

descriptive statistics

November 7, 2021

```
[235]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
# set seaborn theme
sns.set_style(style="whitegrid")
from sklearn.linear_model import LinearRegression
```

```
[236]: from IPython.core.interactiveshell import InteractiveShell

InteractiveShell.ast_node_interactivity = "all"
```

```
[237]: resortHotal = pd.read_csv("H1.csv")
resortHotal.head()
cityHotal = pd.read_csv("H2.csv")
cityHotal.head()
```

```
[237]:
```

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
0	0	342	2015	July	
1	0	737	2015	July	
2	0	7	2015	July	
3	0	13	2015	July	
4	0	14	2015	July	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	0	
4	27	1	0	

	StaysInWeekNights	Adults	Children	...	DepositType	Agent	\
0	0	2	0	...	No Deposit	NULL	
1	0	2	0	...	No Deposit	NULL	

2	1	1	0 ... No Deposit	NULL
3	1	1	0 ... No Deposit	304
4	2	2	0 ... No Deposit	240

	Company	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	NULL		0 Transient	0.0		0
1	NULL		0 Transient	0.0		0
2	NULL		0 Transient	75.0		0
3	NULL		0 Transient	75.0		0
4	NULL		0 Transient	98.0		0

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate
0	0	Check-Out	2015-07-01
1	0	Check-Out	2015-07-01
2	0	Check-Out	2015-07-02
3	0	Check-Out	2015-07-02
4	1	Check-Out	2015-07-03

[5 rows x 31 columns]

[237]:

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
0	0	6	2015	July	
1	1	88	2015	July	
2	1	65	2015	July	
3	1	92	2015	July	
4	1	100	2015	July	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	2	
4	27	2	0	

	StaysInWeekNights	Adults	Children	...	DepositType	Agent	\
0	2	1	0.0	...	No Deposit	6	
1	4	2	0.0	...	No Deposit	9	
2	4	1	0.0	...	No Deposit	9	
3	4	2	0.0	...	No Deposit	9	
4	2	2	0.0	...	No Deposit	9	

	Company	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	NULL		0 Transient	0.0		0
1	NULL		0 Transient	76.5		0
2	NULL		0 Transient	68.0		0
3	NULL		0 Transient	76.5		0
4	NULL		0 Transient	76.5		0

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate
0	0	Check-Out	2015-07-03
1	1	Canceled	2015-07-01
2	1	Canceled	2015-04-30
3	2	Canceled	2015-06-23
4	1	Canceled	2015-04-02

[5 rows x 31 columns]

- 1 add one attribute in each data set such as “Hotel”- It contains the information about the type of hotel after that we combine both datasets (i.e., city hotel or resort hotel)

```
[238]: resortHotal['Hotal'] = 'resort hotal'
resortHotal.head()
cityHotal['Hotal'] = 'city hotal'
cityHotal.head()
```

```
[238]: IsCanceled  LeadTime  ArrivalDateYear  ArrivalDateMonth  \
0           0        342           2015           July
1           0        737           2015           July
2           0          7           2015           July
3           0         13           2015           July
4           0         14           2015           July
```

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	0	
4	27	1	0	

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
0	0	2	0	...	NULL	NULL	
1	0	2	0	...	NULL	NULL	
2	1	1	0	...	NULL	NULL	
3	1	1	0	...	304	NULL	
4	2	2	0	...	240	NULL	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	0	Transient	0.0	0	
1	0	Transient	0.0	0	
2	0	Transient	75.0	0	

3	0	Transient	75.0	0
4	0	Transient	98.0	0

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
0	0	Check-Out	2015-07-01	
1	0	Check-Out	2015-07-01	
2	0	Check-Out	2015-07-02	
3	0	Check-Out	2015-07-02	
4	1	Check-Out	2015-07-03	

	Hotal
0	resort hotal
1	resort hotal
2	resort hotal
3	resort hotal
4	resort hotal

[5 rows x 32 columns]

[238]:

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
0	0	6	2015	July	
1	1	88	2015	July	
2	1	65	2015	July	
3	1	92	2015	July	
4	1	100	2015	July	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	2	
4	27	2	0	

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
0	2	1	0.0	...	6	NULL	
1	4	2	0.0	...	9	NULL	
2	4	1	0.0	...	9	NULL	
3	4	2	0.0	...	9	NULL	
4	2	2	0.0	...	9	NULL	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	0	Transient	0.0	0	
1	0	Transient	76.5	0	
2	0	Transient	68.0	0	
3	0	Transient	76.5	0	
4	0	Transient	76.5	0	

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	Hotal
0	0	Check-Out	2015-07-03	city hotal
1	1	Canceled	2015-07-01	city hotal
2	1	Canceled	2015-04-30	city hotal
3	2	Canceled	2015-06-23	city hotal
4	1	Canceled	2015-04-02	city hotal

[5 rows x 32 columns]

2 concating the both of dataframe

```
[239]: HotalData = pd.concat([resortHotal,cityHotal],ignore_index=True)
HotalData.head()
HotalData.tail()
HotalData['Hotal'].value_counts()
```

```
[239]:   IsCanceled  LeadTime  ArrivalDateYear  ArrivalDateMonth  \
0          0      342          2015          July
1          0      737          2015          July
2          0       7          2015          July
3          0      13          2015          July
4          0      14          2015          July
```

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	0	
4	27	1	0	

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
0	0	2	0.0	...	NULL	NULL	
1	0	2	0.0	...	NULL	NULL	
2	1	1	0.0	...	NULL	NULL	
3	1	1	0.0	...	304	NULL	
4	2	2	0.0	...	240	NULL	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	0	Transient	0.0	0	
1	0	Transient	0.0	0	
2	0	Transient	75.0	0	
3	0	Transient	75.0	0	
4	0	Transient	98.0	0	

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
--	------------------------	-------------------	-----------------------	---

0	0	Check-Out	2015-07-01
1	0	Check-Out	2015-07-01
2	0	Check-Out	2015-07-02
3	0	Check-Out	2015-07-02
4	1	Check-Out	2015-07-03

Hotal

0	resort hotal
1	resort hotal
2	resort hotal
3	resort hotal
4	resort hotal

[5 rows x 32 columns]

[239] :

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
119385	0	23	2017	August	
119386	0	102	2017	August	
119387	0	34	2017	August	
119388	0	109	2017	August	
119389	0	205	2017	August	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
119385	35	30	2	
119386	35	31	2	
119387	35	31	2	
119388	35	31	2	
119389	35	29	2	

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
119385	5	2	0.0	...	394	NULL	
119386	5	3	0.0	...	9	NULL	
119387	5	2	0.0	...	9	NULL	
119388	5	2	0.0	...	89	NULL	
119389	7	2	0.0	...	9	NULL	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
119385	0	Transient	96.14	0	
119386	0	Transient	225.43	0	
119387	0	Transient	157.71	0	
119388	0	Transient	104.40	0	
119389	0	Transient	151.20	0	

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
119385	0	Check-Out	2017-09-06	
119386	2	Check-Out	2017-09-07	
119387	4	Check-Out	2017-09-07	

119388	0	Check-Out	2017-09-07
119389	2	Check-Out	2017-09-07

```

            Hotal
119385  city hotal
119386  city hotal
119387  city hotal
119388  city hotal
119389  city hotal

```

[5 rows x 32 columns]

```

[239]: city hotal      79330
       resort hotal   40060
       Name: Hotal, dtype: int64

```

3 Displaying Dataframe Structure

```
[240]: HotalData.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   IsCanceled                            119390 non-null  int64
1   LeadTime                              119390 non-null  int64
2   ArrivalDateYear                       119390 non-null  int64
3   ArrivalDateMonth                      119390 non-null  object
4   ArrivalDateWeekNumber                 119390 non-null  int64
5   ArrivalDateDayOfMonth                 119390 non-null  int64
6   StaysInWeekendNights                  119390 non-null  int64
7   StaysInWeekNights                     119390 non-null  int64
8   Adults                                119390 non-null  int64
9   Children                              119386 non-null  float64
10  Babies                                119390 non-null  int64
11  Meal                                  119390 non-null  object
12  Country                               118902 non-null  object
13  MarketSegment                         119390 non-null  object
14  DistributionChannel                   119390 non-null  object
15  IsRepeatedGuest                       119390 non-null  int64
16  PreviousCancellations                 119390 non-null  int64
17  PreviousBookingsNotCanceled           119390 non-null  int64
18  ReservedRoomType                      119390 non-null  object
19  AssignedRoomType                      119390 non-null  object

```

```

20 BookingChanges          119390 non-null int64
21 DepositType             119390 non-null object
22 Agent                   119390 non-null object
23 Company                  119390 non-null object
24 DaysInWaitingList       119390 non-null int64
25 CustomerType            119390 non-null object
26 ADR                     119390 non-null float64
27 RequiredCarParkingSpaces 119390 non-null int64
28 TotalOfSpecialRequests   119390 non-null int64
29 ReservationStatus        119390 non-null object
30 ReservationStatusDate    119390 non-null object
31 Hotal                    119390 non-null object
dtypes: float64(2), int64(16), object(14)
memory usage: 29.1+ MB

```

4 Assigning the correct data type of the Categorical attribute in the dataframe

```

[241]: HotalData[ "IsCanceled"] = HotalData[ "IsCanceled"].astype("category")
HotalData[ "ArrivalDateMonth"] = HotalData[ "ArrivalDateMonth"].
↳astype("category")
HotalData[ "IsRepeatedGuest"] = HotalData[ "IsRepeatedGuest"].astype("category")

```

```

[242]: HotalData.info()

```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   IsCanceled                            119390 non-null category
1   LeadTime                             119390 non-null int64
2   ArrivalDateYear                       119390 non-null int64
3   ArrivalDateMonth                      119390 non-null category
4   ArrivalDateWeekNumber                 119390 non-null int64
5   ArrivalDateDayOfMonth                 119390 non-null int64
6   StaysInWeekendNights                  119390 non-null int64
7   StaysInWeekNights                     119390 non-null int64
8   Adults                                119390 non-null int64
9   Children                              119386 non-null float64
10  Babies                                119390 non-null int64
11  Meal                                  119390 non-null object
12  Country                               118902 non-null object
13  MarketSegment                         119390 non-null object
14  DistributionChannel                   119390 non-null object
15  IsRepeatedGuest                       119390 non-null category

```



```

16 PreviousCancellations      119390 non-null int64
17 PreviousBookingsNotCanceled 119390 non-null int64
18 ReservedRoomType          119390 non-null object
19 AssignedRoomType           119390 non-null object
20 BookingChanges              119390 non-null int64
21 DepositType                 119390 non-null object
22 Agent                       119390 non-null object
23 Company                     119390 non-null object
24 DaysInWaitingList           119390 non-null int64
25 CustomerType                119390 non-null object
26 ADR                         119390 non-null float64
27 RequiredCarParkingSpaces    119390 non-null int64
28 TotalOfSpecialRequests      119390 non-null int64
29 ReservationStatus           119390 non-null object
30 ReservationStatusDate       119390 non-null object
31 Hotal                       119390 non-null object
dtypes: category(3), float64(2), int64(14), object(13)
memory usage: 26.8+ MB

```

5 Summary of Categorical Attribute:

```

[243]: HotalData.IsCanceled.unique()
HotalData.groupby('IsCanceled') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# ArrivalDateMonth
HotalData.groupby('ArrivalDateMonth') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# Meal
HotalData.groupby('Meal') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# Country
len(HotalData.Country.unique())
HotalData.groupby('Country') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

```

```

# MarketSegment
len(HotalData.MarketSegment.unique())
HotalData.groupby('MarketSegment') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# DistributionChannel
len(HotalData.DistributionChannel.unique())
HotalData.groupby('DistributionChannel') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# IsRepeatedGuest
len(HotalData.IsRepeatedGuest.unique())
HotalData.groupby('IsRepeatedGuest') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# ReservedRoomType
len(HotalData.ReservedRoomType.unique())
HotalData.groupby('ReservedRoomType') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# AssignedRoomType
len(HotalData.AssignedRoomType.unique())
HotalData.groupby('AssignedRoomType') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# DepositType
len(HotalData.DepositType.unique())
HotalData.groupby('DepositType') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

# Agent
len(HotalData.Agent.unique())
HotalData.groupby('Agent') \
    .size() \
    .reset_index() \

```

```

        .rename(columns={0: 'Counts'}) \
        .sort_values(by='Counts', ascending=False)
# Company
len(HotalData.Company.unique())
HotalData.groupby('Company') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# CustomerType
len(HotalData.CustomerType.unique())
HotalData.groupby('CustomerType') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# CustomerType
len(HotalData.CustomerType.unique())
HotalData.groupby('CustomerType') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# ReservationStatus
len(HotalData.ReservationStatus.unique())
HotalData.groupby('ReservationStatus') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)
# Hotal
len(HotalData.Hotal.unique())
HotalData.groupby('Hotal') \
    .size() \
    .reset_index() \
    .rename(columns={0: 'Counts'}) \
    .sort_values(by='Counts', ascending=False)

```

```

[243]: [0, 1]
Categories (2, int64): [0, 1]

```

```

[243]:   IsCanceled  Counts
0          0    75166
1          1   44224

```

```

[243]:   ArrivalDateMonth  Counts
1          August    13877

```

5	July	12661
8	May	11791
10	October	11160
0	April	11089
6	June	10939
11	September	10508
7	March	9794
3	February	8068
9	November	6794
2	December	6780
4	January	5929

```
[243]:      Meal  Counts
0  BB      92310
2  HB      14463
3  SC      10650
4  Undefined  1169
1  FB       798
```

```
[243]: 178
```

```
[243]:      Country  Counts
135    PRT    48590
59     GBR    12129
56     FRA    10415
51     ESP     8568
43     DEU     7287
..     ...     ...
109    MLI       1
111    MMR       1
114    MRT       1
148    SMR       1
137    PYF       1
```

```
[177 rows x 2 columns]
```

```
[243]: 8
```

```
[243]:      MarketSegment  Counts
6      Online TA    56477
5  Offline TA/TO    24219
4      Groups      19811
3      Direct      12606
2      Corporate     5295
1  Complementary     743
0      Aviation     237
7      Undefined       2
```

[243]: 5

```
[243]: DistributionChannel Counts
      3          TA/TO    97870
      1          Direct   14645
      0        Corporate    6677
      2           GDS      193
      4        Undefined      5
```

[243]: 2

```
[243]: IsRepeatedGuest Counts
      0          0  115580
      1          1   3810
```

[243]: 10

```
[243]: ReservedRoomType Counts
      0 A          85994
      3 D          19201
      4 E          6535
      5 F          2897
      6 G          2094
      1 B          1118
      2 C           932
      7 H           601
      9 P            12
      8 L             6
```

[243]: 12

```
[243]: AssignedRoomType Counts
      0 A          74053
      3 D          25322
      4 E          7806
      5 F          3751
      6 G          2553
      2 C          2375
      1 B          2163
      7 H           712
      8 I           363
      9 K           279
     11 P            12
     10 L             1
```

[243]: 3

```
[243]:      DepositType  Counts
      0  No Deposit    104641
      1  Non Refund     14587
      2  Refundable      162
```

```
[243]: 334
```

```
[243]:      Agent  Counts
      8      9  31961
    333    NULL  16340
    173    240  13922
      0      1   7191
     13     14   3640
     ..     ...    ...
    200    280      1
    298    444      1
    299    446      1
    199    278      1
    302    451      1
```

```
[334 rows x 2 columns]
```

```
[243]: 353
```

```
[243]:      Company  Counts
    352    NULL  112593
     20      40   927
    131     223   784
     37      67   267
     23      45   250
     ..     ...    ...
    172     284      1
    175     288      1
    281     425      1
      1        8      1
      0        6      1
```

```
[353 rows x 2 columns]
```

```
[243]: 4
```

```
[243]:      CustomerType  Counts
      2      Transient  89613
      3  Transient-Party  25124
      0      Contract   4076
      1      Group      577
```

[243]: 4

```
[243]:      CustomerType  Counts
      2      Transient  89613
      3  Transient-Party  25124
      0      Contract   4076
      1          Group    577
```

[243]: 3

```
[243]:  ReservationStatus  Counts
      1      Check-Out   75166
      0      Canceled   43017
      2      No-Show    1207
```

[243]: 2

```
[243]:      Hotal  Counts
      0  city hotal   79330
      1  resort hotal  40060
```

6 Covert the data type object to date datetime format

```
[244]: HotalData['ReservationStatusDate'] = pd.
      ↪to_datetime(HotalData['ReservationStatusDate'])
      HotalData['ReservationStatusDate'].min()
      HotalData['ReservationStatusDate'].idxmin()
      HotalData['ReservationStatusDate'].max()
      #HotalData.loc[73714,]
      HotalData['ReservationStatusDate'].idxmax()
      HotalData.loc[[73714,40058],]
```

[244]: Timestamp('2014-10-17 00:00:00')

[244]: 73714

[244]: Timestamp('2017-09-14 00:00:00')

[244]: 40058

```
[244]:      IsCanceled  LeadTime  ArrivalDateYear  ArrivalDateMonth  \
      73714          1       265             2015              July
      40058          0       211             2017              August

      ArrivalDateWeekNumber  ArrivalDateDayOfMonth  StaysInWeekendNights  \
```

73714	28	9	0
40058	35	31	4

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
73714	2	2	0.0	...	1	NULL	
40058	10	2	0.0	...	40	NULL	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
73714	0	Transient-Party	62.8		0
40058	0	Contract	112.8		0

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
73714	0	Canceled	2014-10-17	
40058	1	Check-Out	2017-09-14	

	Hotal
73714	city hotal
40058	resort hotal

[2 rows x 32 columns]

Summary of numerical Attribute:

```
[245]: round(HotalData.describe(),0)
HotalData.describe()
```

```
[245]:
```

	LeadTime	ArrivalDateYear	ArrivalDateWeekNumber	\
count	119390.0	119390.0	119390.0	
mean	104.0	2016.0	27.0	
std	107.0	1.0	14.0	
min	0.0	2015.0	1.0	
25%	18.0	2016.0	16.0	
50%	69.0	2016.0	28.0	
75%	160.0	2017.0	38.0	
max	737.0	2017.0	53.0	

	ArrivalDateDayOfMonth	StaysInWeekendNights	StaysInWeekNights	\
count	119390.0	119390.0	119390.0	
mean	16.0	1.0	3.0	
std	9.0	1.0	2.0	
min	1.0	0.0	0.0	
25%	8.0	0.0	1.0	
50%	16.0	1.0	2.0	
75%	23.0	2.0	3.0	
max	31.0	19.0	50.0	

Adults	Children	Babies	PreviousCancellations	\
--------	----------	--------	-----------------------	---

count	119390.0	119386.0	119390.0	119390.0
mean	2.0	0.0	0.0	0.0
std	1.0	0.0	0.0	1.0
min	0.0	0.0	0.0	0.0
25%	2.0	0.0	0.0	0.0
50%	2.0	0.0	0.0	0.0
75%	2.0	0.0	0.0	0.0
max	55.0	10.0	10.0	26.0

	PreviousBookingsNotCanceled	BookingChanges	DaysInWaitingList \
count	119390.0	119390.0	119390.0
mean	0.0	0.0	2.0
std	1.0	1.0	18.0
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	0.0	0.0	0.0
max	72.0	21.0	391.0

	ADR	RequiredCarParkingSpaces	TotalOfSpecialRequests
count	119390.0	119390.0	119390.0
mean	102.0	0.0	1.0
std	51.0	0.0	1.0
min	-6.0	0.0	0.0
25%	69.0	0.0	0.0
50%	95.0	0.0	0.0
75%	126.0	0.0	1.0
max	5400.0	8.0	5.0

[245]:

	LeadTime	ArrivalDateYear	ArrivalDateWeekNumber \
count	119390.000000	119390.000000	119390.000000
mean	104.011416	2016.156554	27.165173
std	106.863097	0.707476	13.605138
min	0.000000	2015.000000	1.000000
25%	18.000000	2016.000000	16.000000
50%	69.000000	2016.000000	28.000000
75%	160.000000	2017.000000	38.000000
max	737.000000	2017.000000	53.000000

	ArrivalDateDayOfMonth	StaysInWeekendNights	StaysInWeekNights \
count	119390.000000	119390.000000	119390.000000
mean	15.798241	0.927599	2.500302
std	8.780829	0.998613	1.908286
min	1.000000	0.000000	0.000000
25%	8.000000	0.000000	1.000000
50%	16.000000	1.000000	2.000000
75%	23.000000	2.000000	3.000000

max	31.000000	19.000000	50.000000
-----	-----------	-----------	-----------

	Adults	Children	Babies	PreviousCancellations \
count	119390.000000	119386.000000	119390.000000	119390.000000
mean	1.856403	0.103890	0.007949	0.087118
std	0.579261	0.398561	0.097436	0.844336
min	0.000000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	0.000000
50%	2.000000	0.000000	0.000000	0.000000
75%	2.000000	0.000000	0.000000	0.000000
max	55.000000	10.000000	10.000000	26.000000

	PreviousBookingsNotCanceled	BookingChanges	DaysInWaitingList \
count	119390.000000	119390.000000	119390.000000
mean	0.137097	0.221124	2.321149
std	1.497437	0.652306	17.594721
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000
max	72.000000	21.000000	391.000000

	ADR	RequiredCarParkingSpaces	TotalOfSpecialRequests
count	119390.000000	119390.000000	119390.000000
mean	101.831122	0.062518	0.571363
std	50.535790	0.245291	0.792798
min	-6.380000	0.000000	0.000000
25%	69.290000	0.000000	0.000000
50%	94.575000	0.000000	0.000000
75%	126.000000	0.000000	1.000000
max	5400.000000	8.000000	5.000000

```
[246]: round(HotalData.describe().T,2)
HotalData.describe().T
```

```
[246]:
```

	count	mean	std	min	25% \
LeadTime	119390.0	104.01	106.86	0.00	18.00
ArrivalDateYear	119390.0	2016.16	0.71	2015.00	2016.00
ArrivalDateWeekNumber	119390.0	27.17	13.61	1.00	16.00
ArrivalDateDayOfMonth	119390.0	15.80	8.78	1.00	8.00
StaysInWeekendNights	119390.0	0.93	1.00	0.00	0.00
StaysInWeekNights	119390.0	2.50	1.91	0.00	1.00
Adults	119390.0	1.86	0.58	0.00	2.00
Children	119386.0	0.10	0.40	0.00	0.00
Babies	119390.0	0.01	0.10	0.00	0.00
PreviousCancellations	119390.0	0.09	0.84	0.00	0.00
PreviousBookingsNotCanceled	119390.0	0.14	1.50	0.00	0.00

BookingChanges	119390.0	0.22	0.65	0.00	0.00
DaysInWaitingList	119390.0	2.32	17.59	0.00	0.00
ADR	119390.0	101.83	50.54	-6.38	69.29
RequiredCarParkingSpaces	119390.0	0.06	0.25	0.00	0.00
TotalOfSpecialRequests	119390.0	0.57	0.79	0.00	0.00

	50%	75%	max
LeadTime	69.00	160.0	737.0
ArrivalDateYear	2016.00	2017.0	2017.0
ArrivalDateWeekNumber	28.00	38.0	53.0
ArrivalDateDayOfMonth	16.00	23.0	31.0
StaysInWeekendNights	1.00	2.0	19.0
StaysInWeekNights	2.00	3.0	50.0
Adults	2.00	2.0	55.0
Children	0.00	0.0	10.0
Babies	0.00	0.0	10.0
PreviousCancellations	0.00	0.0	26.0
PreviousBookingsNotCanceled	0.00	0.0	72.0
BookingChanges	0.00	0.0	21.0
DaysInWaitingList	0.00	0.0	391.0
ADR	94.58	126.0	5400.0
RequiredCarParkingSpaces	0.00	0.0	8.0
TotalOfSpecialRequests	0.00	1.0	5.0

[246] :

	count	mean	std	min	\
LeadTime	119390.0	104.011416	106.863097	0.00	
ArrivalDateYear	119390.0	2016.156554	0.707476	2015.00	
ArrivalDateWeekNumber	119390.0	27.165173	13.605138	1.00	
ArrivalDateDayOfMonth	119390.0	15.798241	8.780829	1.00	
StaysInWeekendNights	119390.0	0.927599	0.998613	0.00	
StaysInWeekNights	119390.0	2.500302	1.908286	0.00	
Adults	119390.0	1.856403	0.579261	0.00	
Children	119386.0	0.103890	0.398561	0.00	
Babies	119390.0	0.007949	0.097436	0.00	
PreviousCancellations	119390.0	0.087118	0.844336	0.00	
PreviousBookingsNotCanceled	119390.0	0.137097	1.497437	0.00	
BookingChanges	119390.0	0.221124	0.652306	0.00	
DaysInWaitingList	119390.0	2.321149	17.594721	0.00	
ADR	119390.0	101.831122	50.535790	-6.38	
RequiredCarParkingSpaces	119390.0	0.062518	0.245291	0.00	
TotalOfSpecialRequests	119390.0	0.571363	0.792798	0.00	

	25%	50%	75%	max
LeadTime	18.00	69.000	160.0	737.0
ArrivalDateYear	2016.00	2016.000	2017.0	2017.0
ArrivalDateWeekNumber	16.00	28.000	38.0	53.0
ArrivalDateDayOfMonth	8.00	16.000	23.0	31.0

StaysInWeekendNights	0.00	1.000	2.0	19.0
StaysInWeekNights	1.00	2.000	3.0	50.0
Adults	2.00	2.000	2.0	55.0
Children	0.00	0.000	0.0	10.0
Babies	0.00	0.000	0.0	10.0
PreviousCancellations	0.00	0.000	0.0	26.0
PreviousBookingsNotCanceled	0.00	0.000	0.0	72.0
BookingChanges	0.00	0.000	0.0	21.0
DaysInWaitingList	0.00	0.000	0.0	391.0
ADR	69.29	94.575	126.0	5400.0
RequiredCarParkingSpaces	0.00	0.000	0.0	8.0
TotalOfSpecialRequests	0.00	0.000	1.0	5.0

7 first five row's of the data frame

```
[247]: HotalData.loc[0:5].T
```

```
[247]:
```

	0	1	\
IsCanceled	0	0	
LeadTime	342	737	
ArrivalDateYear	2015	2015	
ArrivalDateMonth	July	July	
ArrivalDateWeekNumber	27	27	
ArrivalDateDayOfMonth	1	1	
StaysInWeekendNights	0	0	
StaysInWeekNights	0	0	
Adults	2	2	
Children	0	0	
Babies	0	0	
Meal	BB	BB	
Country	PRT	PRT	
MarketSegment	Direct	Direct	
DistributionChannel	Direct	Direct	
IsRepeatedGuest	0	0	
PreviousCancellations	0	0	
PreviousBookingsNotCanceled	0	0	
ReservedRoomType	C	C	
AssignedRoomType	C	C	
BookingChanges	3	4	
DepositType	No Deposit	No Deposit	
Agent	NULL	NULL	
Company	NULL	NULL	
DaysInWaitingList	0	0	
CustomerType	Transient	Transient	
ADR	0	0	

RequiredCarParkingSpaces	0	0
TotalOfSpecialRequests	0	0
ReservationStatus	Check-Out	Check-Out
ReservationStatusDate	2015-07-01 00:00:00	2015-07-01 00:00:00
Hotal	resort hotal	resort hotal
	2	3 \
IsCanceled	0	0
LeadTime	7	13
ArrivalDateYear	2015	2015
ArrivalDateMonth	July	July
ArrivalDateWeekNumber	27	27
ArrivalDateDayOfMonth	1	1
StaysInWeekendNights	0	0
StaysInWeekNights	1	1
Adults	1	1
Children	0	0
Babies	0	0
Meal	BB	BB
Country	GBR	GBR
MarketSegment	Direct	Corporate
DistributionChannel	Direct	Corporate
IsRepeatedGuest	0	0
PreviousCancellations	0	0
PreviousBookingsNotCanceled	0	0
ReservedRoomType	A	A
AssignedRoomType	C	A
BookingChanges	0	0
DepositType	No Deposit	No Deposit
Agent	NULL	304
Company	NULL	NULL
DaysInWaitingList	0	0
CustomerType	Transient	Transient
ADR	75	75
RequiredCarParkingSpaces	0	0
TotalOfSpecialRequests	0	0
ReservationStatus	Check-Out	Check-Out
ReservationStatusDate	2015-07-02 00:00:00	2015-07-02 00:00:00
Hotal	resort hotal	resort hotal
	4	5
IsCanceled	0	0
LeadTime	14	14
ArrivalDateYear	2015	2015
ArrivalDateMonth	July	July
ArrivalDateWeekNumber	27	27
ArrivalDateDayOfMonth	1	1

StaysInWeekendNights	0	0
StaysInWeekNights	2	2
Adults	2	2
Children	0	0
Babies	0	0
Meal	BB	BB
Country	GBR	GBR
MarketSegment	Online TA	Online TA
DistributionChannel	TA/TO	TA/TO
IsRepeatedGuest	0	0
PreviousCancellations	0	0
PreviousBookingsNotCanceled	0	0
ReservedRoomType	A	A
AssignedRoomType	A	A
BookingChanges	0	0
DepositType	No Deposit	No Deposit
Agent	240	240
Company	NULL	NULL
DaysInWaitingList	0	0
CustomerType	Transient	Transient
ADR	98	98
RequiredCarParkingSpaces	0	0
TotalOfSpecialRequests	1	1
ReservationStatus	Check-Out	Check-Out
ReservationStatusDate	2015-07-03 00:00:00	2015-07-03 00:00:00
Hotal	resort hotal	resort hotal

```
[248]: HotalData.columns
```

```
[248]: Index(['IsCanceled', 'LeadTime', 'ArrivalDateYear', 'ArrivalDateMonth',
        'ArrivalDateWeekNumber', 'ArrivalDateDayOfMonth',
        'StaysInWeekendNights', 'StaysInWeekNights', 'Adults', 'Children',
        'Babies', 'Meal', 'Country', 'MarketSegment', 'DistributionChannel',
        'IsRepeatedGuest', 'PreviousCancellations',
        'PreviousBookingsNotCanceled', 'ReservedRoomType', 'AssignedRoomType',
        'BookingChanges', 'DepositType', 'Agent', 'Company',
        'DaysInWaitingList', 'CustomerType', 'ADR', 'RequiredCarParkingSpaces',
        'TotalOfSpecialRequests', 'ReservationStatus', 'ReservationStatusDate',
        'Hotal'],
        dtype='object')
```

```
[249]: HotalData.shape
```

```
[249]: (119390, 32)
```

8 Number of duplicated records in a dataset

```
[250]: HotalData.duplicated().sum()
```

```
[250]: 31994
```

9 remove the duplicated records in the data set

```
[251]: hotaldata = HotalData.drop_duplicates()
hotaldata.shape
```

```
[251]: (87396, 32)
```

10 Univariant_statistics of the data frame

```
[252]: output_hotaldata= pd.DataFrame(columns=[
    →['Count', 'NA', 'Null', 'Unique', 'dtype', 'Mean', 'Mode', 'Std', 'Min', 'Quantile-25%', 'Median', 'qu
def Univariant_statistics(hotaldata):
    for col in hotaldata:
        if pd.api.types.is_numeric_dtype(hotaldata[col]):
            output_hotaldata.loc[col] = [hotaldata[col].count(), hotaldata[col].
    →isna().sum(), hotaldata[col].isnull().sum(), hotaldata[col].nunique(),
                hotaldata[col].dtype, hotaldata[col].
    →mean(), hotaldata[col].mode().values[0], hotaldata[col].std(), hotaldata[col].
    →min(),
                hotaldata[col].quantile(0.25), hotaldata[col].
    →median(), hotaldata[col].quantile(0.75), hotaldata[col].max(), hotaldata[col].
    →skew(axis=0, skipna=True)]
        else:
            output_hotaldata.loc[col] = [hotaldata[col].count(), hotaldata[col].
    →isna().sum(), hotaldata[col].isnull().sum(), hotaldata[col].nunique(),
                hotaldata[col].
    →dtype, '-', '-', '-', '-', '-', '-', '-', '-', '-', '-']
            return output_hotaldata

Univariant_statistics(hotaldata)
```

```
[252]:
```

	Count	NA	Null	Unique	dtype \
IsCanceled	87396	0	0	2	category
LeadTime	87396	0	0	479	int64
ArrivalDateYear	87396	0	0	3	int64
ArrivalDateMonth	87396	0	0	12	category

ArrivalDateWeekNumber	87396	0	0	53	int64
ArrivalDateDayOfMonth	87396	0	0	31	int64
StaysInWeekendNights	87396	0	0	17	int64
StaysInWeekNights	87396	0	0	35	int64
Adults	87396	0	0	14	int64
Children	87392	4	4	5	float64
Babies	87396	0	0	5	int64
Meal	87396	0	0	5	object
Country	86944	452	452	177	object
MarketSegment	87396	0	0	8	object
DistributionChannel	87396	0	0	5	object
IsRepeatedGuest	87396	0	0	2	category
PreviousCancellations	87396	0	0	15	int64
PreviousBookingsNotCanceled	87396	0	0	73	int64
ReservedRoomType	87396	0	0	10	object
AssignedRoomType	87396	0	0	12	object
BookingChanges	87396	0	0	21	int64
DepositType	87396	0	0	3	object
Agent	87396	0	0	334	object
Company	87396	0	0	353	object
DaysInWaitingList	87396	0	0	128	int64
CustomerType	87396	0	0	4	object
ADR	87396	0	0	8879	float64
RequiredCarParkingSpaces	87396	0	0	5	int64
TotalOfSpecialRequests	87396	0	0	6	int64
ReservationStatus	87396	0	0	3	object
ReservationStatusDate	87396	0	0	926	datetime64[ns]
Hotal	87396	0	0	2	object

	Mean	Mode	Std	Min	Quantile-25%	\
IsCanceled	-	-	-	-	-	-
LeadTime	79.8914	0	86.0523	0	11	
ArrivalDateYear	2016.21	2016	0.686102	2015	2016	
ArrivalDateMonth	-	-	-	-	-	-
ArrivalDateWeekNumber	26.8383	33	13.6746	1	16	
ArrivalDateDayOfMonth	15.8155	17	8.83515	1	8	
StaysInWeekendNights	1.00526	0	1.03192	0	0	
StaysInWeekNights	2.62539	1	2.05358	0	1	
Adults	1.8758	2	0.6265	0	2	
Children	0.13864	0	0.455881	0	0	
Babies	0.0108243	0	0.113597	0	0	
Meal	-	-	-	-	-	-
Country	-	-	-	-	-	-
MarketSegment	-	-	-	-	-	-
DistributionChannel	-	-	-	-	-	-
IsRepeatedGuest	-	-	-	-	-	-
PreviousCancellations	0.0304133	0	0.369145	0	0	

PreviousBookingsNotCanceled	0.18399	0	1.73189	0	0
ReservedRoomType	-	-	-	-	-
AssignedRoomType	-	-	-	-	-
BookingChanges	0.271603	0	0.727245	0	0
DepositType	-	-	-	-	-
Agent	-	-	-	-	-
Company	-	-	-	-	-
DaysInWaitingList	0.749565	0	10.0157	0	0
CustomerType	-	-	-	-	-
ADR	106.337	0	55.014	-6.38	72
RequiredCarParkingSpaces	0.0842258	0	0.281533	0	0
TotalOfSpecialRequests	0.698567	0	0.831946	0	0
ReservationStatus	-	-	-	-	-
ReservationStatusDate	-	-	-	-	-
Hotal	-	-	-	-	-

	Median	quantile-75%	Max	Skewness
IsCanceled	-	-	-	-
LeadTime	49	125	737	1.43177
ArrivalDateYear	2016	2017	2017	-0.297194
ArrivalDateMonth	-	-	-	-
ArrivalDateWeekNumber	27	37	53	0.0225036
ArrivalDateDayOfMonth	16	23	31	0.000308252
StaysInWeekendNights	1	2	19	1.41164
StaysInWeekNights	2	4	50	2.69132
Adults	2	2	55	19.8596
Children	0	0	10	3.46357
Babies	0	0	10	21.1485
Meal	-	-	-	-
Country	-	-	-	-
MarketSegment	-	-	-	-
DistributionChannel	-	-	-	-
IsRepeatedGuest	-	-	-	-
PreviousCancellations	0	0	26	34.3237
PreviousBookingsNotCanceled	0	0	72	20.4597
ReservedRoomType	-	-	-	-
AssignedRoomType	-	-	-	-
BookingChanges	0	0	21	5.54633
DepositType	-	-	-	-
Agent	-	-	-	-
Company	-	-	-	-
DaysInWaitingList	0	0	391	19.3964
CustomerType	-	-	-	-
ADR	98.1	134	5400	10.9214
RequiredCarParkingSpaces	0	0	8	3.49025
TotalOfSpecialRequests	0	1	5	1.08285
ReservationStatus	-	-	-	-

ReservationStatusDate	-	-	-	-
Hotal	-	-	-	-

11 Check No of null values in each attribute

```
[253]: hotaldata.isnull().sum()
```

```
[253]: IsCanceled          0
      LeadTime             0
      ArrivalDateYear      0
      ArrivalDateMonth     0
      ArrivalDateWeekNumber 0
      ArrivalDateDayOfMonth 0
      StaysInWeekendNights  0
      StaysInWeekNights     0
      Adults               0
      Children             4
      Babies               0
      Meal                 0
      Country              452
      MarketSegment        0
      DistributionChannel   0
      IsRepeatedGuest       0
      PreviousCancellations 0
      PreviousBookingsNotCanceled 0
      ReservedRoomType      0
      AssignedRoomType      0
      BookingChanges        0
      DepositType           0
      Agent                 0
      Company               0
      DaysInWaitingList     0
      CustomerType          0
      ADR                   0
      RequiredCarParkingSpaces 0
      TotalOfSpecialRequests 0
      ReservationStatus      0
      ReservationStatusDate  0
      Hotal                 0
      dtype: int64
```

12 check the unique values in Meal Attribute

```
[254]: hoteldata["Meal"].unique()  
hoteldata.Meal.unique()
```

```
[254]: array(['BB', 'FB', 'HB', 'SC', 'Undefined'],  
          dtype=object)
```

```
[254]: array(['BB', 'FB', 'HB', 'SC', 'Undefined'],  
          dtype=object)
```

13 check the unique values in DistributionChannel Attribute

```
[255]: hoteldata.DistributionChannel.unique()
```

```
[255]: array(['Direct', 'Corporate', 'TA/TO', 'Undefined', 'GDS'], dtype=object)
```

14 check the unique values in Country Attribute

```
[256]: hoteldata.Country.unique()
```

```
[256]: array(['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'], dtype=object)
```

15 check the unique values in CustomerType Attribute

```
[257]: hoteldata.CustomerType.unique()
```

```
[257]: array(['Transient', 'Contract', 'Transient-Party', 'Group'], dtype=object)
```

16 Check data with null values

```
[258]: hoteldata[hoteldata.isnull().any(axis=1)].head(20)
```

```
[258]:
```

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
30	0	118	2015	July	
4127	1	0	2016	February	
7092	1	8	2016	July	
7860	1	39	2016	August	
8779	1	0	2016	October	
9376	1	0	2016	November	
9610	1	4	2016	December	
11262	1	87	2017	April	
13803	0	7	2016	January	
13804	0	7	2016	January	
13805	1	2	2016	October	
13807	0	7	2016	January	
13809	0	7	2016	September	
13810	0	66	2015	September	
13811	1	71	2016	July	
13816	0	6	2016	April	
13819	0	11	2016	February	
13820	0	1	2016	March	
13823	0	7	2016	January	
13828	0	5	2016	March	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
30	27	1	4	
4127	8	15	0	
7092	30	21	0	
7860	36	30	0	
8779	42	13	0	
9376	48	21	0	
9610	52	20	0	
11262	18	30	2	
13803	3	13	0	
13804	4	18	1	
13805	42	12	0	
13807	3	13	0	

13809	38	13	0
13810	39	26	4
13811	29	14	3
13816	15	6	0
13819	6	1	1
13820	11	8	0
13823	4	20	0
13828	11	8	0

	StaysInWeekNights	Adults	Children	...	Agent	Company	\
30	10	1	0.0	...	NULL	NULL	
4127	0	0	0.0	...	NULL	383	
7092	1	1	0.0	...	NULL	204	
7860	5	2	0.0	...	NULL	NULL	
8779	1	1	0.0	...	NULL	457	
9376	0	0	0.0	...	NULL	386	
9610	1	1	0.0	...	NULL	204	
11262	3	2	0.0	...	6	NULL	
13803	2	1	0.0	...	94	NULL	
13804	3	1	0.0	...	NULL	94	
13805	2	1	0.0	...	NULL	94	
13807	2	1	0.0	...	94	NULL	
13809	2	1	0.0	...	94	NULL	
13810	6	2	0.0	...	NULL	NULL	
13811	8	2	0.0	...	NULL	NULL	
13816	2	1	0.0	...	NULL	94	
13819	3	1	0.0	...	94	NULL	
13820	2	1	0.0	...	94	NULL	
13823	1	1	0.0	...	NULL	94	
13828	2	1	0.0	...	NULL	94	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
30	0	Transient	62.0	0	
4127	0	Transient	0.0	0	
7092	0	Transient	73.0	0	
7860	0	Transient	159.0	0	
8779	0	Transient	50.0	0	
9376	0	Group	0.0	0	
9610	0	Transient	30.0	0	
11262	0	Transient	48.0	0	
13803	0	Transient	27.0	0	
13804	0	Transient	28.5	1	
13805	0	Transient	37.0	0	
13807	0	Transient	27.0	0	
13809	0	Transient	66.0	1	
13810	0	Transient	115.0	1	
13811	0	Transient	126.6	0	

13816	0	Transient	39.0	0
13819	0	Transient	27.0	0
13820	0	Transient	27.0	0
13823	0	Transient	25.0	0
13828	0	Transient	27.0	1

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
30	2	Check-Out	2015-07-15	
4127	0	Canceled	2016-02-15	
7092	2	Canceled	2016-07-20	
7860	5	Canceled	2016-07-22	
8779	0	Canceled	2016-10-13	
9376	0	Canceled	2016-11-21	
9610	2	Canceled	2016-12-19	
11262	0	Canceled	2017-02-02	
13803	0	Check-Out	2016-01-15	
13804	0	Check-Out	2016-01-22	
13805	0	Canceled	2016-10-10	
13807	0	Check-Out	2016-01-15	
13809	0	Check-Out	2016-09-15	
13810	2	Check-Out	2015-10-06	
13811	1	Canceled	2016-07-12	
13816	0	Check-Out	2016-04-08	
13819	0	Check-Out	2016-02-05	
13820	0	Check-Out	2016-03-10	
13823	0	Check-Out	2016-01-21	
13828	0	Check-Out	2016-03-10	

	Hotal
30	resort hotal
4127	resort hotal
7092	resort hotal
7860	resort hotal
8779	resort hotal
9376	resort hotal
9610	resort hotal
11262	resort hotal
13803	resort hotal
13804	resort hotal
13805	resort hotal
13807	resort hotal
13809	resort hotal
13810	resort hotal
13811	resort hotal
13816	resort hotal
13819	resort hotal
13820	resort hotal

```
13823 resort hotal
13828 resort hotal
```

```
[20 rows x 32 columns]
```

17 Replace missing values:

18 agent: If no agency is given, booking was most likely made without one.

19 company feature has many null values. So this column can be dropped.

20 rest schould be self-explanatory.

```
[259]: hotaldata_clean=hotaldata.drop(columns='Company')
      hotaldata_clean
```

```
[259]:
```

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
0	0	342	2015	July	
1	0	737	2015	July	
2	0	7	2015	July	
3	0	13	2015	July	
4	0	14	2015	July	
...	
119385	0	23	2017	August	
119386	0	102	2017	August	
119387	0	34	2017	August	
119388	0	109	2017	August	
119389	0	205	2017	August	

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	0	
4	27	1	0	
...	
119385	35	30	2	
119386	35	31	2	
119387	35	31	2	
119388	35	31	2	

119389

35

29

2

	StaysInWeekNights	Adults	Children	...	DepositType \
0	0	2	0.0	...	No Deposit
1	0	2	0.0	...	No Deposit
2	1	1	0.0	...	No Deposit
3	1	1	0.0	...	No Deposit
4	2	2	0.0	...	No Deposit
...
119385	5	2	0.0	...	No Deposit
119386	5	3	0.0	...	No Deposit
119387	5	2	0.0	...	No Deposit
119388	5	2	0.0	...	No Deposit
119389	7	2	0.0	...	No Deposit

	Agent	DaysInWaitingList	CustomerType	ADR \
0	NULL		0 Transient	0.00
1	NULL		0 Transient	0.00
2	NULL		0 Transient	75.00
3	304		0 Transient	75.00
4	240		0 Transient	98.00
...
119385	394		0 Transient	96.14
119386	9		0 Transient	225.43
119387	9		0 Transient	157.71
119388	89		0 Transient	104.40
119389	9		0 Transient	151.20

	RequiredCarParkingSpaces	TotalOfSpecialRequests	ReservationStatus \
0	0	0	Check-Out
1	0	0	Check-Out
2	0	0	Check-Out
3	0	0	Check-Out
4	0	1	Check-Out
...
119385	0	0	Check-Out
119386	0	2	Check-Out
119387	0	4	Check-Out
119388	0	0	Check-Out
119389	0	2	Check-Out

	ReservationStatusDate	Hotal
0	2015-07-01	resort hotal
1	2015-07-01	resort hotal
2	2015-07-02	resort hotal
3	2015-07-02	resort hotal
4	2015-07-03	resort hotal


```

...
119385      2017-09-06    city hotal
119386      2017-09-07    city hotal
119387      2017-09-07    city hotal
119388      2017-09-07    city hotal
119389      2017-09-07    city hotal

```

[87396 rows x 31 columns]

```
[260]: hotaldata_clean["Agent"] = pd.to_numeric(hotaldata_clean["Agent"].str.
        ↪replace(",", ""), errors='coerce')
```

```
[261]: hotaldata_clean['Agent'].isna()
```

```

[261]: 0          True
       1          True
       2          True
       3         False
       4         False
       ...
119385    False
119386    False
119387    False
119388    False
119389    False
Name: Agent, Length: 87396, dtype: bool

```

```

[262]: hotaldata_clean['Agent'].fillna(value=0,inplace=True)

hotaldata_clean['Agent'].isna()

```

```

[262]: 0          False
       1          False
       2          False
       3          False
       4          False
       ...
119385    False
119386    False
119387    False
119388    False
119389    False
Name: Agent, Length: 87396, dtype: bool

```

```
[263]: hotaldata_clean["Agent"] = hotaldata_clean["Agent"].astype("category")
```

```
[264]: nan_replacements = {"Country": "Unknown"}
      hotaldata_clean = hotaldata_clean.fillna(nan_replacements)
```

21 “meal” contains values “Undefined”, which is equal to SC.

22 Meal types

23 Undefined/SC = No meal package

24 BB = Breakfast and dinner

25 HB = Half board (Breakfast and one other meal – usually dinner)

26 FB = Full board (Breakfast, lunch and dinner)

```
[265]: hotaldata_clean["Meal"] = hotaldata_clean["Meal"].replace("Undefined", "SC")
      hotaldata_clean["Meal"].replace("Undefined", "SC", inplace=True)
      hotaldata_clean.isnull().sum()
```

```
[265]: IsCanceled          0
      LeadTime            0
      ArrivalDateYear     0
      ArrivalDateMonth    0
      ArrivalDateWeekNumber 0
      ArrivalDateDayOfMonth 0
      StaysInWeekendNights 0
      StaysInWeekNights    0
      Adults              0
      Children            4
      Babies              0
      Meal                0
      Country             0
      MarketSegment       0
      DistributionChannel  0
      IsRepeatedGuest      0
      PreviousCancellations 0
      PreviousBookingsNotCanceled 0
      ReservedRoomType     0
      AssignedRoomType     0
      BookingChanges       0
```

```

DepositType          0
Agent                0
DaysInWaitingList    0
CustomerType         0
ADR                  0
RequiredCarParkingSpaces 0
TotalOfSpecialRequests 0
ReservationStatus     0
ReservationStatusDate 0
Hotal                 0
dtype: int64

```

```

[266]: hotaldata_clean_no_adult = hotaldata_clean[(hotaldata_clean["Adults"]==0 )]
hotaldata_clean_no_adult_no_children =
↳hotaldata_clean_no_adult[hotaldata_clean_no_adult["Children"]==0]
hotaldata_clean_no_adult_no_children_no_babies =
↳hotaldata_clean_no_adult_no_children[hotaldata_clean_no_adult_no_children["Babies"]==0]
hotaldata_clean.drop(hotaldata_clean_no_adult_no_children_no_babies.index,
↳inplace = True)

```

27 check unique values in Children attribute

```

[267]: hotaldata_clean["Children"].unique()

```

```

[267]: array([ 0.,  1.,  2., 10.,  3., nan])

```

28 total number of na values in Children attribute

```

[268]: hotaldata_clean["Children"].isnull().sum()

```

```

[268]: 4

```

```

[269]: hotaldata_clean["Children"].isna().sum()

```

```

[269]: 4

```

calculating mean to inplace with null values in children attribute

```

[270]: mean_age= round(hotaldata_clean['Children'].mean(),0)
print(mean_age)

```

```

0.0

```

29 Replacing na values with mean of the children attribute

```
[271]: hotaldata_clean['Children'].fillna(value=mean_age,inplace=True)
hotaldata_clean.isnull().values.any().sum()
```

```
[271]: 0
```

30 Correlation matrix

```
[272]: corr = hotaldata_clean.corr()
corr
```

```
[272]:
```

	LeadTime	ArrivalDateYear	ArrivalDateWeekNumber	\
LeadTime	1.000000	0.139408	0.101295	
ArrivalDateYear	0.139408	1.000000	-0.513895	
ArrivalDateWeekNumber	0.101295	-0.513895	1.000000	
ArrivalDateDayOfMonth	0.009899	-0.009872	0.093273	
StaysInWeekendNights	0.236316	0.005277	0.027466	
StaysInWeekNights	0.312506	0.003765	0.028498	
Adults	0.139033	0.038744	0.025322	
Children	0.028395	0.041178	0.013542	
Babies	-0.003705	-0.023366	0.014302	
PreviousCancellations	0.005370	-0.054284	0.007038	
PreviousBookingsNotCanceled	-0.079021	0.027266	-0.020938	
BookingChanges	0.080822	0.008944	0.012943	
DaysInWaitingList	0.131909	-0.027490	0.013159	
ADR	0.021948	0.176594	0.099092	
RequiredCarParkingSpaces	-0.086736	-0.040041	0.009054	
TotalOfSpecialRequests	0.034048	0.064210	0.046759	

	ArrivalDateDayOfMonth	StaysInWeekendNights	\
LeadTime	0.009899	0.236316	
ArrivalDateYear	-0.009872	0.005277	
ArrivalDateWeekNumber	0.093273	0.027466	
ArrivalDateDayOfMonth	1.000000	-0.017636	
StaysInWeekendNights	-0.017636	1.000000	
StaysInWeekNights	-0.028375	0.550471	
Adults	-0.001235	0.091450	
Children	0.015826	0.028911	
Babies	-0.000397	0.013802	
PreviousCancellations	-0.008553	-0.020641	
PreviousBookingsNotCanceled	0.000148	-0.056891	
BookingChanges	0.007006	0.034624	
DaysInWaitingList	0.006088	-0.031980	
ADR	0.022682	0.040417	

RequiredCarParkingSpaces	0.009040	-0.042979			
TotalOfSpecialRequests	-0.001677	0.032816			
	StaysInWeekNights	Adults	Children	Babies	\
LeadTime	0.312506	0.139033	0.028395	-0.003705	
ArrivalDateYear	0.003765	0.038744	0.041178	-0.023366	
ArrivalDateWeekNumber	0.028498	0.025322	0.013542	0.014302	
ArrivalDateDayOfMonth	-0.028375	-0.001235	0.015826	-0.000397	
StaysInWeekendNights	0.550471	0.091450	0.028911	0.013802	
StaysInWeekNights	1.000000	0.099323	0.030936	0.016201	
Adults	0.099323	1.000000	0.022171	0.016236	
Children	0.030936	0.022171	1.000000	0.017370	
Babies	0.016201	0.016236	0.017370	1.000000	
PreviousCancellations	-0.018696	-0.042389	-0.019208	-0.005394	
PreviousBookingsNotCanceled	-0.058714	-0.122105	-0.029446	-0.009199	
BookingChanges	0.066425	-0.036446	0.033409	0.083425	
DaysInWaitingList	0.001855	-0.014925	-0.020370	-0.006825	
ADR	0.055084	0.241996	0.326359	0.023206	
RequiredCarParkingSpaces	-0.044484	0.007015	0.036267	0.030873	
TotalOfSpecialRequests	0.038336	0.112509	0.044503	0.094857	

	PreviousCancellations	\
LeadTime	0.005370	
ArrivalDateYear	-0.054284	
ArrivalDateWeekNumber	0.007038	
ArrivalDateDayOfMonth	-0.008553	
StaysInWeekendNights	-0.020641	
StaysInWeekNights	-0.018696	
Adults	-0.042389	
Children	-0.019208	
Babies	-0.005394	
PreviousCancellations	1.000000	
PreviousBookingsNotCanceled	0.391766	
BookingChanges	-0.010267	
DaysInWaitingList	0.003733	
ADR	-0.050359	
RequiredCarParkingSpaces	-0.003480	
TotalOfSpecialRequests	0.001695	

	PreviousBookingsNotCanceled	BookingChanges	\
LeadTime	-0.079021	0.080822	
ArrivalDateYear	0.027266	0.008944	
ArrivalDateWeekNumber	-0.020938	0.012943	
ArrivalDateDayOfMonth	0.000148	0.007006	
StaysInWeekendNights	-0.056891	0.034624	
StaysInWeekNights	-0.058714	0.066425	
Adults	-0.122105	-0.036446	

Children	-0.029446	0.033409
Babies	-0.009199	0.083425
PreviousCancellations	0.391766	-0.010267
PreviousBookingsNotCanceled	1.000000	0.005774
BookingChanges	0.005774	1.000000
DaysInWaitingList	-0.006334	0.024143
ADR	-0.086052	0.010144
RequiredCarParkingSpaces	0.040712	0.050659
TotalOfSpecialRequests	0.027044	0.018308

	DaysInWaitingList	ADR \
LeadTime	0.131909	0.021948
ArrivalDateYear	-0.027490	0.176594
ArrivalDateWeekNumber	0.013159	0.099092
ArrivalDateDayOfMonth	0.006088	0.022682
StaysInWeekendNights	-0.031980	0.040417
StaysInWeekNights	0.001855	0.055084
Adults	-0.014925	0.241996
Children	-0.020370	0.326359
Babies	-0.006825	0.023206
PreviousCancellations	0.003733	-0.050359
PreviousBookingsNotCanceled	-0.006334	-0.086052
BookingChanges	0.024143	0.010144
DaysInWaitingList	1.000000	-0.032522
ADR	-0.032522	1.000000
RequiredCarParkingSpaces	-0.015995	0.038733
TotalOfSpecialRequests	-0.049438	0.137461

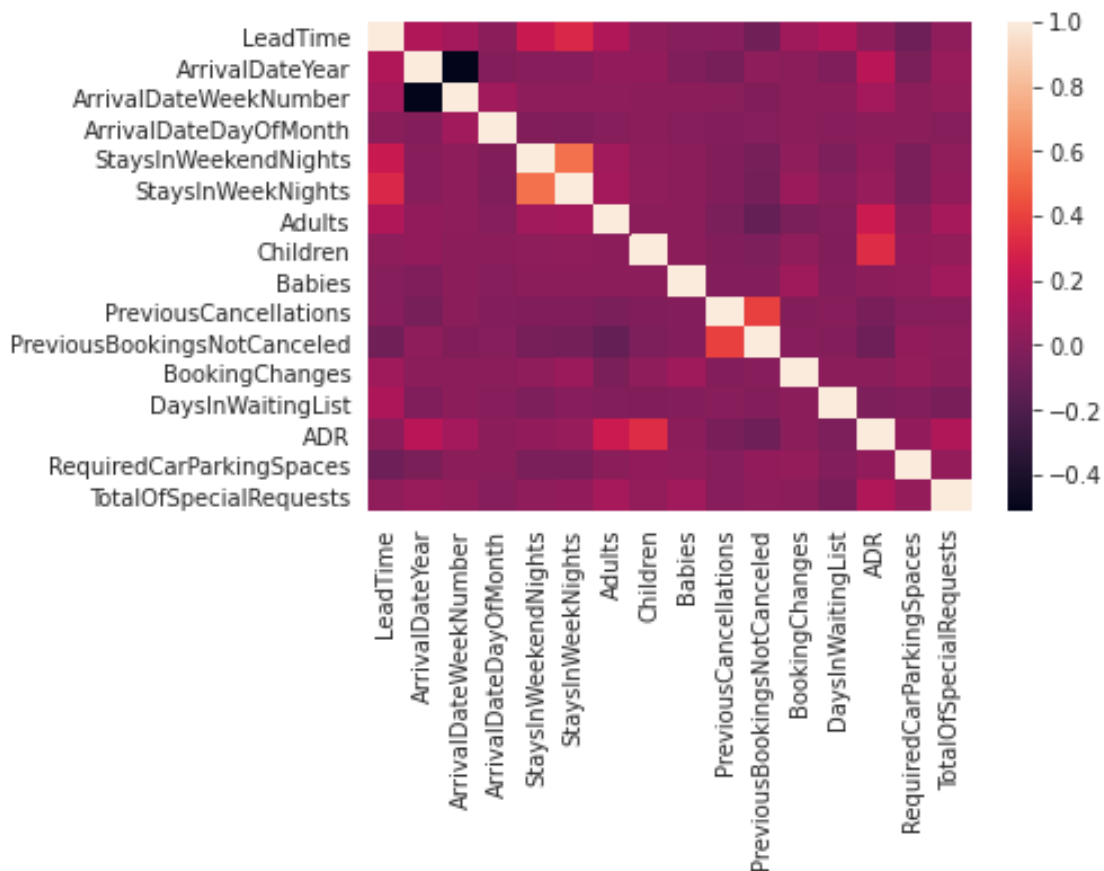
	RequiredCarParkingSpaces	TotalOfSpecialRequests
LeadTime	-0.086736	0.034048
ArrivalDateYear	-0.040041	0.064210
ArrivalDateWeekNumber	0.009054	0.046759
ArrivalDateDayOfMonth	0.009040	-0.001677
StaysInWeekendNights	-0.042979	0.032816
StaysInWeekNights	-0.044484	0.038336
Adults	0.007015	0.112509
Children	0.036267	0.044503
Babies	0.030873	0.094857
PreviousCancellations	-0.003480	0.001695
PreviousBookingsNotCanceled	0.040712	0.027044
BookingChanges	0.050659	0.018308
DaysInWaitingList	-0.015995	-0.049438
ADR	0.038733	0.137461
RequiredCarParkingSpaces	1.000000	0.048479
TotalOfSpecialRequests	0.048479	1.000000

[]:

31 Graphical Representation of Correlation matrix

```
[273]: sns.heatmap(corr,
                    xticklabels=corr.columns.values,
                    yticklabels=corr.columns.values)
```

```
[273]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6c9a9aab90>
```



```
[274]: hotaldata_clean.to_csv("hotaldataClean1.csv",index=False)
hotaldata_clean
```

```
[274]:
```

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	\
0	0	342	2015	July	
1	0	737	2015	July	
2	0	7	2015	July	
3	0	13	2015	July	
4	0	14	2015	July	
...	
119385	0	23	2017	August	

119386	0	102	2017	August
119387	0	34	2017	August
119388	0	109	2017	August
119389	0	205	2017	August

	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	\
0	27	1	0	
1	27	1	0	
2	27	1	0	
3	27	1	0	
4	27	1	0	
...	
119385	35	30	2	
119386	35	31	2	
119387	35	31	2	
119388	35	31	2	
119389	35	29	2	

	StaysInWeekNights	Adults	Children	...	DepositType	Agent	\
0	0	2	0.0	...	No Deposit	0.0	
1	0	2	0.0	...	No Deposit	0.0	
2	1	1	0.0	...	No Deposit	0.0	
3	1	1	0.0	...	No Deposit	304.0	
4	2	2	0.0	...	No Deposit	240.0	
...	
119385	5	2	0.0	...	No Deposit	394.0	
119386	5	3	0.0	...	No Deposit	9.0	
119387	5	2	0.0	...	No Deposit	9.0	
119388	5	2	0.0	...	No Deposit	89.0	
119389	7	2	0.0	...	No Deposit	9.0	

	DaysInWaitingList	CustomerType	ADR	RequiredCarParkingSpaces	\
0	0	Transient	0.00	0	
1	0	Transient	0.00	0	
2	0	Transient	75.00	0	
3	0	Transient	75.00	0	
4	0	Transient	98.00	0	
...	
119385	0	Transient	96.14	0	
119386	0	Transient	225.43	0	
119387	0	Transient	157.71	0	
119388	0	Transient	104.40	0	
119389	0	Transient	151.20	0	

	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate	\
0	0	Check-Out	2015-07-01	
1	0	Check-Out	2015-07-01	

2		0	Check-Out	2015-07-02
3		0	Check-Out	2015-07-02
4		1	Check-Out	2015-07-03
...
119385		0	Check-Out	2017-09-06
119386		2	Check-Out	2017-09-07
119387		4	Check-Out	2017-09-07
119388		0	Check-Out	2017-09-07
119389		2	Check-Out	2017-09-07

		Hotal
0	resort	hotal
1	resort	hotal
2	resort	hotal
3	resort	hotal
4	resort	hotal
...		...
119385	city	hotal
119386	city	hotal
119387	city	hotal
119388	city	hotal
119389	city	hotal

[87230 rows x 31 columns]