HOTEL BOOKING ANALYSIS : EDA

In [118…
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [119…
```
hotel = pd.read_csv("hotel_bookings 2.csv")
hotel.head()
```

Out[119…

	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	count
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	0.0	0	BB	P
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	0.0	0	BB	P
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	0.0	0	BB	G
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	0.0	0	BB	G
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	0.0	0	BB	G

In [120…
```
df = hotel.copy()
```

In [121…
```
df.head(2)
```

Out[121…

	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	count
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	0.0	0	BB	P
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	0.0	0	BB	P

In [122…
```
df.columns
```

Out[122…

```
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')
```

In [123…

```
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
```

In [ ]:

```
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'], dayfirst=True)
```

In [125…

```
# Show summary of all Objects using describe
```

```
object_describe = df.describe(include = 'object')
object_describe
```

Out[125…

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166

In [126…

```
# Show unique values of each Object

for x in object_describe:
    display(x)
    print(df[x].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']
--------------------------------------------------------------------------------------------------------
```

In [127…
```
df.describe()
```

Out[127…

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repea
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119386.000000	119390.000000	119
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	1.856403	0.103890	0.007949	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.000000	10.000000	10.000000	
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.579261	0.398561	0.097436	

In [128…
```
# Show only those columns which have null values

df.isna().sum()[df.isna().sum() > 0]
```

Out[128…
```
children         4
country        488
agent        16340
company     112593
dtype: int64
```

In [129…
```
mode_children = df['children'].mode()[0]
```

In [130…
```
mode_country = df['country'].mode()[0]
```

In [131…
```
df['children'].fillna(mode_children, inplace=True)
df['country'].fillna(mode_country, inplace=True)
df.drop(['company', 'agent'], axis=1, inplace=True)
```

In [313…
```
df.isna().sum()[df.isna().sum() > 0]
```

Out[313…
```
Series([], dtype: int64)
```

In [133…
```
df.describe()
```

Out[133...

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repea
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	1.856403	0.103886	0.007949	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.000000	10.000000	10.000000	
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.579261	0.398555	0.097436	

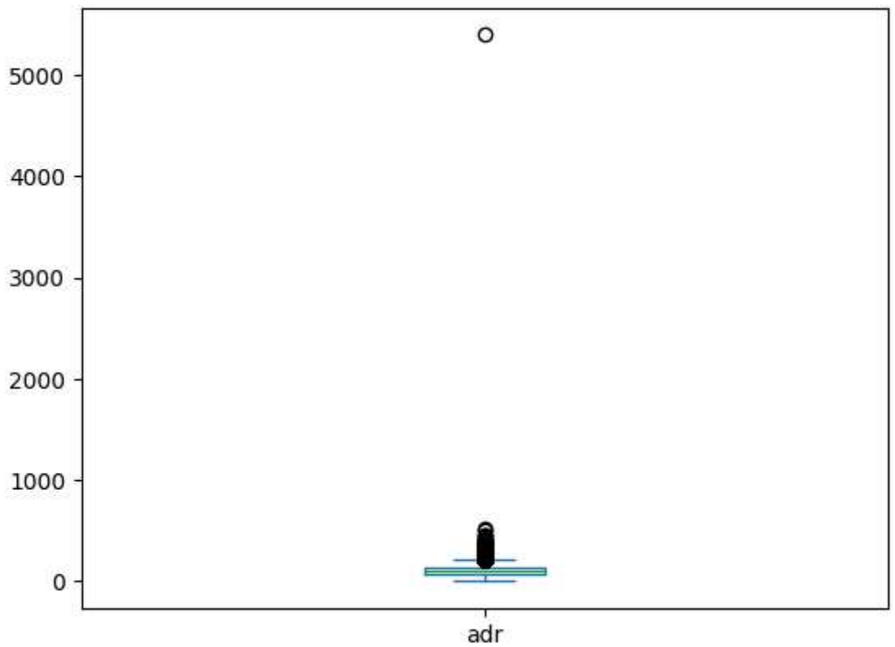
In [ ]:
```
# average daily rate

df['adr'].describe()
```

Out[ ]:
```
count    119390.000000
mean        101.831122
std          50.535790
min          -6.380000
25%          69.290000
50%          94.575000
75%         126.000000
max        5400.000000
Name: adr, dtype: float64
```

In [ ]:
```
df['adr'].plot(kind='box')
# This shows value above 5000 are outliers
```

Out[ ]:
```
<Axes: >
```



In [141…
```
df['adr'].nlargest(5).reset_index()
```

Out[141…

	index	adr
0	48515	5400.0
1	111403	510.0
2	15083	508.0
3	103912	451.5
4	13142	450.0

In [156…
```
outlier_adr = df[df['adr'] > 5000].index
outlier_adr
```

Out[156… `Index([48515], dtype='int64')`

In [157…
```
df.drop(index=outlier_adr, inplace=True)
```

In [158…
```
df['adr'].nlargest(5).reset_index()
```

Out[158…

	index	adr
0	111403	510.0
1	15083	508.0
2	103912	451.5
3	13142	450.0
4	13391	437.0