotel_booking_df1['hotel'].value_counts().plot.pie(figsize=(5,7),fontsize=25, explode=[0.05,0.05], autopct='%1.1f%%', shadow=T

# add a title
plt.title('pie chart for more preffered hotel ')

# show the chart
plt.show()
```

1. Why did you pick the specific chart?

This chart present the data in which hotel have more booking.

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

Here, we have found that city hotel(61.1%) have more booking than resort hotel(38.9%).Hence, using this chart we have say that city hotel have more consumption.

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

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

Yes, for both hotel this chart show some positive bussiness impact.

City hotel: Provided more services to atract more people for more revenue.

Resort hotel: Increase services to atract more people and also see City Hotel for how they are atracts more people for more revenue.

Chart - 2 For Reapeted Guest

```
# Chart - 2 visualization code
# create a pie chart
hotel_booking_df1['is_repeated_guest'].value_counts().plot.pie(figsize=(5,7),fontsize=25, explode=[0.05,0.05], autopct='%1.1f%

# add a title
plt.title('pie chart for repeated guest ')

# show the chart
plt.show()


# repeated guest=1
#not repeated guest=0
#groupby hotel
repeated_guests_df=hotel_booking_df1[hotel_booking_df1['is_repeated_guest']==1].groupby('hotel').size().reset_index().rename(c

#set plot size adn plot barplot
plt.figure(figsize=(10,8))
sns.barplot(x=repeated_guests_df['hotel'],y=repeated_guests_df['number_of_repated_guests'])

# set labels
plt.xlabel('Hotel type')
plt.ylabel('count of repeated guests')
plt.title("Most repeated guests for each hotel")
```

1. Why did you pick the specific chart?

The first chart show percentage of repeated guest and second chart show repeated guest for both hotel.

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

In first chart we find only 3.9 percentage repeated guest and In second chart we see city hotel and resort hotel have almost equal repeated guest.

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

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

Yes, for both hotel these chart make some positive bussiness impact.

By seeing, the charts of repeated guests and most repeated guests for both hotel show that hotel have 3.9% of guests who are repeats and almost equal for both type of hotel. suggesting an opportunity to focus on customer retention and take feedback of customer to improve our services.

Chart - 3 Required Car Parking Spaces

```
# Chart - 3 visualization code
# create a pie chart
hotel_booking_df1['required_car_parking_spaces'].value_counts().plot.pie(figsize=(6,8), legend = (0.85,1), fontsize=12, explod
# add a label
labels = hotel_booking_df1['required_car_parking_spaces'].value_counts().index
# add a title
plt.title('required_car_parking_spaces ')
# show the chart
plt.show()
```

1. Why did you pick the specific chart?

This pie chart show how many percentage people wants car parking space.

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

This pie chart shows 91.6% guest does not required car parking space only 8.3% people wants required car parking spaces.

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

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

Yes, this chart create a positive bussiness impact for both hotel type

Only 8.3% people wants required car parking space. So, the majority of guests do not require a car parking space. hotels may want to consider strategies such as alternative transportation methods and dynamic pricing strategies to optimize the use of parking spaces and generate additional revenue.

Chart - 4 Most Preffered meal Type

```
# Chart - 4 visualization code
# create a bar chart
plt.figure(figsize=(20,6))
plt.subplot(1,2,1)
plt.title('preffered meal type',fontweight='bold',size=18)
sns.barplot (y= list(hotel_booking_df1.meal.value_counts()), x= list(hotel_booking_df1.meal.value_counts().index))
plt.subplot(1,2,2)
ax=hotel_booking_df1.meal.value_counts().plot(kind='bar')
for p in ax.patches:
  ax.annotate(str(p.get_height()),(p.get_x()*1.005,p.get_height()*1.005))

plt.show()
```

1. Why did you pick the specific chart?

This chart shows which meal type is more preffered.

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

Using this chart see that most preffered meal type is BB(67900) followed by SC(9391),HB(9080),Undefined(492) and FB(360).

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

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

Yes, this chart show positive bussiness impact for both hotel.

Most people like BB types meal. So need to improve quality of other meal type and serve more BB type meal for more people come and more revenue.

Chart - 5 Avg ADR of each Hotel type and Avg ADR across Distribution type

```
# Chart - 5 visualization code
group_by_hotel = hotel_booking_df1.groupby('hotel')


# group by hotel adr
highest_adr = group_by_hotel['adr'].mean().reset_index()
plt.figure(figsize=(5,4))
plt.xlabel('hotel type', fontsize=10)
plt.ylabel('adr', fontsize=10)
plt.title("avg adr of each hotel type", fontsize=10)
sns.barplot(x=highest_adr['hotel'],y=highest_adr['adr'])


# Using groupby distribution channel
distribution_channel_df=hotel_booking_df1.groupby(['distribution_channel','hotel'])['adr'].mean().reset_index()
# plot bar chart
plt.figure(figsize=(12,5))
```

```
sns.barplot(x='distribution_channel', y='adr',data=distribution_channel_df,hue='hotel')
plt.title('ADR across Distribution channel')
```

1. Why did you pick the specific chart?

This chart shows average daily rate of both type of hotels.

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

City hotel has the highest adr. This means city hotel generates more revenue than the resort hotel. More the adr more the revenue.

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

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

Double-click (or enter) to edit

Chart - 6 Which agent made highest booking

```
# Chart - 6 visualization code
# return highest bookings made by agents
highest_bookings= hotel_booking_df1.groupby(['agent'])['agent'].agg({'count'}).reset_index().rename(columns={'count': "Most_Bo

 # as agent 0 was NAN value and we replaced it with 0 and indicates no bookings.so droping.
highest_bookings.drop(highest_bookings[highest_bookings['agent']==0].index,inplace=True)

# taking top 10 bookings made by agent
top_ten_highest_bookings=highest_bookings[:10]

top_ten_highest_bookings


#Visualizaing the graph
plt.figure(figsize=(12,5))
sns.barplot(x=top_ten_highest_bookings['agent'],y=top_ten_highest_bookings['Most_Bookings'],order=top_ten_highest_bookings['ag
plt.xlabel('Agent No')
plt.ylabel('Number of Bookings')
plt.title("Most Bookings Made by the agent")
```

1. Why did you pick the specific chart?

This chart show which agent made highest booking.

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

Using this chart see that top ten booking agent where agent 9 made highest booking followed by agent 240,14,7,250,241,28,8,1 and 6.

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

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

Yess, this chart show positive bussiness impact.

Since agent 9 made highest booking. So other agent need to improve their skill to impress more people for hotel booking if possible learn agent 9. Also hotel need to give some offer to agent. Then agent try more people comes to hotel, more revenue wiil be generate.

Chart - 7 Distribution of Customer Type

```
# Chart - 7 visualization
hotel_booking_df1['customer_type'].value_counts().plot.pie(explode=[0.09]*4,shadow=True,autopct='%1.2f%%',figsize=(12,8),fonts
```

```
labels=hotel_booking_df1['customer_type'].value_counts().index.tolist()
plt.title('% Distribution of Customer Type')
plt.legend(bbox_to_anchor=(0.85, 1), loc='upper left', labels=labels)
```

1. Why did you pick the specific chart?

This chart show percentage distribution of customer type.

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

Answer Here

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

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

Answer Here

Chart - 8 Booking by month and Optimal Stay length in hotels

```
# Chart - 8 visualization code
# groupby arrival_date_month and taking the hotel count
bookings_by_months_df=hotel_booking_df1.groupby(['arrival_date_month'])['hotel'].count().reset_index().rename(columns={'hotel'
# Create list of months in order
months = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'Decem
# creating df which will map the order of above months list without changing its values.
bookings_by_months_df['arrival_date_month']=pd.Categorical(bookings_by_months_df['arrival_date_month'],categories=months,order
# sorting by arrival_date_month
bookings_by_months_df=bookings_by_months_df.sort_values('arrival_date_month')

bookings_by_months_df
```

```
# set plot size
plt.figure(figsize=(15,6))

#pltting lineplot on x- months & y- booking counts
sns.lineplot(x=bookings_by_months_df['arrival_date_month'],y=bookings_by_months_df['Counts'])

# set title for the plot
plt.title('Number of bookings across each month')
#set x label
plt.xlabel('Month')
#set y label
plt.ylabel('Number of bookings')
```

```
# Using group by function on total_stay and hotel
stay = hotel_booking_df1.groupby(['total_stay', 'hotel']).agg('count').reset_index()
# Taking only first three columns
stay = stay.iloc[:,:3]
# Remaining the columns
stay = stay.rename(columns={'is_canceled':'Number of stays'})
```

```
# Setting plot size for bar chart
plt.figure(figsize=(20,10))
sns.barplot(x='total_stay', y='Number of stays', hue='hotel',data=stay)
# Set labels
plt.title('Optimal Stay Length in Both Hotel types', fontsize=15)
plt.ylabel('Count of Stay',fontsize=15)
plt.xlabel('Total stay(days)',fontsize=15)
```

1. Why did you pick the specific chart?

For first chart we have picked the line chart here because it helps to show small shifts that may be getting hard to spot in other graphs.It helps show trends for different periods.They are easy to understand so,here we can easily track the change of number of bookings with respect to month.

In second chart the bar plot has been used this chart shown clear view in understanding the relation between total stay in terms of days and count of stays.

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

In the first chart we have found that July and August had the most booking.

In the second chart we found optimal stay in both type hotel is less than 7 days. And after that staying number is declined.

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

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

Yes this provides good insights that hotels should well prepared for the month of July and August as maximum booking take place in this month. So better the preparation and good approach will definately adda to the growth of Hotels.

While second chart also have positive impact,the insight gathered from this chart,hotels can work in the domain to increase the staying length of customer to increase their revenue.The other understanding is that customers usually prefers a one week stay in hotels needs to work efficiently in these seven days so that customers would return to the same hotels again so this will increase the hotel revenue.

Chart - 9 Which year had the highest booking

```
# Chart - 9 visualization code
# set plot size
plt.figure(figsize=(12,5))

#  plot with countplot
sns.countplot(x=hotel_booking_df1['arrival_date_year'],hue=df['hotel'])
plt.title("Year Wise bookings")
```

1. Why did you pick the specific chart?

Answer Here.

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

Answer Here

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

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

Answer Here

Chart - 10 From which country most guest are coming?

```
# Chart - 10 visualization code
guest_country = hotel_booking_df1[hotel_booking_df1['is_canceled'] == 0]['country'].value_counts().reset_index()
guest_country.columns = ['Country', 'No of guests']
guest_country
```

```
basemap = folium.Map()
ax = px.choropleth(guest_country, locations = guest_country['Country'],
                           color = guest_country['No of guests'], hover_name = guest_country['Country'])
ax.show()


# Counting the guests from various countries.
country_df=hotel_booking_df1['country'].value_counts().reset_index().rename(columns={'index': 'country','country': 'count of g

# Visualizing by  plotting the graph
plt.figure(figsize=(20,8))
sns.barplot(x=country_df['country'],y=country_df['count of guests'])
plt.xlabel('Country')
plt.ylabel('Number of guests',fontsize=12)
plt.title("Number of guests from diffrent Countries")
```

1. Why did you pick the specific chart?

This chart shown highest booking made by top 10 country.

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

This chart shown, people from about 166 countries visited the hotels,most of the guests are local i.e. from Portugal. Among overseas visitors, European neighbours like United Kingdom, France, Spain, Germany, Italy had a lion's share.

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

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

Double-click (or enter) to edit

Chart - 11 Which distribution channel had highest booking and cancellation

```
# Chart - 11 visualization code
#Creating labels
labels=hotel_booking_df1['distribution_channel'].value_counts().index.tolist()

# creating new df of distribution channel
distribution_channel_df=hotel_booking_df1['distribution_channel'].value_counts().reset_index().rename(columns={'index':"distri

#adding percentage columns to the distribution_channel_df
distribution_channel_df['percentage']=round(distribution_channel_df['count']*100/df.shape[0],1)

#Creating list of percentage
sizes=distribution_channel_df['percentage'].values.tolist()

#plotting the piw chart
hotel_booking_df['distribution_channel'].value_counts().plot.pie(explode=[0.05]*5, shadow=False, figsize=(15,8),fontsize=10,la

# setting legends with the percentage values
labels = [f'{l}, {s}%' for l, s in zip(labels, sizes)]
plt.legend(bbox_to_anchor=(0.85, 1), loc='upper left', labels=labels)
plt.title(' Mostly Used Distribution Channel for Hotel Bookings ')
```

```
canceled_df=hotel_booking_df1[hotel_booking_df1['is_canceled']==1] # 1= canceled

#group by distribution channel
canceled_df=canceled_df.groupby(['distribution_channel','hotel']).size().reset_index().rename(columns={0:'Counts'})
canceled_df

#set plot size and plot barchart
plt.figure(figsize=(8,5))
sns.barplot(x='distribution_channel',y='Counts',hue="hotel",data=canceled_df)

# set labels
plt.xlabel('Distribution channel')
plt.ylabel('counts')
plt.title('Cancellation Rate Vs Distribution channel')
```

1. Why did you pick the specific chart?

The following chart represent maximum percentage of booking done through which channel and which channel have more cancellation rate for both hotel.

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

As clearly seen TA/TO (Tour Agent/Tour Operator) made highest booking and also TA/TO have highest cancellation rate for both 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.

The most common distribution channel for hotel booking is TA/TO(57.9%) followed by Direct(10.9%) and corporate(4.3%). And also highest cancellation rate for both hotel is TA/TO. The cancellation rate is lower for bookings made directly with the hotel. These insights suggest that hotels may want to focus on strategies to encourage more direct bookings and reduce reliance on TA/TO, which may help to improve the overall cancellation rate and increase customer loyalty.

Chart - 12 Relationship between repeated guests and previous booking not cancelled

```
# Chart - 12 visualization code
repeated_guests_df=hotel_booking_df1[hotel_booking_df1['is_repeated_guest']==1]
repeated_guests_df_1=hotel_booking_df1[hotel_booking_df1['is_repeated_guest']==0]
plt.figure(figsize=(8,5))
sns.barplot(x=hotel_booking_df1['is_repeated_guest'],y= hotel_booking_df1['previous_bookings_not_canceled'])
plt.xticks([0,1],['Not_repeated_guests','repeated_guests'],fontsize=15)
plt.title('Relationship Between repeated guests and previous bookings not cancelled.')
plt.show()
```

Percentage of booking cancellation

```
# booking canceled=1
# booking not canceled= 0

# creating new DataFrame where bookings are cancelled.
canceled_df=hotel_booking_df1[hotel_booking_df1['is_canceled']==1]

# Grouping by hotel
canceled_df=canceled_df.groupby('hotel').size().reset_index().rename(columns={0: "no_of_cancelled_bookings"})

# adding 'total booking column for calculating the percentage.
canceled_df['total_booiknngs']=hotel_booking_df1.groupby('hotel').size().reset_index().rename(columns={0:"total_bookings"}).dro
canceled_df

#plotting the barchat
plt.figure(figsize=(8,5))
sns.barplot(x=canceled_df['hotel'],y=canceled_df['no_of_cancelled_bookings']*100/canceled_df['total_booiknngs'])

#set labels
plt.xlabel('Hotel type')
plt.ylabel('Percentage(%)')
```

```
plt.title("Percentage of booking cancellation")
```

1. Why did you pick the specific chart?

Answer Here.

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

Answer Here

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

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

Answer Here

Chart - 13 Plotting Histogram

```
# Chart - 13 visualization code
hotel_booking_df1.hist(figsize=(24,18))
plt.show()
```

1. Why did you pick the specific chart?

To understanding the data in clear way with proper insights,I have used the histogram here.It is used to summarize discrete or continuous data that are measured on an interval scale.It is often used to illustrate the major features of the distribution of the data in convenient form.It is also useful when dealing with large data sets.It can help detect any unusual observation (outlier) or any gaps in the data.Thus we have used the histogram plot to analysis the variable distribution over the whole dataset whether it's symmetric or not.

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

Some insights found the chart as follows:

- We can see that the maximum guest came in the year 2016.
- Maximum arrival week number is 30.
- Maximum arrival happens in the last of the month.
- Maximum guests comes with no children.
- There is very less requirement of car parking space.

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

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

Histogram can not define business impact. it's just to see the distribution of the column data over the dataset

Chart - 14 - Correlation Heatmap

```
# Correlation Heatmap visualization code
plt.figure(figsize=(18,10))
sns.heatmap(hotel_booking_df1.corr(),annot=True)
plt.title('Co-relation of the columns')
```

1. Why did you pick the specific chart?

1 cell hidden

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

1 cell hidden

Chart - 15 - Pair Plot

```
# Pair Plot visualization code

# Select the columns for the pair plot
columns_for_pairplot = ['lead_time', 'adr', 'total_stay', 'is_repeated_guest', 'previous_cancellations', 'booking_changes', 't

# Create a new DataFrame with only the selected columns
df_selected = hotel_booking_df1[columns_for_pairplot]

# Create the pair plot
sns.pairplot(df_selected)

# Display the plot
plt.show()
```

1. Why did you pick the specific chart?

1 cell hidden

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

1 cell hidden

5. Solution to Business Objective

What do you suggest the client to achieve Business Objective ?

Explain Briefly.

--> Our business objective was to increase bookings, decrease cancellations, and increase customer retention while also extending stays. Based on our investigation, some recommendations we came up with are as follows:

- Additional public marketing can help raise the number of visitors from certain nations. Even after they depart, more effort can be taken to keep them by keeping in touch with them by personalised emails, phone calls, etc.
- Agents and market sectors that bring in more clients should also be recognised with awards and incentives.
- Cancellations had a strong connection with new clients. If new clients are prone to cancel, further efforts should be taken to retain them by providing discounts and offers. Additionally, greater efforts should be made to retain clients as Repeated clients generally cancel less bookings.
- Data from each visitor's stay can be used to send them personalized offers to maximize their chances of booking again.
- Launching customer loyalty programmes to reward loyal customers.
- The period of each visitor's stay can be utilized to create a clever pricing model that should prolong one's stay.
- Extra efforts should be made to foster positive relationships with clients which can be done by sending them engaging emails, such as "Thank you" and "Happy Holidays",etc.
- Good customer evaluations can have a significant impact on a hotel's brand value, and it is important to consider customer feedback and reviews in order to improve hotel amenities and the guest experience.

Conclusion

1. City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
2. 27.5 % bookings were got cancelled out of all the bookings.
3. Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
4. The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.
5. Most of the customers (91.6%) do not require car parking spaces.
6. 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
7. BB(Bed & Breakfast) is the most preferred type of meal by the guests.
8. Maximum number of guests were from Portugal, i.e. more than 25000 guests.
9. Most of the bookings for City hotels and Resort hotel were happened in 2016.
10. Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
11. Booking cancellation rate is high for City hotels which almost 30 %.
12. Average lead time for resort hotel is high.
13. Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
14. Resort hotels have the most repeated guests.
15. Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
16. Almost 19 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.