Importing libraries

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import preprocessing
import numpy as np
import seaborn as sns
```

Data collection

Importing dataset

```
df = pd.read_csv(r"hotel_bookings.csv")
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
...                   ...                       ...
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                       0
1                               1                       0
2                               1                       0
3                               1                       0
4                               1                       0
...                           ...                     ...
```

```
119385                           30                           2
119386                           31                           2
119387                           31                           2
119388                           31                           2
119389                           29                           2

        stays_in_week_nights  adults  ...  deposit_type  agent company
\
0                          0       2  ...    No Deposit    NaN     NaN

1                          0       2  ...    No Deposit    NaN     NaN

2                          1       1  ...    No Deposit    NaN     NaN

3                          1       1  ...    No Deposit  304.0     NaN

4                          2       2  ...    No Deposit  240.0     NaN

...                      ...     ...  ...           ...    ...     ...

119385                     5       2  ...    No Deposit  394.0     NaN

119386                     5       3  ...    No Deposit    9.0     NaN

119387                     5       2  ...    No Deposit    9.0     NaN

119388                     5       2  ...    No Deposit   89.0     NaN

119389                     7       2  ...    No Deposit    9.0     NaN


       days_in_waiting_list customer_type     adr  \
0                         0     Transient    0.00
1                         0     Transient    0.00
2                         0     Transient   75.00
3                         0     Transient   75.00
4                         0     Transient   98.00
...                     ...           ...     ...
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     Transient  151.20

        required_car_parking_spaces  total_of_special_requests  \
0                                 0                          0
1                                 0                          0
2                                 0                          0
3                                 0                          0
4                                 0                          1
...                             ...                        ...
```

```
119385                              0                            0
119386                              0                            2
119387                              0                            4
119388                              0                            0
119389                              0                            2

        reservation_status reservation_status_date
0                Check-Out              2015-07-01
1                Check-Out              2015-07-01
2                Check-Out              2015-07-02
3                Check-Out              2015-07-02
4                Check-Out              2015-07-03
...                    ...                     ...
119385           Check-Out              2017-09-06
119386           Check-Out              2017-09-07
119387           Check-Out              2017-09-07
119388           Check-Out              2017-09-07
119389           Check-Out              2017-09-07

[119390 rows x 32 columns]
```

Basic details

```
df.size
```

```
3820480
```

```
df.shape
```

```
(119390, 32)
```

```
list(df.columns)
```

```
['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']
```

```
df.dtypes
```

```
hotel                            object
is_canceled                       int64
lead_time                         int64
arrival_date_year                 int64
arrival_date_month               object
arrival_date_week_number          int64
arrival_date_day_of_month         int64
stays_in_weekend_nights           int64
stays_in_week_nights              int64
adults                            int64
children                        float64
babies                            int64
meal                             object
country                          object
market_segment                   object
distribution_channel             object
is_repeated_guest                 int64
previous_cancellations            int64
previous_bookings_not_canceled    int64
reserved_room_type               object
assigned_room_type               object
booking_changes                   int64
deposit_type                     object
agent                           float64
company                         float64
days_in_waiting_list              int64
customer_type                    object
adr                             float64
required_car_parking_spaces       int64
total_of_special_requests         int64
reservation_status               object
reservation_status_date          object
dtype: object
```

```
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
```

Statistical details

```
df.describe()
```

```
         is_canceled      lead_time  arrival_date_year  \
count  119390.000000  119390.000000      119390.000000   
mean        0.370416     104.011416        2016.156554   
std         0.482918     106.863097           0.707476   
```

```
min          0.000000       0.000000        2015.000000
25%          0.000000      18.000000        2016.000000
50%          0.000000      69.000000        2016.000000
75%          1.000000     160.000000        2017.000000
max          1.000000     737.000000        2017.000000

       arrival_date_week_number  arrival_date_day_of_month  \
count             119390.000000              119390.000000
mean                  27.165173                  15.798241
std                   13.605138                   8.780829
min                    1.000000                   1.000000
25%                   16.000000                   8.000000
50%                   28.000000                  16.000000
75%                   38.000000                  23.000000
max                   53.000000                  31.000000

       stays_in_weekend_nights  stays_in_week_nights         adults  \
count            119390.000000         119390.000000  119390.000000
mean                  0.927599              2.500302       1.856403
std                   0.998613              1.908286       0.579261
min                   0.000000              0.000000       0.000000
25%                   0.000000              1.000000       2.000000
50%                   1.000000              2.000000       2.000000
75%                   2.000000              3.000000       2.000000
max                  19.000000             50.000000      55.000000

            children         babies  is_repeated_guest  \
count  119386.000000  119390.000000      119390.000000
mean        0.103890       0.007949           0.031912
std         0.398561       0.097436           0.175767
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        10.000000      10.000000           1.000000

       previous_cancellations  previous_bookings_not_canceled  \
count           119390.000000                   119390.000000
mean                0.087118                        0.137097
std                 0.844336                        1.497437
min                 0.000000                        0.000000
25%                 0.000000                        0.000000
50%                 0.000000                        0.000000
75%                 0.000000                        0.000000
max                26.000000                       72.000000

       booking_changes          agent      company
days_in_waiting_list  \
count    119390.000000  103050.000000  6797.000000
119390.000000
```

```
mean           0.221124      86.693382   189.266735
2.321149
std            0.652306     110.774548   131.655015
17.594721
min            0.000000       1.000000     6.000000
0.000000
25%            0.000000       9.000000    62.000000
0.000000
50%            0.000000      14.000000   179.000000
0.000000
75%            0.000000     229.000000   270.000000
0.000000
max           21.000000     535.000000   543.000000
391.000000

                   adr  required_car_parking_spaces
total_of_special_requests
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
```

```
df.count()
```

```
hotel                           119390
is_canceled                     119390
lead_time                       119390
arrival_date_year               119390
arrival_date_month              119390
arrival_date_week_number        119390
arrival_date_day_of_month       119390
stays_in_weekend_nights         119390
stays_in_week_nights            119390
adults                          119390
children                        119386
babies                          119390
meal                            119390
country                         118902
```

```
market_segment                  119390
distribution_channel            119390
is_repeated_guest               119390
previous_cancellations          119390
previous_bookings_not_canceled  119390
reserved_room_type              119390
assigned_room_type              119390
booking_changes                 119390
deposit_type                    119390
agent                           103050
company                           6797
days_in_waiting_list            119390
customer_type                   119390
adr                             119390
required_car_parking_spaces     119390
total_of_special_requests       119390
reservation_status              119390
reservation_status_date         119390
dtype: int64
```

```
df.isnull()
```

```
        hotel  is_canceled  lead_time  arrival_date_year
arrival_date_month  \
0       False        False      False              False
False
1       False        False      False              False
False
2       False        False      False              False
False
3       False        False      False              False
False
4       False        False      False              False
False
...       ...          ...        ...                ...
...
119385  False        False      False              False
False
119386  False        False      False              False
False
119387  False        False      False              False
False
119388  False        False      False              False
False
119389  False        False      False              False
False

        arrival_date_week_number  arrival_date_day_of_month  \
0                          False                      False
1                          False                      False
```

```
2                            False                         False
3                            False                         False
4                            False                         False
...                            ...                           ...
119385                       False                         False
119386                       False                         False
119387                       False                         False
119388                       False                         False
119389                       False                         False

        stays_in_weekend_nights  stays_in_week_nights  adults  ...  \
0                         False                False   False  ...
1                         False                False   False  ...
2                         False                False   False  ...
3                         False                False   False  ...
4                         False                False   False  ...
...                         ...                  ...     ...  ...
119385                    False                False   False  ...
119386                    False                False   False  ...
119387                    False                False   False  ...
119388                    False                False   False  ...
119389                    False                False   False  ...

        deposit_type  agent  company  days_in_waiting_list
customer_type  \
0              False   True     True                 False
False
1              False   True     True                 False
False
2              False   True     True                 False
False
3              False  False     True                 False
False
4              False  False     True                 False
False
...              ...    ...      ...                   ...
...
119385         False  False     True                 False
False
119386         False  False     True                 False
False
119387         False  False     True                 False
False
119388         False  False     True                 False
False
119389         False  False     True                 False
False

          adr  required_car_parking_spaces  total_of_special_requests
\
```

```
0       False                           False                           False

1       False                           False                           False

2       False                           False                           False

3       False                           False                           False

4       False                           False                           False

...       ...                             ...                             ...

119385  False                           False                           False

119386  False                           False                           False

119387  False                           False                           False

119388  False                           False                           False

119389  False                           False                           False
```

```
        reservation_status  reservation_status_date
0                    False                    False
1                    False                    False
2                    False                    False
3                    False                    False
4                    False                    False
...                    ...                      ...
119385               False                    False
119386               False                    False
119387               False                    False
119388               False                    False
119389               False                    False

[119390 rows x 32 columns]
```

```
df.isnull().sum()
```

```
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
```

```
df.duplicated()
```

```
0         False
1         False
2         False
3         False
4         False
          ...
119385    False
119386    False
119387    False
119388    False
119389    False
Length: 119390, dtype: bool
```

Preprocessing

Data cleaning

```
df.fillna(0,inplace = True) # removing null values
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
```

```
2       Resort Hotel            0          7               2015
3       Resort Hotel            0         13               2015
4       Resort Hotel            0         14               2015
...              ...          ...        ...                ...
119385    City Hotel            0         23               2017
119386    City Hotel            0        102               2017
119387    City Hotel            0         34               2017
119388    City Hotel            0        109               2017
119389    City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
...                   ...                       ...
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                       0
1                               1                       0
2                               1                       0
3                               1                       0
4                               1                       0
...                           ...                     ...
119385                         30                       2
119386                         31                       2
119387                         31                       2
119388                         31                       2
119389                         29                       2

        stays_in_week_nights  adults  ...  deposit_type  agent company
\
0                          0       2  ...    No Deposit    0.0     0.0

1                          0       2  ...    No Deposit    0.0     0.0

2                          1       1  ...    No Deposit    0.0     0.0

3                          1       1  ...    No Deposit  304.0     0.0

4                          2       2  ...    No Deposit  240.0     0.0

...                      ...     ...  ...           ...    ...     ...
```

```
119385                        5       2  ...    No Deposit  394.0     0.0

119386                        5       3  ...    No Deposit    9.0     0.0

119387                        5       2  ...    No Deposit    9.0     0.0

119388                        5       2  ...    No Deposit   89.0     0.0

119389                        7       2  ...    No Deposit    9.0     0.0


       days_in_waiting_list customer_type     adr  \
0                         0     Transient    0.00
1                         0     Transient    0.00
2                         0     Transient   75.00
3                         0     Transient   75.00
4                         0     Transient   98.00
...                     ...           ...     ...
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     Transient  151.20

        required_car_parking_spaces  total_of_special_requests  \
0                                 0                          0
1                                 0                          0
2                                 0                          0
3                                 0                          0
4                                 0                          1
...                             ...                        ...
119385                            0                          0
119386                            0                          2
119387                            0                          4
119388                            0                          0
119389                            0                          2

        reservation_status reservation_status_date
0                Check-Out              2015-07-01
1                Check-Out              2015-07-01
2                Check-Out              2015-07-02
3                Check-Out              2015-07-02
4                Check-Out              2015-07-03
...                    ...                     ...
119385           Check-Out              2017-09-06
119386           Check-Out              2017-09-07
119387           Check-Out              2017-09-07
119388           Check-Out              2017-09-07
119389           Check-Out              2017-09-07
```

```
[119390 rows x 32 columns]
```

```
df.drop(df[df['meal'] == 'Undefined'].index, inplace = True)
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

        arrival_date_month  arrival_date_week_number  \
0                     July                        27
1                     July                        27
2                     July                        27
3                     July                        27
4                     July                        27
...                    ...                       ...
119385              August                        35
119386              August                        35
119387              August                        35
119388              August                        35
119389              August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                       0
1                               1                       0
2                               1                       0
3                               1                       0
4                               1                       0
...                           ...                     ...
119385                         30                       2
119386                         31                       2
119387                         31                       2
119388                         31                       2
119389                         29                       2

        stays_in_week_nights  adults  ...  deposit_type  agent company
\
0                          0       2  ...    No Deposit    0.0     0.0

1                          0       2  ...    No Deposit    0.0     0.0
```

```
2                            1       1  ...    No Deposit    0.0     0.0
3                            1       1  ...    No Deposit  304.0     0.0
4                            2       2  ...    No Deposit  240.0     0.0
...                        ...     ...  ...           ...    ...     ...
119385                       5       2  ...    No Deposit  394.0     0.0
119386                       5       3  ...    No Deposit    9.0     0.0
119387                       5       2  ...    No Deposit    9.0     0.0
119388                       5       2  ...    No Deposit   89.0     0.0
119389                       7       2  ...    No Deposit    9.0     0.0

       days_in_waiting_list customer_type     adr  \
0                         0     Transient    0.00
1                         0     Transient    0.00
2                         0     Transient   75.00
3                         0     Transient   75.00
4                         0     Transient   98.00
...                     ...           ...     ...
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     Transient  151.20

        required_car_parking_spaces  total_of_special_requests  \
0                                 0                          0
1                                 0                          0
2                                 0                          0
3                                 0                          0
4                                 0                          1
...                             ...                        ...
119385                            0                          0
119386                            0                          2
119387                            0                          4
119388                            0                          0
119389                            0                          2

        reservation_status reservation_status_date
0                Check-Out              2015-07-01
1                Check-Out              2015-07-01
2                Check-Out              2015-07-02
```

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

[118221 rows x 32 columns]
```

```
df.drop(df[df['market_segment'] == 'Undefined'].index, inplace = True)
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
...                   ...                       ...
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                       0
1                               1                       0
2                               1                       0
3                               1                       0
4                               1                       0
...                           ...                     ...
119385                         30                       2
119386                         31                       2
119387                         31                       2
```

```
119388                          31                        2
119389                          29                        2

        stays_in_week_nights  adults  ...  deposit_type  agent company
\
0                          0       2  ...    No Deposit    0.0     0.0

1                          0       2  ...    No Deposit    0.0     0.0

2                          1       1  ...    No Deposit    0.0     0.0

3                          1       1  ...    No Deposit  304.0     0.0

4                          2       2  ...    No Deposit  240.0     0.0

...                      ...     ...  ...           ...    ...     ...

119385                     5       2  ...    No Deposit  394.0     0.0

119386                     5       3  ...    No Deposit    9.0     0.0

119387                     5       2  ...    No Deposit    9.0     0.0

119388                     5       2  ...    No Deposit   89.0     0.0

119389                     7       2  ...    No Deposit    9.0     0.0


       days_in_waiting_list customer_type     adr  \
0                         0     Transient    0.00
1                         0     Transient    0.00
2                         0     Transient   75.00
3                         0     Transient   75.00
4                         0     Transient   98.00
...                     ...           ...     ...
119385                    0     Transient   96.14
119386                    0     Transient  225.43
119387                    0     Transient  157.71
119388                    0     Transient  104.40
119389                    0     Transient  151.20

        required_car_parking_spaces  total_of_special_requests  \
0                                 0                          0
1                                 0                          0
2                                 0                          0
3                                 0                          0
4                                 0                          1
...                             ...                        ...
119385                            0                          0
119386                            0                          2
119387                            0                          4
```

```
119388                              0                           0
119389                              0                           2

        reservation_status reservation_status_date
0                Check-Out              2015-07-01
1                Check-Out              2015-07-01
2                Check-Out              2015-07-02
3                Check-Out              2015-07-02
4                Check-Out              2015-07-03
...                    ...                     ...
119385           Check-Out              2017-09-06
119386           Check-Out              2017-09-07
119387           Check-Out              2017-09-07
119388           Check-Out              2017-09-07
119389           Check-Out              2017-09-07

[118219 rows x 32 columns]
```

```
df.isnull().sum()
```

```
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                          0
babies                            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
agent                             0
company                           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
dtype: int64
```

DImensionality reduction

```
# combining similar columns
cols =
["arrival_date_year","arrival_date_month","arrival_date_day_of_month"]
df['arrival_date'] = df[cols].apply(lambda row:
'-'.join(row.values.astype(str)), axis=1)
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

       arrival_date_month  arrival_date_week_number  \
0                    July                        27
1                    July                        27
2                    July                        27
3                    July                        27
4                    July                        27
...                   ...                       ...
119385             August                        35
119386             August                        35
119387             August                        35
119388             August                        35
119389             August                        35

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                               1                       0
1                               1                       0
2                               1                       0
3                               1                       0
4                               1                       0
...                           ...                     ...
119385                         30                       2
119386                         31                       2
119387                         31                       2
```

```
119388                          31                          2
119389                          29                          2

        stays_in_week_nights  adults  ...  agent  company  \
0                          0       2  ...    0.0      0.0
1                          0       2  ...    0.0      0.0
2                          1       1  ...    0.0      0.0
3                          1       1  ...  304.0      0.0
4                          2       2  ...  240.0      0.0
...                      ...     ...  ...    ...      ...
119385                     5       2  ...  394.0      0.0
119386                     5       3  ...    9.0      0.0
119387                     5       2  ...    9.0      0.0
119388                     5       2  ...   89.0      0.0
119389                     7       2  ...    9.0      0.0

       days_in_waiting_list customer_type     adr
required_car_parking_spaces  \
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

        total_of_special_requests  reservation_status  \
0                               0           Check-Out
1                               0           Check-Out
2                               0           Check-Out
3                               0           Check-Out
4                               1           Check-Out
...                           ...                 ...
119385                          0           Check-Out
119386                          2           Check-Out
```

```
119387                           4           Check-Out
119388                           0           Check-Out
119389                           2           Check-Out

        reservation_status_date    arrival_date
0                    2015-07-01     2015-July-1
1                    2015-07-01     2015-July-1
2                    2015-07-02     2015-July-1
3                    2015-07-02     2015-July-1
4                    2015-07-03     2015-July-1
...                         ...             ...
119385               2017-09-06  2017-August-30
119386               2017-09-07  2017-August-31
119387               2017-09-07  2017-August-31
119388               2017-09-07  2017-August-31
119389               2017-09-07  2017-August-29

[118219 rows x 33 columns]
```

```
# removing same columns
df.drop(columns=['arrival_date_month','arrival_date_week_number','arri
val_date_day_of_month'],axis=1,inplace = True)
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

        stays_in_weekend_nights  stays_in_week_nights  adults
children  \
0                             0                     0       2
0.0
1                             0                     0       2
0.0
2                             0                     1       1
0.0
3                             0                     1       1
0.0
4                             0                     2       2
0.0
...                         ...                   ...     ...       ..
```

```
.
119385                          2                       5       2
0.0
119386                          2                       5       3
0.0
119387                          2                       5       2
0.0
119388                          2                       5       2
0.0
119389                          2                       7       2
0.0

        babies meal  ...  agent company days_in_waiting_list
customer_type  \
0            0   BB  ...    0.0     0.0                    0
Transient
1            0   BB  ...    0.0     0.0                    0
Transient
2            0   BB  ...    0.0     0.0                    0
Transient
3            0   BB  ...  304.0     0.0                    0
Transient
4            0   BB  ...  240.0     0.0                    0
Transient
...        ...  ...  ...    ...     ...                  ...
...
119385       0   BB  ...  394.0     0.0                    0
Transient
119386       0   BB  ...    9.0     0.0                    0
Transient
119387       0   BB  ...    9.0     0.0                    0
Transient
119388       0   BB  ...   89.0     0.0                    0
Transient
119389       0   HB  ...    9.0     0.0                    0
Transient

            adr  required_car_parking_spaces total_of_special_requests
\
0          0.00                            0                         0

1          0.00                            0                         0

2         75.00                            0                         0

3         75.00                            0                         0

4         98.00                            0                         1

...         ...                          ...                       ...
```

```
119385   96.14                                0                           0

119386  225.43                                0                           2

119387  157.71                                0                           4

119388  104.40                                0                           0

119389  151.20                                0                           2


       reservation_status  reservation_status_date    arrival_date
0               Check-Out               2015-07-01     2015-July-1
1               Check-Out               2015-07-01     2015-July-1
2               Check-Out               2015-07-02     2015-July-1
3               Check-Out               2015-07-02     2015-July-1
4               Check-Out               2015-07-03     2015-July-1
...                   ...                      ...             ...
119385          Check-Out               2017-09-06  2017-August-30
119386          Check-Out               2017-09-07  2017-August-31
119387          Check-Out               2017-09-07  2017-August-31
119388          Check-Out               2017-09-07  2017-August-31
119389          Check-Out               2017-09-07  2017-August-29

[118219 rows x 30 columns]
```

Data transformation

Label encoding

```
label_encoder = preprocessing.LabelEncoder()
print("Distinct hotel values: ",df['hotel'].unique())
df1 =
pd.DataFrame(label_encoder.fit_transform(df['hotel']),columns=['Hotel'
])
print("\nCoded hotel values (respectively): ",df['hotel'].unique())
print("\nDataset:\n")
df1
```

```
Distinct hotel values:  ['Resort Hotel' 'City Hotel']

Coded hotel values (respectively):  ['Resort Hotel' 'City Hotel']

Dataset:


        Hotel
0           1
1           1
```

```
2            1
3            1
4            1
...        ...
118214       0
118215       0
118216       0
118217       0
118218       0

[118219 rows x 1 columns]
```

One hot coding

```
enc = preprocessing.OneHotEncoder(handle_unknown='ignore')
print("Distinct reservation status values:
",df['reservation_status'].unique())
enc_df =
pd.DataFrame(enc.fit_transform(df[['reservation_status']]).toarray())
enc_df
```

```
Distinct reservation status values:  ['Check-Out' 'Canceled' 'No-
Show']
```

```
           0    1    2
0        0.0  1.0  0.0
1        0.0  1.0  0.0
2        0.0  1.0  0.0
3        0.0  1.0  0.0
4        0.0  1.0  0.0
...      ...  ...  ...
118214   0.0  1.0  0.0
118215   0.0  1.0  0.0
118216   0.0  1.0  0.0
118217   0.0  1.0  0.0
118218   0.0  1.0  0.0

[118219 rows x 3 columns]
```

Data discretization

```
# categorising lead time
m1 = min(df['lead_time'])
m2 = max(df['lead_time'])
bins = np.linspace(m1,m2,3)
names = ['low','high']
df['Lead_time_binned'] =
pd.cut(df['lead_time'],bins,labels=names,include_lowest = True)
df
```

```
                hotel  is_canceled  lead_time  arrival_date_year  \
0        Resort Hotel            0        342               2015
1        Resort Hotel            0        737               2015
2        Resort Hotel            0          7               2015
3        Resort Hotel            0         13               2015
4        Resort Hotel            0         14               2015
...               ...          ...        ...                ...
119385     City Hotel            0         23               2017
119386     City Hotel            0        102               2017
119387     City Hotel            0         34               2017
119388     City Hotel            0        109               2017
119389     City Hotel            0        205               2017

        stays_in_weekend_nights  stays_in_week_nights  adults
children  \
0                             0                     0       2
0.0
1                             0                     0       2
0.0
2                             0                     1       1
0.0
3                             0                     1       1
0.0
4                             0                     2       2
0.0
...                         ...                   ...     ...       ..
.
119385                        2                     5       2
0.0
119386                        2                     5       3
0.0
119387                        2                     5       2
0.0
119388                        2                     5       2
0.0
119389                        2                     7       2
0.0

        babies meal  ... company days_in_waiting_list customer_type
adr  \
0            0   BB  ...     0.0                    0     Transient
0.00
1            0   BB  ...     0.0                    0     Transient
0.00
2            0   BB  ...     0.0                    0     Transient
75.00
3            0   BB  ...     0.0                    0     Transient
75.00
4            0   BB  ...     0.0                    0     Transient
98.00
```

```
...        ...  ...  ...     ...                   ...           ...
...
119385       0   BB  ...     0.0                     0     Transient
96.14
119386       0   BB  ...     0.0                     0     Transient
225.43
119387       0   BB  ...     0.0                     0     Transient
157.71
119388       0   BB  ...     0.0                     0     Transient
104.40
119389       0   HB  ...     0.0                     0     Transient
151.20

        required_car_parking_spaces  total_of_special_requests  \
0                                 0                          0
1                                 0                          0
2                                 0                          0
3                                 0                          0
4                                 0                          1
...                             ...                        ...
119385                            0                          0
119386                            0                          2
119387                            0                          4
119388                            0                          0
119389                            0                          2

       reservation_status reservation_status_date    arrival_date  \
0               Check-Out              2015-07-01     2015-July-1
1               Check-Out              2015-07-01     2015-July-1
2               Check-Out              2015-07-02     2015-July-1
3               Check-Out              2015-07-02     2015-July-1
4               Check-Out              2015-07-03     2015-July-1
...                   ...                     ...             ...
119385          Check-Out              2017-09-06  2017-August-30
119386          Check-Out              2017-09-07  2017-August-31
119387          Check-Out              2017-09-07  2017-August-31
119388          Check-Out              2017-09-07  2017-August-31
119389          Check-Out              2017-09-07  2017-August-29

       Lead_time_binned
0                   low
1                  high
2                   low
3                   low
4                   low
...                 ...
119385              low
119386              low
119387              low
119388              low
```

```
119389                low

[118219 rows x 31 columns]
```

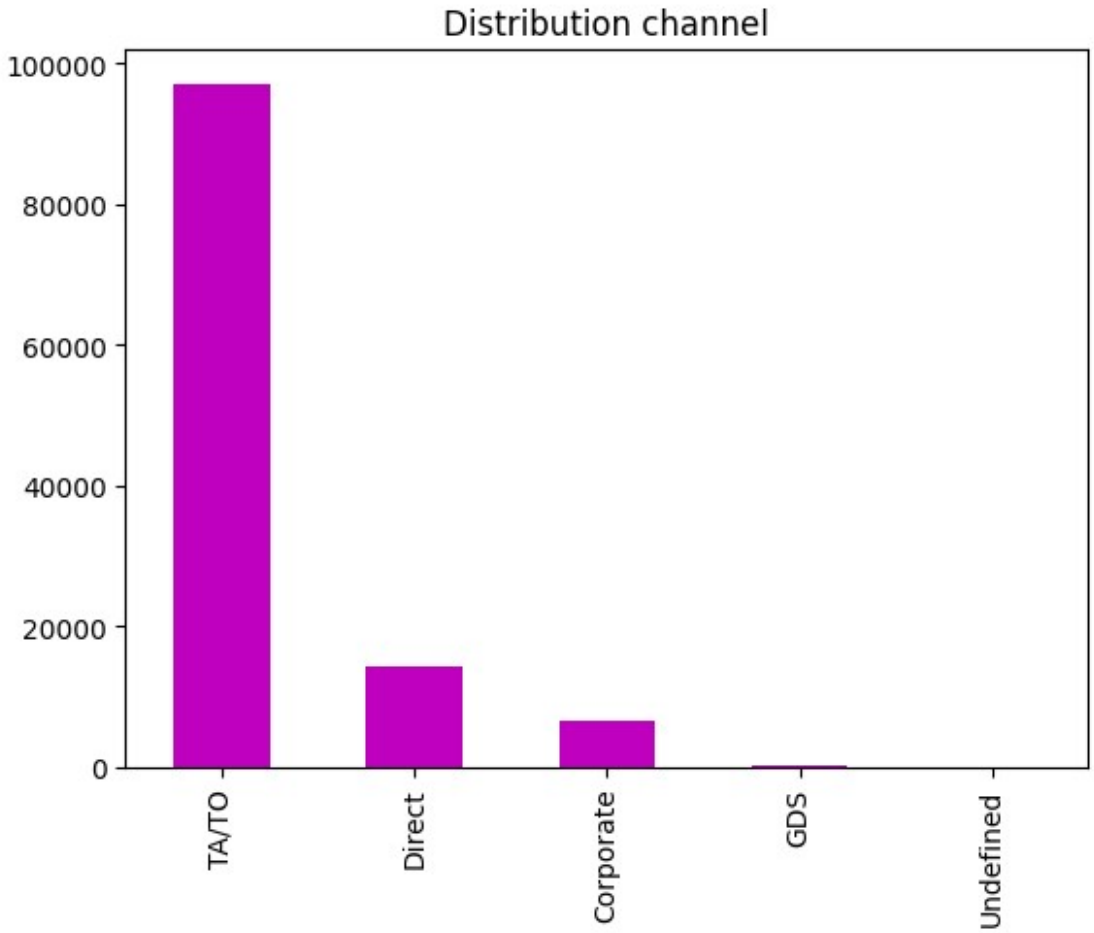
Data visualization

Exploratory Data Analysis (EDA)

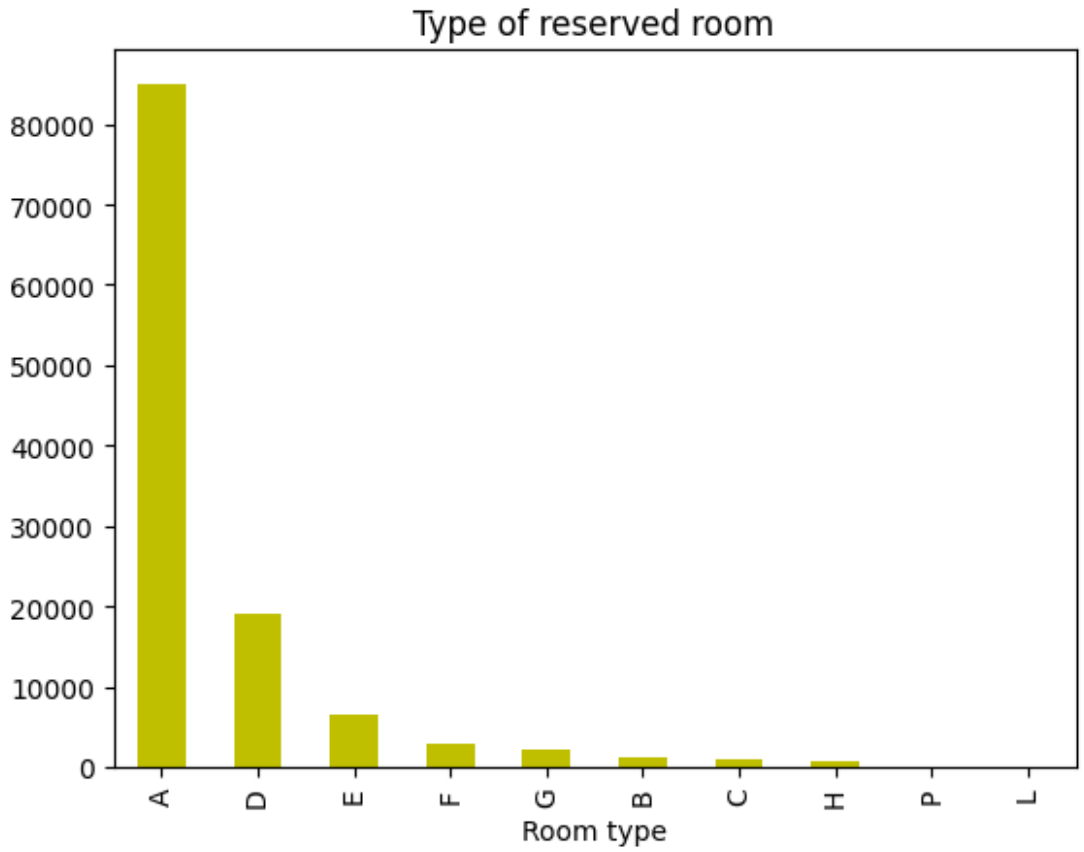
Univariate visualization

Bar chart

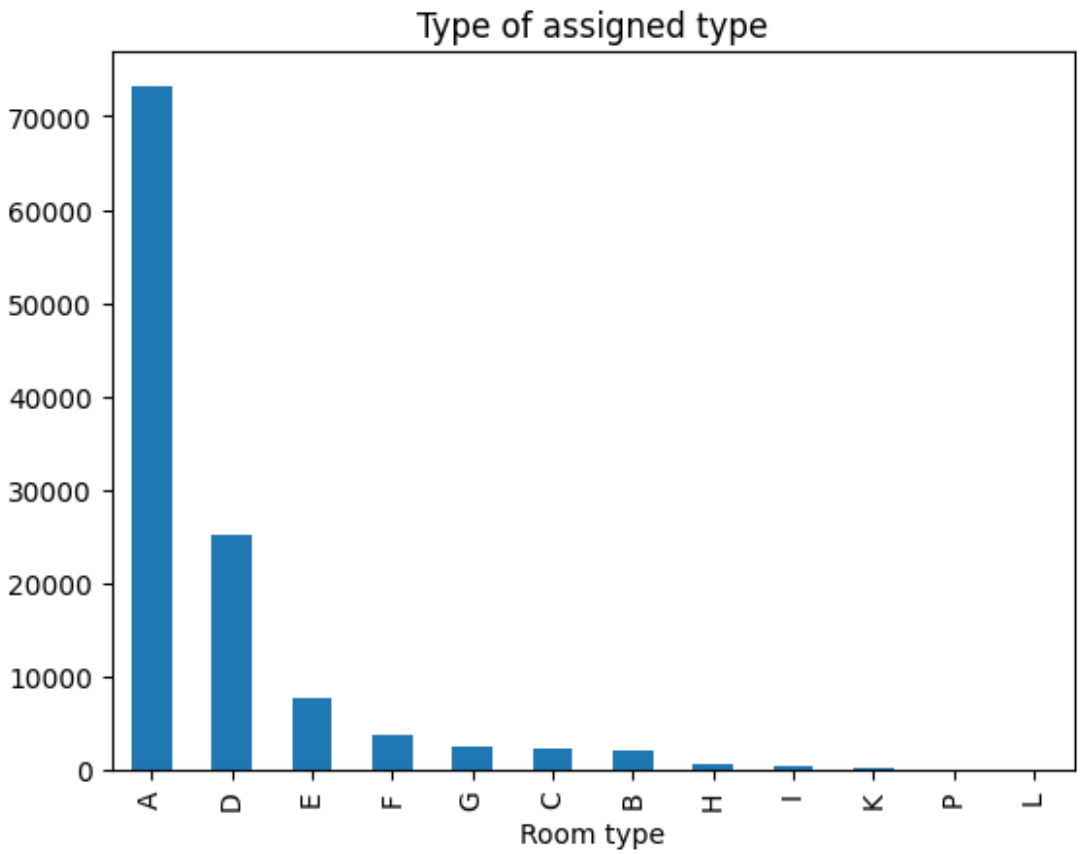
```
df['distribution_channel'].value_counts().plot(kind='bar',color='m')
plt.title("Distribution channel")
plt.show()
```



```
df['reserved_room_type'].value_counts().plot(kind='bar',color='y')
plt.title("Type of reserved room")
plt.xlabel("Room type")
plt.show()
```

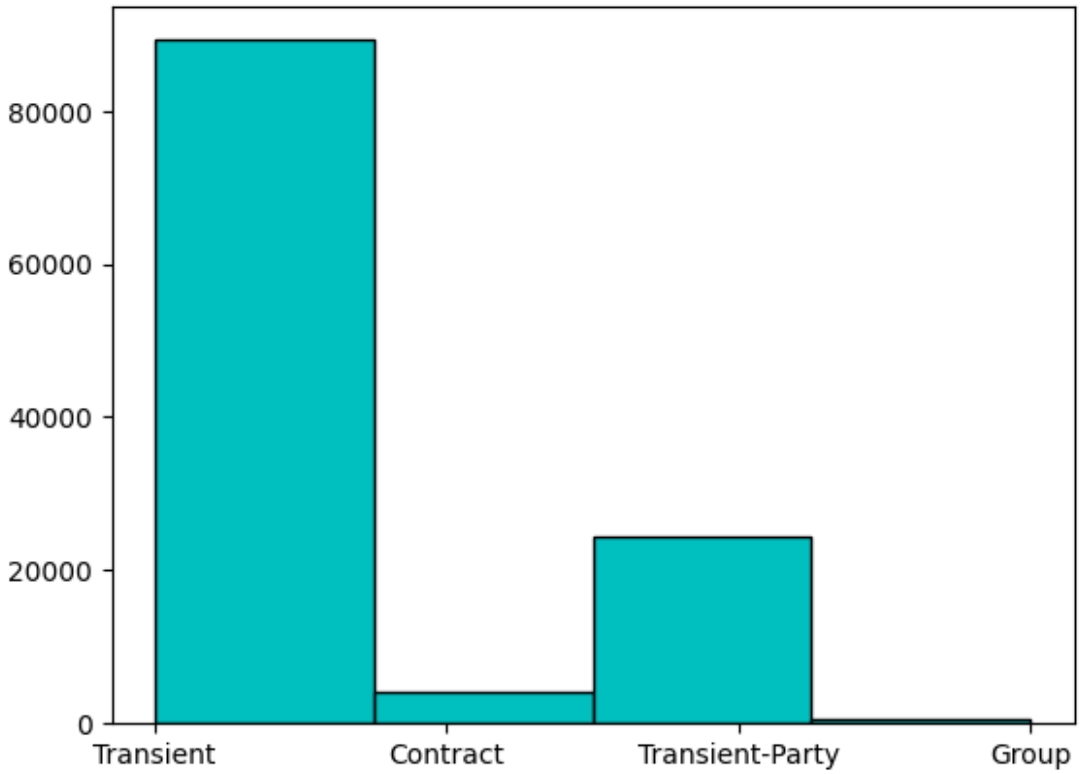



```
df['assigned_room_type'].value_counts().plot(kind='bar')
plt.title("Type of assigned type")
plt.xlabel("Room type")
plt.show()
```

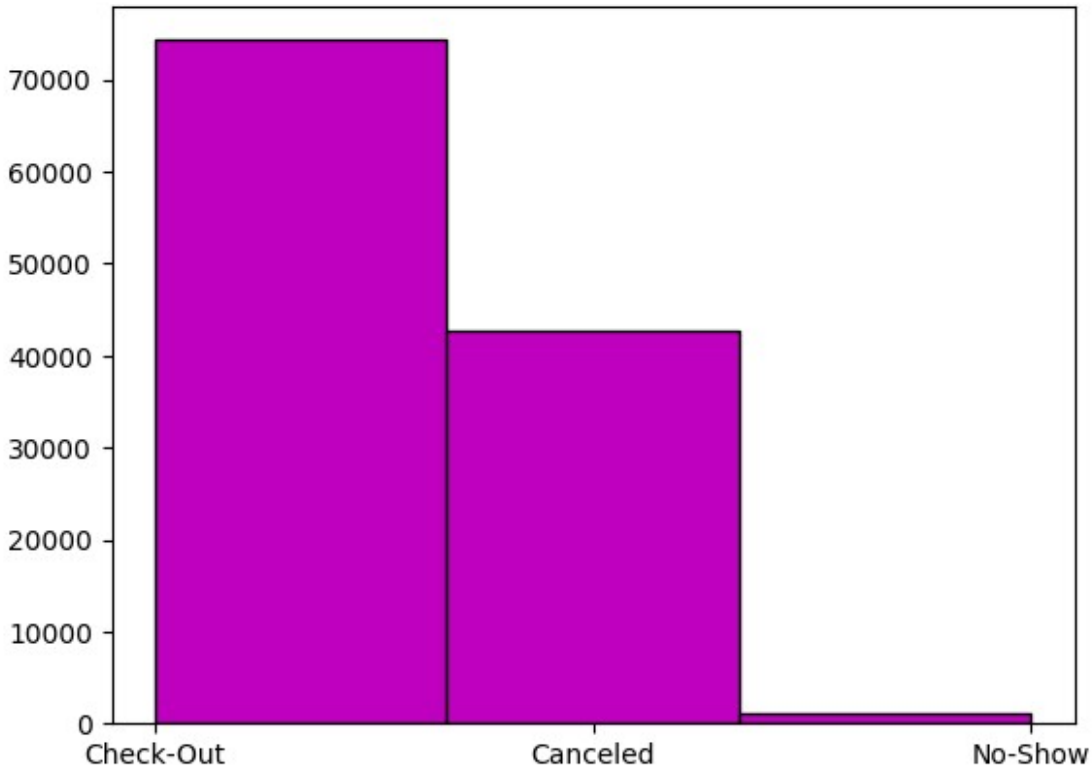


Histogram

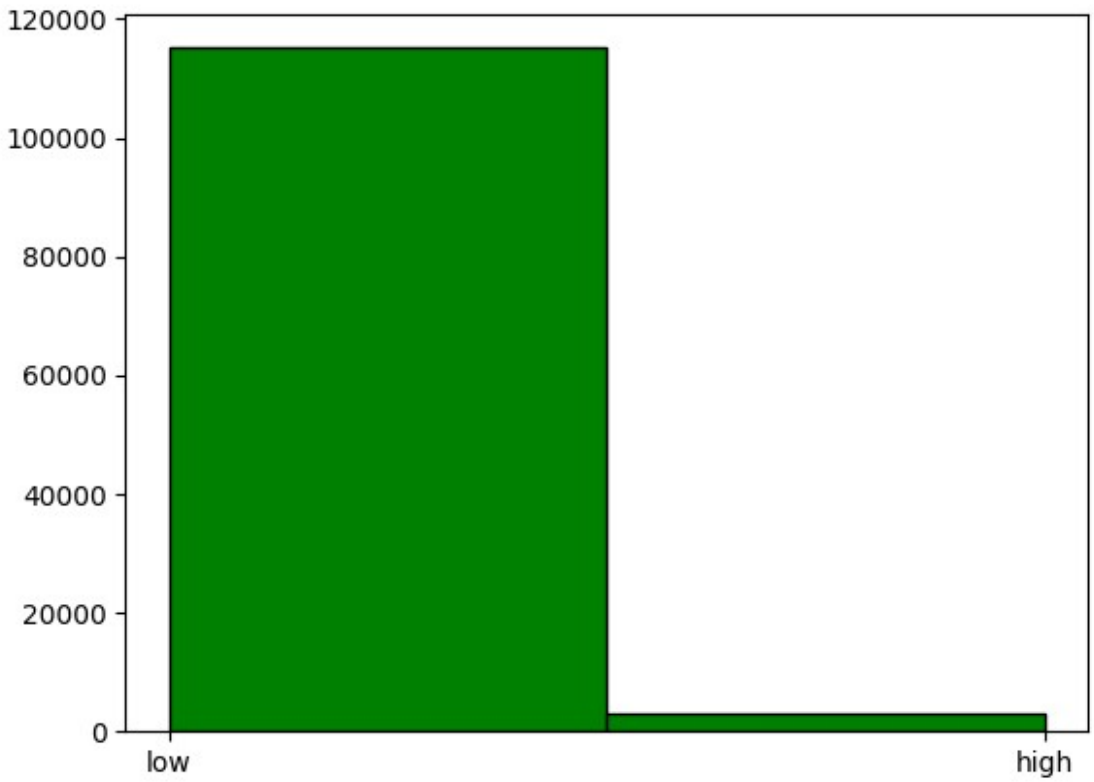
```
plt.hist(df['customer_type'],bins=4,edgecolor="k",color='c')
plt.show()
```



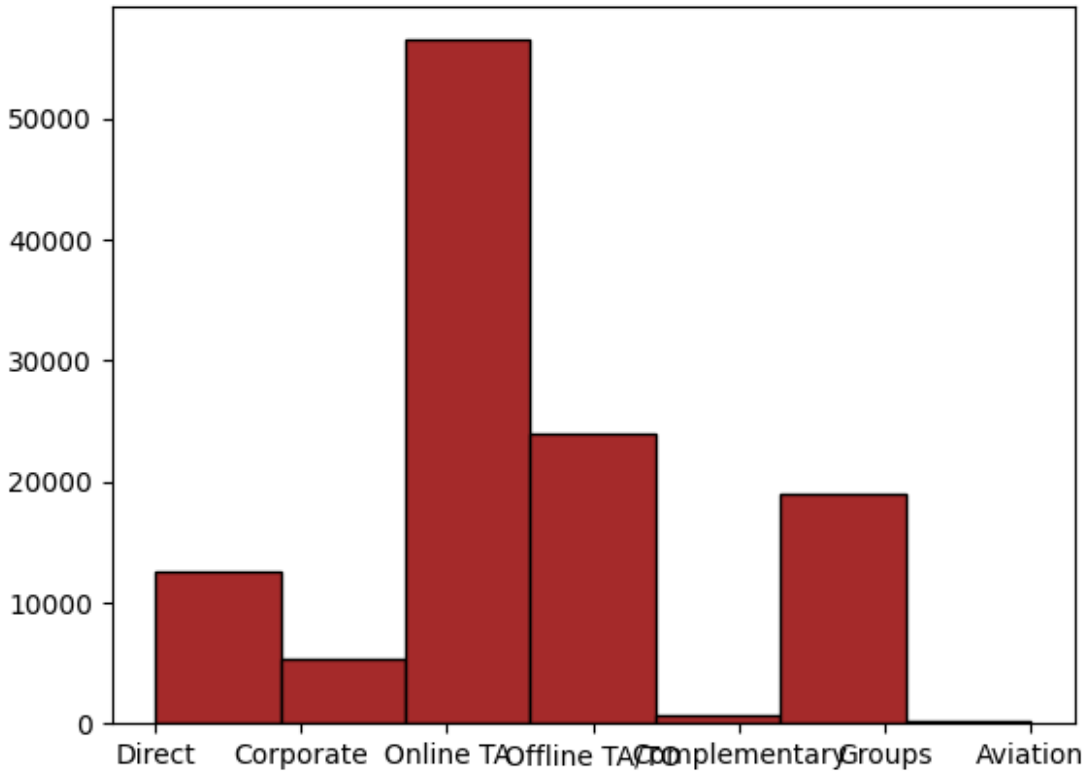

```
plt.hist(df['reservation_status'],edgecolor="k",bins=3,color='m')
plt.show()
```




```
plt.hist(df['Lead_time_binned'],bins=2,edgecolor="k",color='g')
plt.show()
```

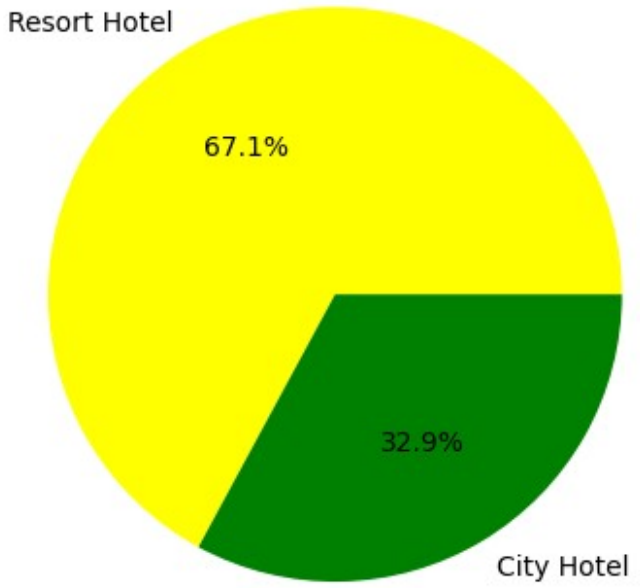



```
plt.hist(df['market_segment'],bins=7,edgecolor="k",color='brown')
plt.show()
```

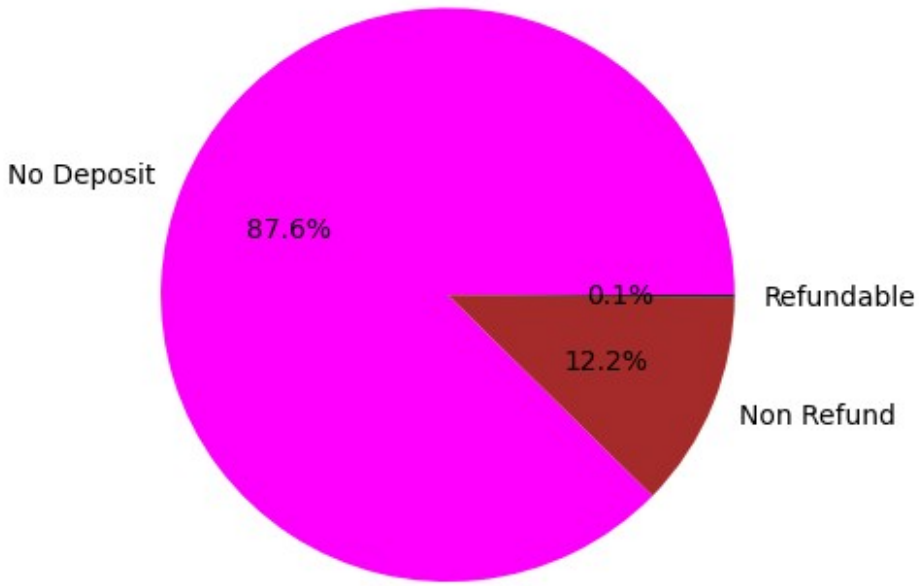


Pie chart

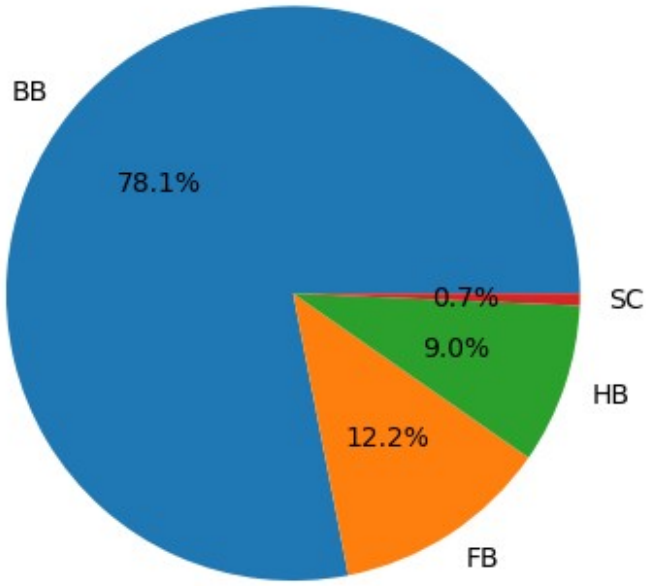
```
plt.pie(df.hotel.value_counts(), autopct = '%.1f%%', labels = ["Resort
Hotel","City Hotel"],colors=['yellow','green'])
plt.show()
```




```
plt.pie(df.deposit_type.value_counts(), autopct = '%.1f%%', labels =
["No Deposit", "Non Refund",
"Refundable"],colors=['magenta','brown','black'])
plt.show()
```



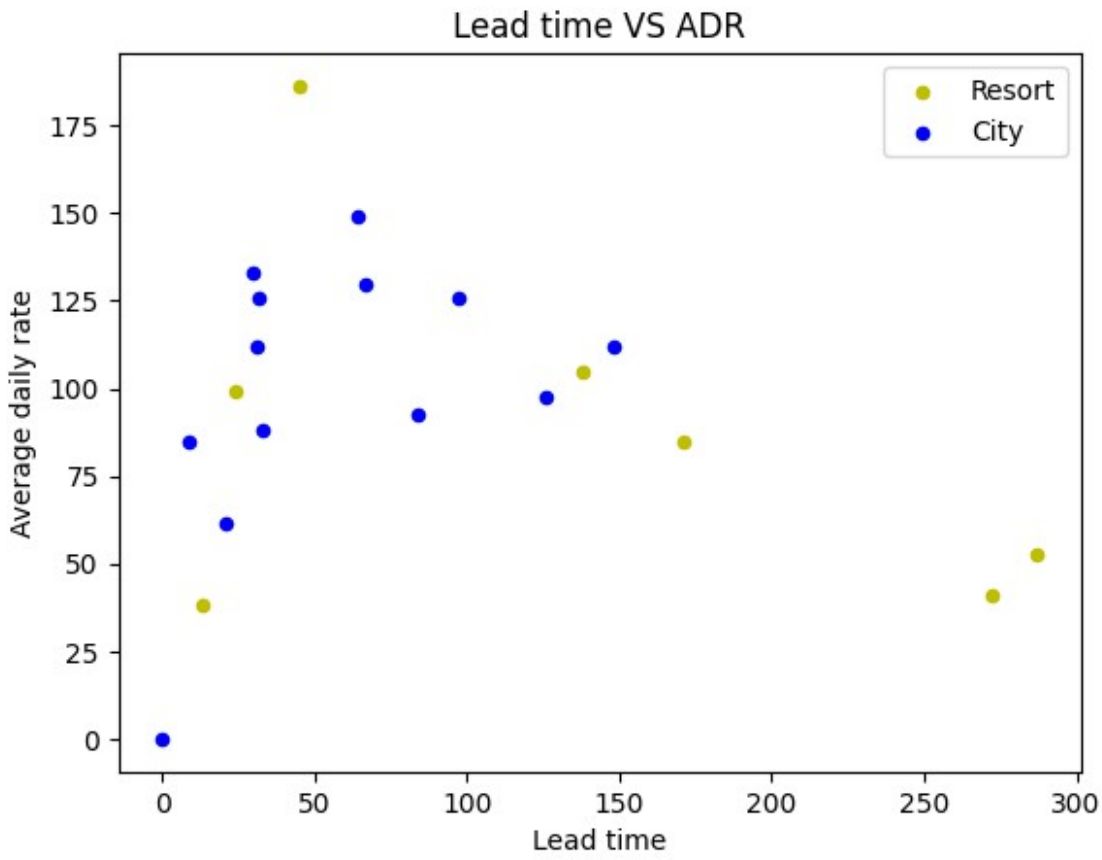
```
plt.pie(df.meal.value_counts(), autopct = '%.1f%%', labels =
["BB","FB","HB","SC"])
plt.show()
```



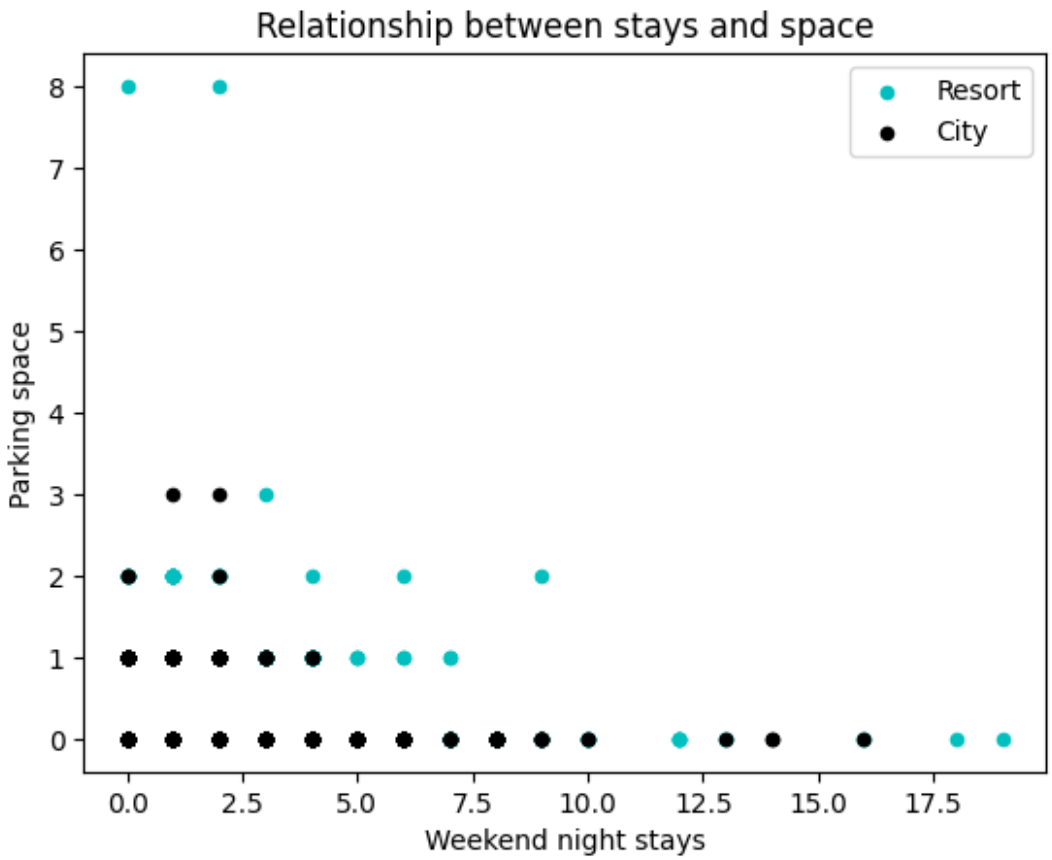
Bivariate visualization

Scatter plot

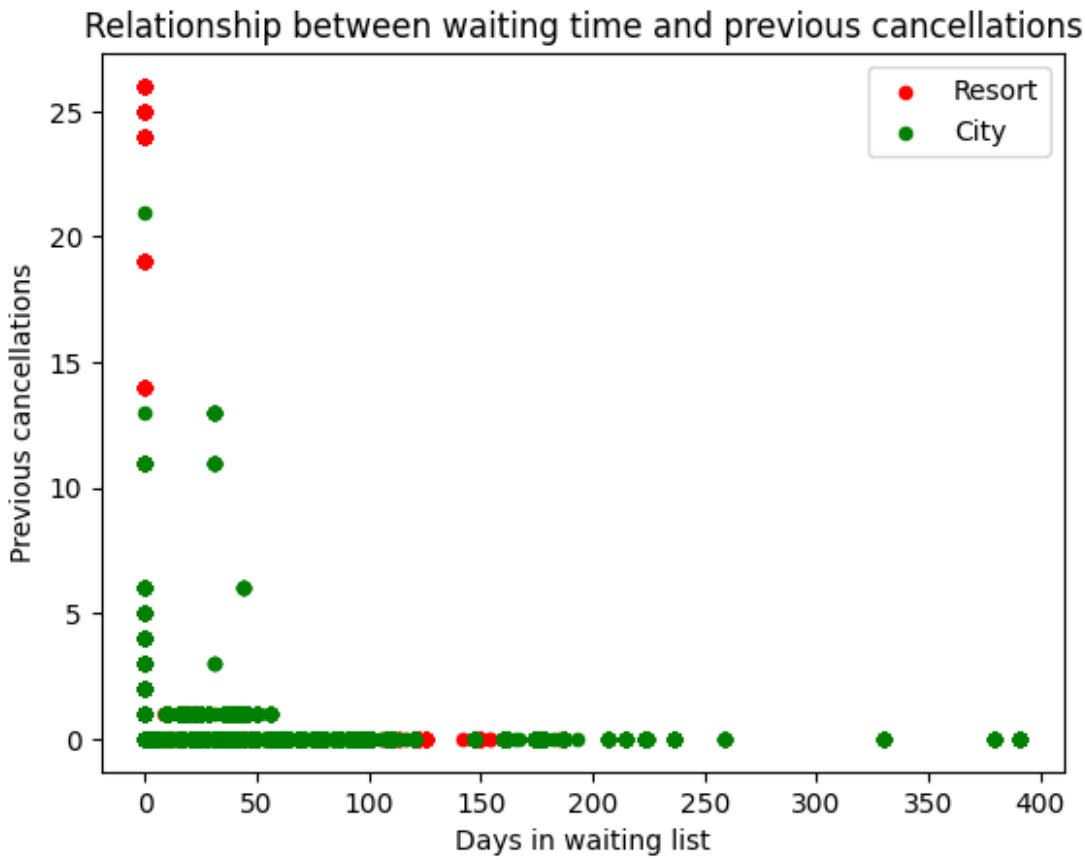
```
df1 = df.sample(20)
fig = df1[df1.hotel=='Resort
Hotel'].plot.scatter(x='lead_time',y='adr',color='y', label='Resort')
df1[df1.hotel=='City
Hotel'].plot.scatter(x='lead_time',y='adr',color='b',
label='City',ax=fig)
fig.set_xlabel("Lead time")
fig.set_ylabel("Average daily rate")
fig.set_title("Lead time VS ADR")
fig=plt.gcf()
plt.show()
```




```
fig = df[df.hotel=='Resort
Hotel'].plot.scatter(x='stays_in_weekend_nights',y='required_car_parki
ng_spaces',color='c', label='Resort')
df[df.hotel=='City
Hotel'].plot.scatter(x='stays_in_weekend_nights',y='required_car_parki
ng_spaces',color='k', label='City',ax=fig)
fig.set_xlabel("Weekend night stays")
fig.set_ylabel("Parking space")
fig.set_title("Relationship between stays and space")
fig=plt.gcf()
plt.show()
```

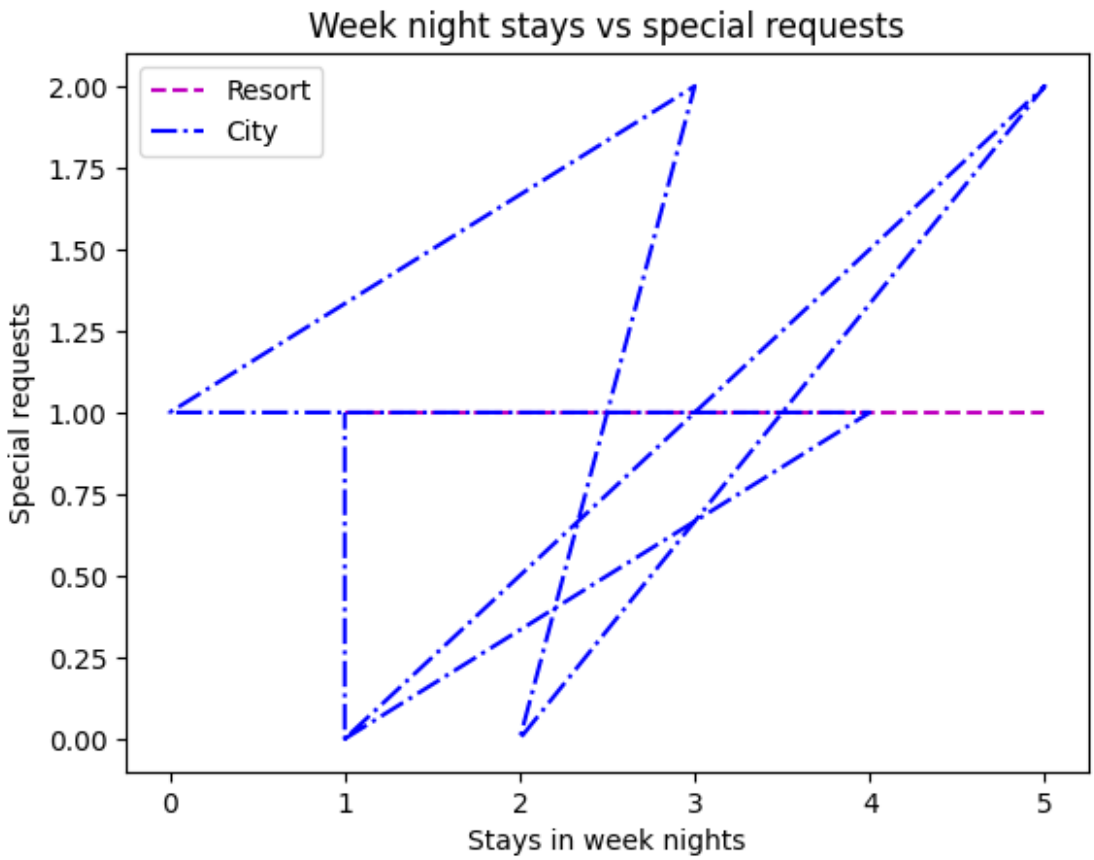



```
fig = df[df.hotel=='Resort
Hotel'].plot.scatter(x='days_in_waiting_list',y='previous_cancellation
s',color='r', label='Resort')
df[df.hotel=='City
Hotel'].plot.scatter(x='days_in_waiting_list',y='previous_cancellation
s',color='g', label='City',ax=fig)
fig.set_xlabel("Days in waiting list")
fig.set_ylabel("Previous cancellations")
fig.set_title("Relationship between waiting time and previous
cancellations")
fig=plt.gcf()
plt.show()
```

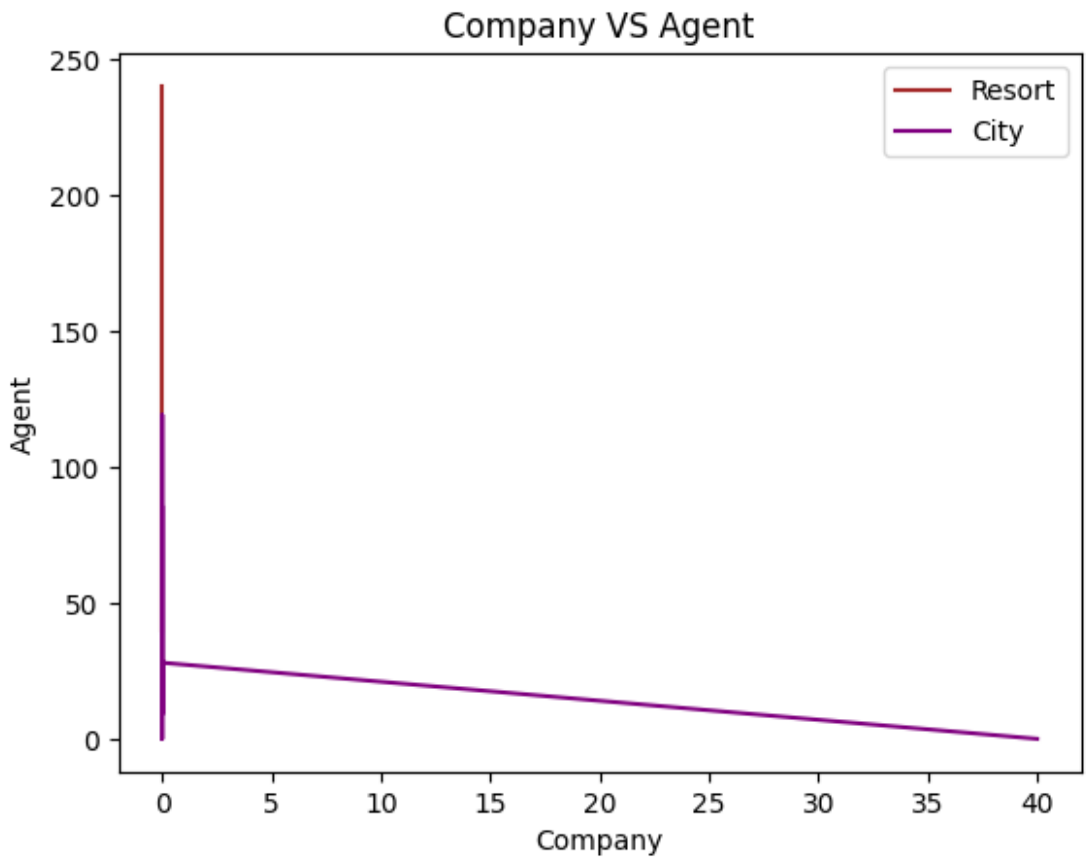


Line plot

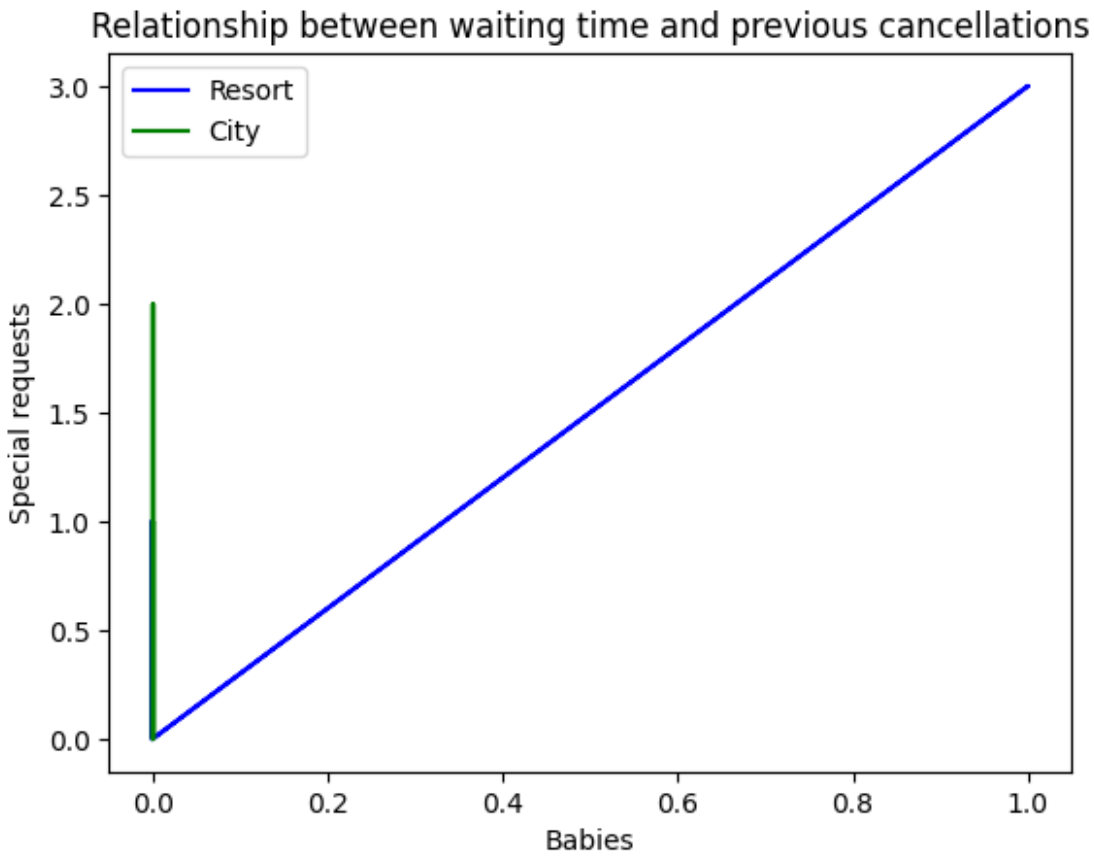
```
df1 = df.sample(10)
fig = df1[df1.hotel=='Resort
Hotel'].plot.line(x='stays_in_week_nights',y='total_of_special_request
s',color='m',style="--", label='Resort')
df1[df1.hotel=='City
Hotel'].plot.line(x='stays_in_week_nights',y='total_of_special_request
s',color='b',style="-.", label='City',ax=fig)
fig.set_xlabel("Stays in week nights")
fig.set_ylabel("Special requests")
fig.set_title("Week night stays vs special requests")
fig=plt.gcf()
plt.show()
```




```
df1 = df.sample(10)
fig = df1[df1.hotel=='Resort
Hotel'].plot.line(x='company',y='agent',color='brown', label='Resort')
df1[df1.hotel=='City
Hotel'].plot.line(x='company',y='agent',color='purple',
label='City',ax=fig)
fig.set_xlabel("Company")
fig.set_ylabel("Agent")
fig.set_title("Company VS Agent")
fig=plt.gcf()
plt.show()
```




```
df1 = df.sample(50)
fig = df1[df1.hotel=='Resort
Hotel'].plot.line(x='babies',y='total_of_special_requests',color='b',
label='Resort')
df1[df1.hotel=='City
Hotel'].plot.line(x='babies',y='total_of_special_requests',color='g',
label='City',ax=fig)
fig.set_xlabel("Babies")
fig.set_ylabel("Special requests")
fig.set_title("Relationship between waiting time and previous
cancellations")
fig=plt.gcf()
plt.show()
```

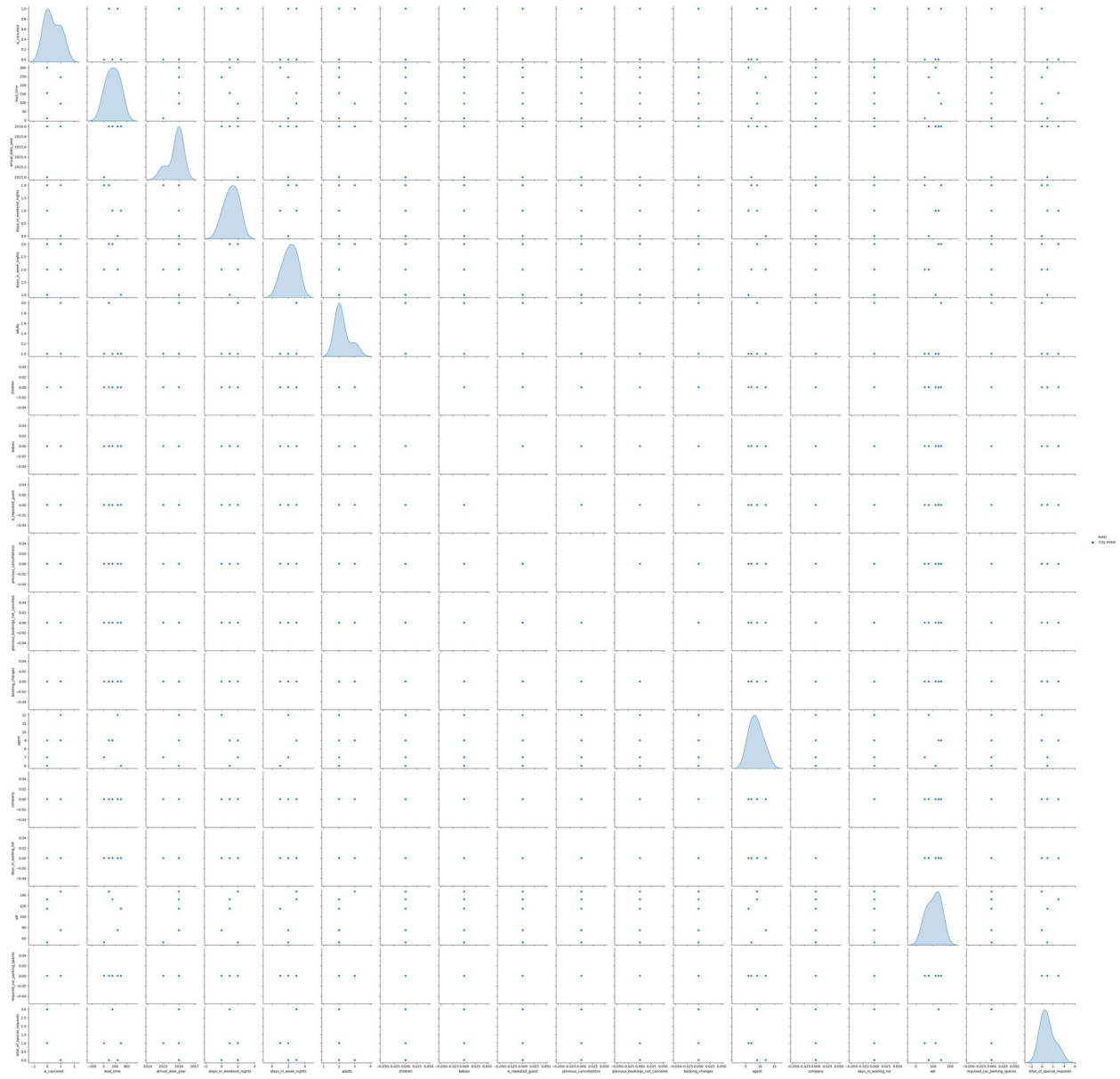


Multivariate visualization

Pair plot

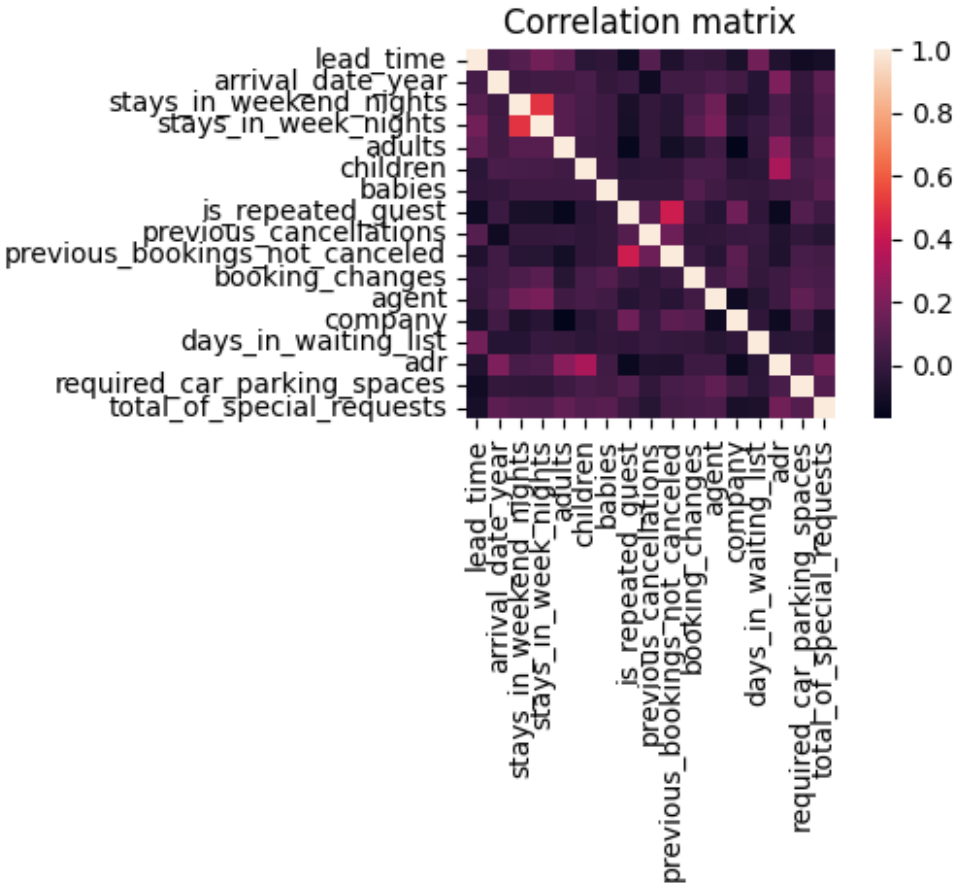
```
sns.pairplot(df.sample(5),hue='hotel')
```

```
<seaborn.axisgrid.PairGrid at 0x7dcc11613b50>
```



Heat map

```
correlation = df.select_dtypes(include=['float64','int64']).iloc[:,
1:].corr()
sns.heatmap(correlation, vmax=1, square=True)
plt.xticks(rotation=90)
plt.yticks(rotation=360)
plt.title('Correlation matrix')
plt.tight_layout()
plt.show()
```


Correlation matrix
lead_time
arrival_date_year
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
agent
company
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
1.0
0.8
0.6
0.4
0.2
0.0