```
# Import the required Libraries
import numpy as np,pandas as pd, matplotlib.pyplot as plt, seaborn as sns,plotly.express as px
```

```
# Importing the dataset
data = pd.read_csv('/content/hotel_bookings.csv')
data.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_w
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	

5 rows × 32 columns

```
data.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
```

```
data.describe(include='all')
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
count	119390	119390.000000	119390.000000	119390.000000	119390	119390.000000	119390.000000
unique	2	NaN	NaN	NaN	12	NaN	NaN
top	City Hotel	NaN	NaN	NaN	August	NaN	NaN
freq	79330	NaN	NaN	NaN	13877	NaN	NaN
mean	NaN	0.370416	104.011416	2016.156554	NaN	27.165173	15.798241
std	NaN	0.482918	106.863097	0.707476	NaN	13.605138	8.780829
min	NaN	0.000000	0.000000	2015.000000	NaN	1.000000	1.000000
25%	NaN	0.000000	18.000000	2016.000000	NaN	16.000000	8.000000
50%	NaN	0.000000	69.000000	2016.000000	NaN	28.000000	16.000000
75%	NaN	1.000000	160.000000	2017.000000	NaN	38.000000	23.000000
max	NaN	1.000000	737.000000	2017.000000	NaN	53.000000	31.000000

11 rows × 32 columns

Data Cleaning

```
data.isnull().sum()
```

	0
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

```
# Replacing the missing values in 'country' with the most frequented values
data['country'] = data['country'].fillna(data['country'].mode()[0])

# Replacing the null values in 'agent' and 'children' with 0
data['agent'] = data['agent'].fillna(0)
data['children'] = data['children'].fillna(0)

# Droping the 'company' column
data.drop('company', axis=1, inplace=True)
print(data.isnull().sum().sum())
```

```
0
```

```
# Droping the remaining missing values
data.dropna(inplace=True)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
```

```
Data columns (total 31 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                        119390 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         119390 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                           119390 non-null  float64
 24  days_in_waiting_list            119390 non-null  int64
 25  customer_type                   119390 non-null  object
 26  adr                             119390 non-null  float64
 27  required_car_parking_spaces     119390 non-null  int64
 28  total_of_special_requests       119390 non-null  int64
 29  reservation_status              119390 non-null  object
 30  reservation_status_date         119390 non-null  object
dtypes: float64(3), int64(16), object(12)
memory usage: 28.2+ MB
```

```
# Check for duplicates
data.duplicated().sum()
```

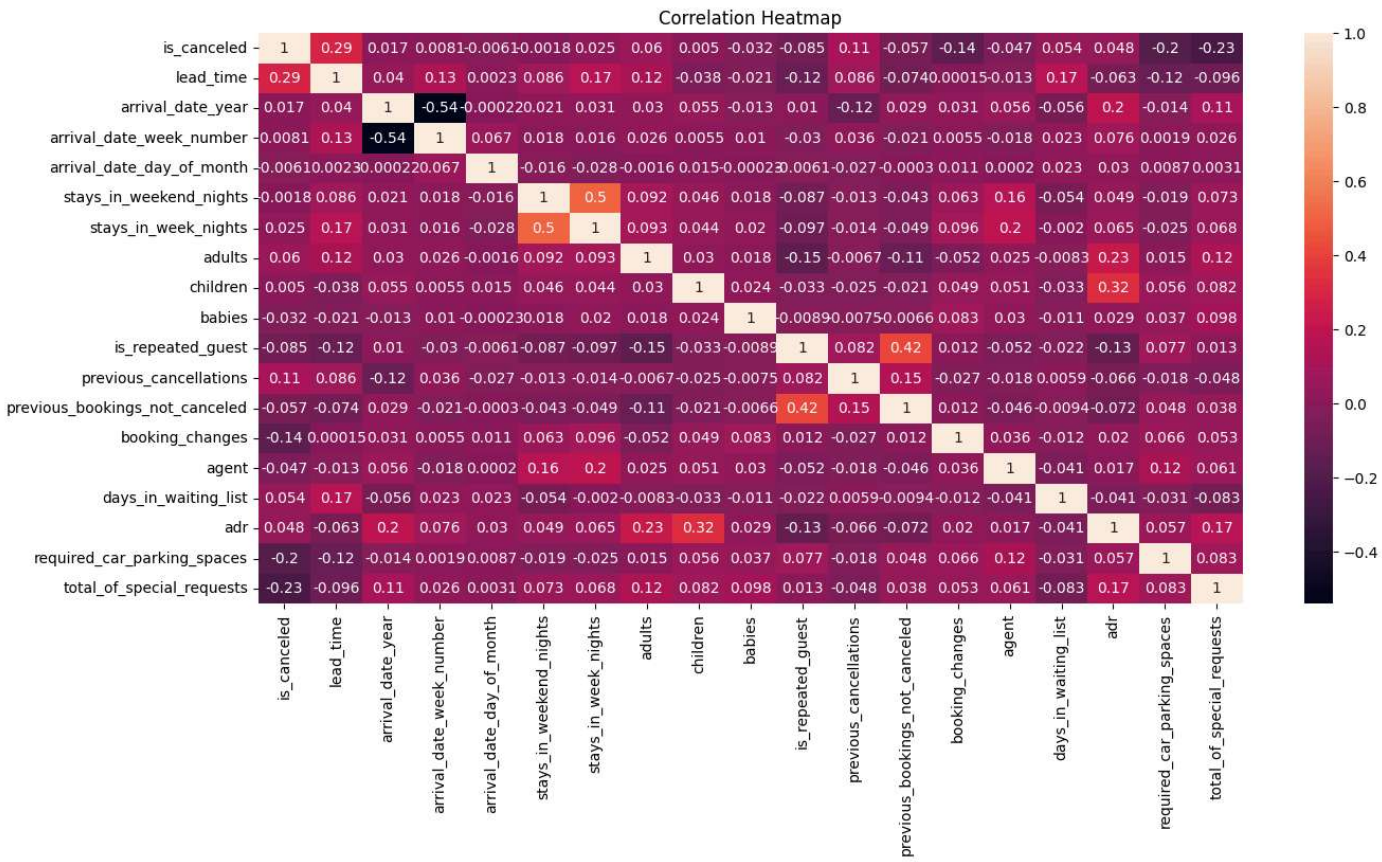
```
# Converting the data type of reservation_status_date to datetime
data['reservation_status_date'] = pd.to_datetime(data['reservation_status_date'])
print(data['reservation_status_date'].dtype)
```

```
# Removing the rows where there are no guests.
data = data[~((data['adults'] == 0) & (data['children'] == 0) & (data['babies'] == 0))]
data.shape
```

```
# Value counts for each columns
for col in data.columns:
  print(data[col].value_counts())
  print('*'*70)
```

EDA

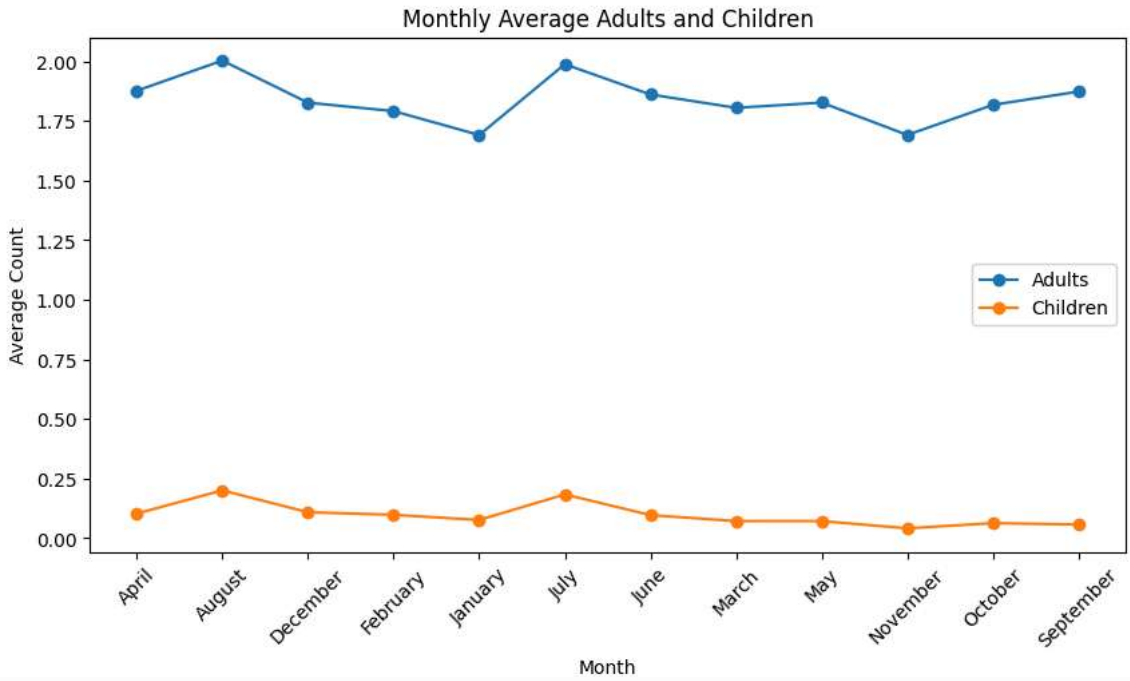
```
# Correalation between different numerical variables
data_num = data.select_dtypes(include=['int64', 'float64'])
Correlation_matrix = data_num.corr()
plt.figure(figsize=(15, 7))
sns.heatmap(Correlation_matrix, annot=True)
plt.title('Correlation Heatmap')
plt.show()
```



- **Total Length of Stay**:
  - The analysis indicates a positive correlation between the `stays_in_weekend_nights` and `stays_in_week_nights` columns.
  - By merging these two features, we can create a new variable that represents the total length of stay, facilitating better insights into booking patterns and enhancing our ability to analyze guest behaviors.
- **Repeat Guest Behavior**:
  - There is a positive correlation observed between `previous_bookings_not_cancelled` and `is_repeated_guest`.
  - This correlation implies that guests who have successfully completed prior bookings are more likely to return as repeat guests, suggesting potential areas for focusing our guest retention strategies.

Booking Trends Patterns

```
# Arrival Month pattern with Number of Adults and Chlidren
monthly_data = data.groupby('arrival_date_month')
plt.figure(figsize=(10, 5))
plt.plot(monthly_data['adults'].mean(), label='Adults',marker='o')
plt.plot(monthly_data['children'].mean(), label='Children',marker='o')
plt.title('Monthly Average Adults and Children')
plt.xlabel('Month')
plt.ylabel('Average Count')
plt.xticks(rotation=45)
plt.legend()
plt.show()
```




```
# Monthly trend on booking
data['reservation_status_date'] = pd.to_datetime(data['reservation_status_date'])
data['month'] = data['reservation_status_date'].dt.month_name()
data['month'] = data['month'].str.capitalize()
monthly_bookings = data.groupby('month').size()
plt.figure(figsize=(10, 5))
monthly_bookings.plot(kind='bar', color='skyblue')
plt.title('Monthly Bookings')
plt.xlabel('Month')
plt.ylabel('Number of Bookings')
plt.xticks(rotation=45)
plt.show()
# Yearly trend on booking
yearly_bookings = data.groupby(data['reservation_status_date'].dt.year).size()
plt.figure(figsize=(10, 5))
yearly_bookings.plot(kind='bar', color='lightcoral')
plt.title('Yearly Bookings')
plt.xlabel('Year')
plt.ylabel('Number of Bookings')
plt.show()
```


```
# Booking trends based on  Countries

guests_by_country = data[data['is_canceled'] == 0]['country'].value_counts().reset_index()
guests_by_country.columns = ['Country', 'Number of guests']
guests_by_country
```

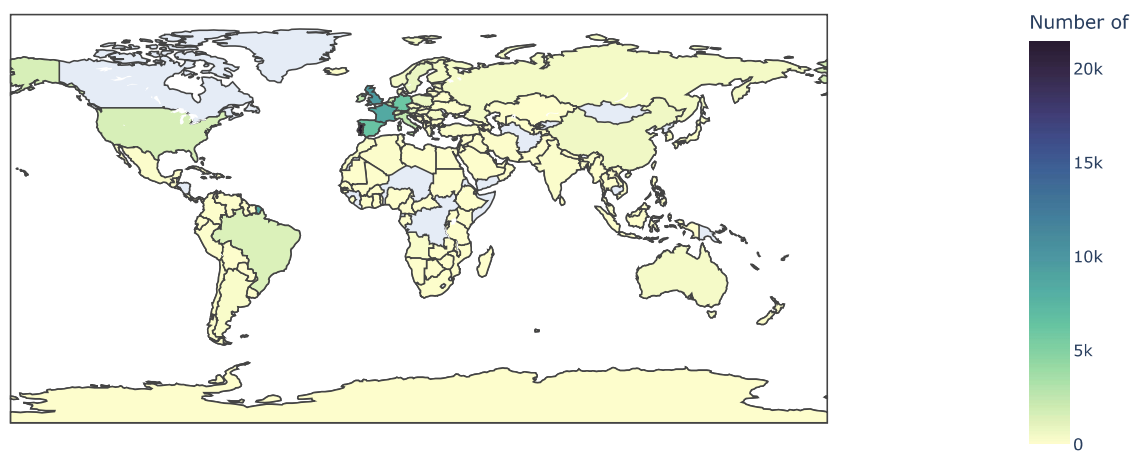
	Country	Number of guests
0	PRT	21492
1	GBR	9676
2	FRA	8481
3	ESP	6391
4	DEU	6069
...	...	...
160	KIR	1
161	ATF	1
162	TJK	1
163	SLE	1
164	FRO	1

165 rows × 2 columns

Next steps: Generate code with guests_by_country | View recommended plots | New interactive sheet

```
guests_map = px.choropleth(
    guests_by_country,
    locations = guests_by_country ['Country'],
    color = guests_by_country ['Number of guests'],
    hover_name = guests_by_country ['Country'],
    title='Guest Distribution by Country',
    color_continuous_scale=px.colors.sequential.deep
)

guests_map.show()
```



- It is observed from the guests to country that the most guests are coming from portugal followed by other europe countries.
- From the yearly trend of booking it is observed that the number of booking is more in the year 2016.
- From the monthly trend of booking it is observed that the number of booking is more in the month July.

Checking Cancellation patttern

```
# Cancelling based on hotel type
plt.figure()
sns.countplot(x='hotel', hue='is_canceled', data=data)
```

```
plt.title('Cancellation based on Hotel Type')
plt.xlabel('Hotel Type')
plt.ylabel('Count')
plt.show()

# Cancellation pattern over months
plt.figure(figsize=(13, 5))
sns.countplot(hue='is_canceled', x='arrival_date_month', data=data
plt.title('Cancellation Status over Arrival Months')
plt.xlabel('Arrival Month')
plt.ylabel('Count')
plt.show()
```