Data Preprocessing: Hotel Booking Dataset

Importing Libaries for Data Preprocessing

```
import numpy as np #used for handling numbers
import pandas as pd #used for handling datasets
import datetime
import seaborn as sn
import matplotlib.pyplot as plt
```

Importing dataset and printing the first few values

```
hotel = pd.read_csv(r"F:\DA imp\DA course\projects\Python\Hotel Booking DATA Preprocessing\hotel_bookings.csv\hotel_bookings.csv")
hotel.head()
```

	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	...	deposit_type
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit

5 rows × 32 columns

```
hotel.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
```

Checking For missing values

```
X = hotel.isna().sum()
print(X)
```

```
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 Missing Children value with 0 and Missing Country with not available ¶

```
replacing = {'children':0,'country': 'Not Available'}
hotel_new = hotel.fillna(replacing)
```

Dropping company and agent columns

```
hotel_new.drop(['company','agent'], axis = 1, inplace = True)

hotel_new.head()
```

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

Changing datatype from float to int

```
hotel_new['children'] = hotel_new['children'].astype(int)
```

Including Babies in Children column

```
hotel_new['Children'] = hotel_new['children'] + hotel_new['babies']
hotel_new.drop(['babies'],axis = 1, inplace = True)
```

Undefined is equal to SC that is no meal package.

```
hotel_new['meal'].replace("Undefined", "SC")
```

```
0         BB
1         BB
2         BB
3         BB
4         BB
          ..
119385    BB
119386    BB
119387    BB
119388    BB
119389    HB
Name: meal, Length: 119390, dtype: object
```

Making two new columns and dropping the redundant column.

```
hotel_new['total_guests'] = hotel_new['adults'] + hotel_new['children']
hotel_new['total_stays']= hotel_new['stays_in_weekend_nights'] + hotel_new['stays_in_week_nights']
```

```
hotel_new.drop(['stays_in_week_nights'],axis = 1, inplace = True)
```

Dropping entries that include 0 Total Guests.

```
hotel_new = hotel_new[hotel_new['total_guests'] != 0]
```

Making column arrival date.

Converting string month to numerical one.

```
datetime_object = hotel['arrival_date_month'].str[0:3]
month_number = np.zeros(len(datetime_object))
```

Creating a new column based on numerical representation of the months.

```
for i in range(0,len(datetime_object)):
    datetime[i] = datetime.datetime.strptime(datetime_object[i],"%b")
    month_number[i] - datetime_object[i].month
```

Float to integer conversion.

```
month_number = pd.DataFrame(month_number).astype(int)
```

Creating Arrival date.

```
hotel_new.loc[:, 'arrival_date'] = hotel['arrival_date_year'].map(str) + '-' + \
                                   month_number[0].map(str) + '-' + \
                                   hotel['arrival_date_day_of_month'].map(str)
```

Dropping the redundant column

```
hotel_new=hotel_new.drop(['arrival_date_year', 'arrival_date_month', 'arrival_date_day_of_month',
                  'arrival_date_week_number'], axis=1)
```

Converting wrong datatype columns

```
hotel_new['arrival_date'] = pd.to_datetime(hotel_new['arrival_date'])
hotel_new['reservation_status_date'] = pd.to_datetime(hotel_new['reservation_status_date'])
```

Final info

```
hotel_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 119210 entries, 0 to 119389
Data columns (total 28 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119210 non-null  object
 1   is_canceled                     119210 non-null  int64
 2   lead_time                       119210 non-null  int64
 3   stays_in_weekend_nights         119210 non-null  int64
 4   adults                          119210 non-null  int64
 5   children                        119210 non-null  int64
 6   meal                            119210 non-null  object
 7   country                         119210 non-null  object
 8   market_segment                  119210 non-null  object
 9   distribution_channel            119210 non-null  object
 10  is_repeated_guest               119210 non-null  int64
 11  previous_cancellations          119210 non-null  int64
 12  previous_bookings_not_canceled  119210 non-null  int64
 13  reserved_room_type              119210 non-null  object
 14  assigned_room_type              119210 non-null  object
 15  booking_changes                 119210 non-null  int64
 16  deposit_type                    119210 non-null  object
 17  days_in_waiting_list            119210 non-null  int64
 18  customer_type                   119210 non-null  object
 19  adr                             119210 non-null  float64
 20  required_car_parking_spaces     119210 non-null  int64
 21  total_of_special_requests       119210 non-null  int64
 22  reservation_status              119210 non-null  object
 23  reservation_status_date         119210 non-null  datetime64[ns]
 24  Children                        119210 non-null  int64
 25  total_guests                    119210 non-null  int64
 26  total_stays                     119210 non-null  int64
 27  arrival_date                    119210 non-null  datetime64[ns]
dtypes: datetime64[ns](2), float64(1), int64(15), object(10)
memory usage: 26.4+ MB
```

Checking if any missing values remaining

```
hotel_new.isna().sum()
```

```
hotel                           0
is_canceled                     0
lead_time                       0
stays_in_weekend_nights         0
adults                          0
children                        0
meal                            0
country                         0
market_segment                  0
distribution_channel            0
is_repeated_guest               0
```

```
previous_cancellations            0
previous_bookings_not_canceled    0
reserved_room_type                0
assigned_room_type                0
booking_changes                   0
deposit_type                      0
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
Children                          0
total_guests                      0
total_stays                       0
arrival_date                      0
```