Hotel booking demand

Introduction

This is the order of my work:

- First exploring
- Preprocessing
- Supervised learning
- Unsupervised learning

In [1]:
```python
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

Let's read the data set.

In [2]:
```python
try:
    hotel_data = pd.read_csv('C:/Users/Matan/Documents/Python/hotel_bookings.csv')
except: print("Something got wrong")
```

In [3]:
```python
hotel_data
```

Out[3]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
0	Resort Hotel	0	342	2015	July	27	1
1	Resort Hotel	0	737	2015	July	27	1
2	Resort Hotel	0	7	2015	July	27	1
3	Resort Hotel	0	13	2015	July	27	1
4	Resort Hotel	0	14	2015	July	27	1
...	...	...	...	...	...	...	...
119385	City Hotel	0	23	2017	August	35	30
119386	City Hotel	0	102	2017	August	35	31
119387	City Hotel	0	34	2017	August	35	31
119388	City Hotel	0	109	2017	August	35	31
119389	City Hotel	0	205	2017	August	35	29

119390 rows × 32 columns

The data is ready.

First exploring

Out[5]:
```
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [6]:
```python
hotel_data.dtypes
```

Out[6]:
```
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
```

In [7]:
```python
hotel_data['country'].drop_duplicates().count()
```

Out[7]: 177

In [8]:
```python
hotel_data['distribution_channel'].drop_duplicates().count()
```

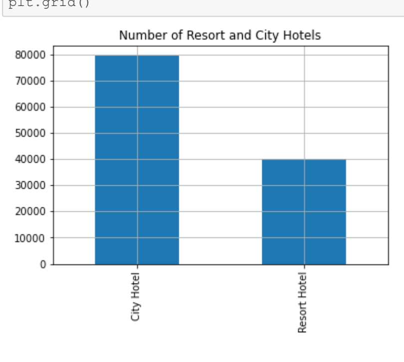
Out[8]: 5

In [9]:
```python
hotel_data['hotel'].drop_duplicates()
```

Out[9]:
```
0        Resort Hotel
40060      City Hotel
Name: hotel, dtype: object
```

There are only two types of hotels.

In [10]:
```python
hotel_data['hotel'].value_counts().plot(kind="bar", title="Number of Resort and City Hotels")
plt.grid()
```



The number of city hotels is twice as large as in resort hotel in this data set.

In [11]:
```python
hotel_kinds = hotel_data.groupby('hotel')
for name, kind in hotel_kinds:
    print("The average number of children and babies in",name, "are:", kind['children'].mean(), kind['babies'].mean())
```

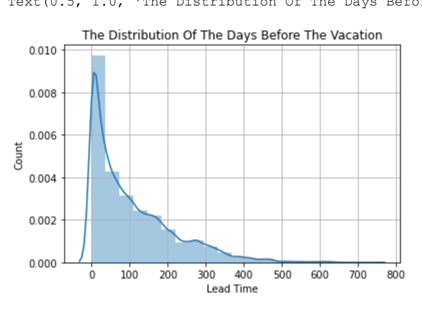
```
The average number of children and babies in City Hotel are: 0.09136979048483473 0.0049413840917685616
The average number of children and babies in Resort Hotel are: 0.12868197703444834 0.013904143784323515
```

We can learn from this result that children and babies are more likely to go to resort hotels.

Now let's check the distribution of the number of days that elapsed between the entering date of the booking and the arrival date.

In [12]:
```python
sns.distplot(hotel_data['lead_time'], bins=20)
plt.grid()
plt.xlabel('Lead Time')
plt.ylabel("Count")
plt.title("The Distribution Of The Days Before The Vacation")
```

Out[12]: Text(0.5, 1.0, 'The Distribution Of The Days Before The Vacation')



We can learn from this graph that most of people choose to book their hotel up to 100 days from their arrival date.

Let's check which country consists the most hotels in this data set.

In [13]:
```python
countries = hotel_data.groupby('country')['country'].value_counts().head(10)
countries.plot.bar()
plt.axhline(32,color = "m", linewidth=2)
plt.xlabel("Countries")
plt.ylabel("Count")
plt.title("The number of hotels in each country")
```

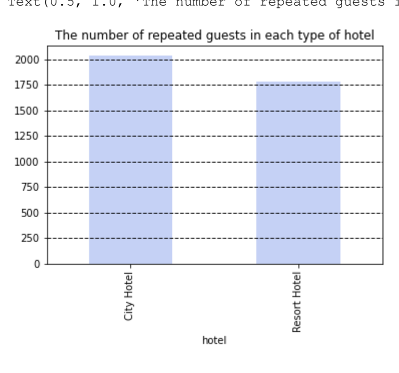
Out[13]: Text(0.5, 1.0, 'The number of hotels in each country')

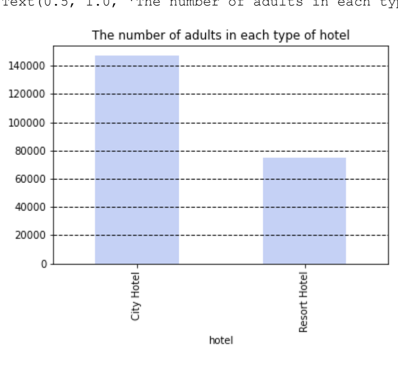
We can see that "AGO" and "ARG" consists of a lot more hotels than the others in this data set. Moreover, The average number of hotels in each country in this data set is a bit less than 50.

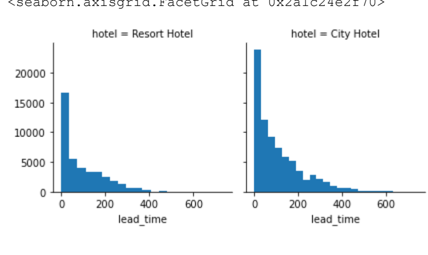
Checking the numbers of repeated guests in each type of hotel.

In [14]:
```python
hotel_data.groupby('hotel')['is_repeated_guest'].sum().plot(kind='bar',color='royalblue', alpha=0.3)
plt.grid(color='black', linestyle='--', linewidth=1, axis='y', alpha=0.9)
plt.title('The number of repeated guests in each type of hotel')
```

Out[14]: Text(0.5, 1.0, 'The number of repeated guests in each type of hotel')







Both are pretty the same.

Checking NaN values.

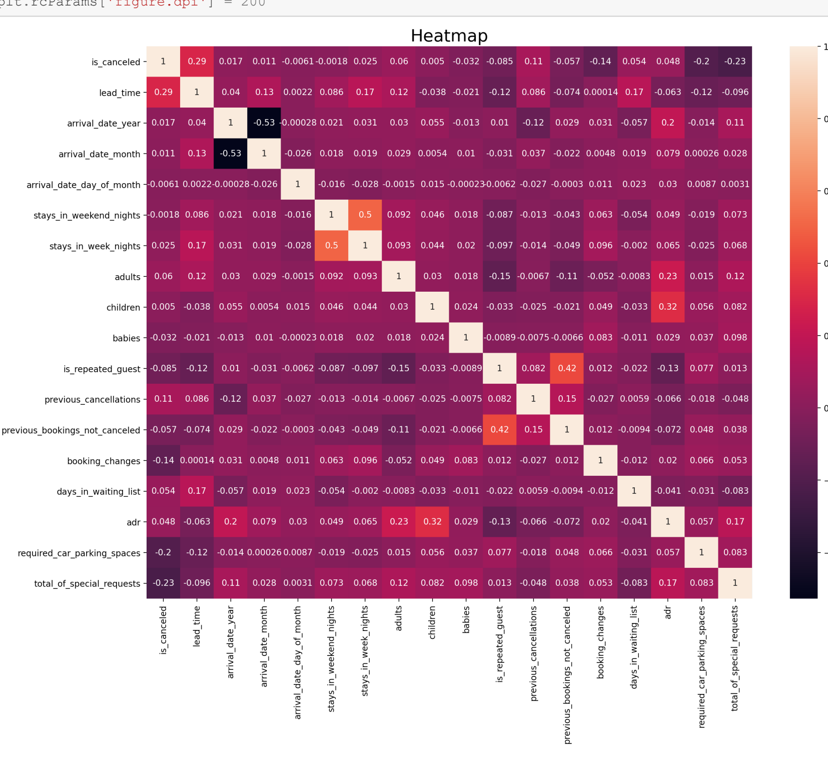
In [17]:
```python
hotel_data.isnull().sum()
```

Out[17]:
```
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
```

Most of the values in the company column are missing.
13.68% of the values in the agent column are missing as well.

Now let's check if there is correlation between the features.

In [64]:
```python
sns.heatmap(hotel_data.corr(), annot = True)
plt.title('Heatmap', fontsize = 20)
plt.rcParams['figure.figsize'] = [16, 12]
plt.rcParams['figure.dpi'] = 200
```



The correlation of most of the features is around 0. However, there is a positive correlation (0.5) between the stays in a week to weekend nights and between the market segment to distribution (0.77).
In addition, there is a negative correlation (-0.53) between the arrival date year to the month.

After I explored the data, my two main questions are:

- The question for the supervised learning is: Can we predict a type of hotel by the feature of booking?

- The question for the unsupervised learning is: Can we split our data to N clusters such that every cluster is including similar booking features?

Preprocessing

In [18]:
```python
from sklearn import preprocessing
```

First, I will delete unnecessary columns and rows which consists of NaN values.

In [19]:
```python
hotel_data = hotel_data.drop(columns=['arrival_date_week_number','country','reserved_room_type','assigned_room_type','agent','company','reservation_status_date'])
hotel_data = hotel_data.dropna(subset=['children'])
```

Deleting explantion

Columns:

- Arrival date week number - Two many different values that couldn't help to solve our main questions.
- Country - Two many different countries which can cause us to overfitting.
- Reserved and assigned room type - not belong to our main questions.
- Agent - Missing values.
- Company - Missing values.
- Reservation status date - not belong to our main questions.

Changing the name of the months to their numbers.

In [20]:
```python
hotel_data['arrival_date_month'] = hotel_data.arrival_date_month.apply(lambda x: x[:3])
months = {'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6, 'Jul': 7, 'Aug': 8, 'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12}
hotel_data.arrival_date_month = hotel_data.arrival_date_month.map(months)
```

Now I will do the preprocessing. One data will fit supervised learning and the other to unsupervised.

In [21]:
```python
def first_preprocessing(data_frame, supervised_or_unsupervised):
    try:
        new_data = data_frame.copy()
        if supervised_or_unsupervised == 'supervised':
            encoders = {'hotel': preprocessing.LabelEncoder(),
                        'meal': preprocessing.LabelEncoder(),
                        'market_segment': preprocessing.LabelEncoder(),
                        'distribution_channel': preprocessing.LabelEncoder(),
                        'deposit_type': preprocessing.LabelEncoder(),
                        'customer_type': preprocessing.LabelEncoder(),
                        'reservation_status': preprocessing.LabelEncoder()}
            encoder_columns_list = ['hotel','meal','market_segment','distribution_channel','deposit_type','customer_type','reservation_status']
            for column in encoder_columns_list:
                new_data[column] = encoders[column].fit_transform(new_data[column].astype(str))
        elif supervised_or_unsupervised == 'unsupervised':
            new_data = pd.get_dummies(new_data)
            columns = new_data.columns
            min_max_scaler = preprocessing.MinMaxScaler()
            new_data = min_max_scaler.fit_transform(new_data)
            new_data = pd.DataFrame(new_data)
            new_data.columns = columns
        return new_data
    except: print("Somthing got wrong - first_preprocessing")
```

Preprocessing explantion:

Supervised data:

- Changing the columns which contain object to int values. #### Unsupervised data:
- Using "One Hot Vector" to the columns that contain object values.
- Normalize the whole data set.

In [22]:
```python
hotel_supervised = first_preprocessing(hotel_data,'supervised')
hotel_supervised
```

Out[22]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_day_of_month	stays_in_weekend_nights
0	1	0	342	2015	7	1	0
1	1	0	737	2015	7	1	0
2	1	0	7	2015	7	1	0
3	1	0	13	2015	7	1	0
4	1	0	14	2015	7	1	0
...	...	...	...	...	...	...	...
119385	0	0	23	2017	8	30	2
119386	0	0	102	2017	8	31	2
119387	0	0	34	2017	8	31	2
119388	0	0	109	2017	8	31	2
119389	0	0	205	2017	8	29	2

119386 rows × 25 columns

Supervised data is ready.

In [23]:
```python
hotel_unsupervised = first_preprocessing(hotel_data,'unsupervised')
hotel_unsupervised
```

Out[23]:

	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_day_of_month	stays_in_weekend_nights
0	0.0	0.464043	0.0	0.545455	0.000000	0.000000
1	0.0	1.000000	0.0	0.545455	0.000000	0.000000
2	0.0	0.009498	0.0	0.545455	0.000000	0.000000
3	0.0	0.017639	0.0	0.545455	0.000000	0.000000
4	0.0	0.018996	0.0	0.545455	0.000000	0.000000
...	...	...	...	...	...	...
119381	0.0	0.031208	1.0	0.636364	0.966667	0.105263
119382	0.0	0.138399	1.0	0.636364	1.000000	0.105263
119383	0.0	0.046133	1.0	0.636364	1.000000	0.105263
119384	0.0	0.147897	1.0	0.636364	1.000000	0.105263
119385	0.0	0.278155	1.0	0.636364	0.933333	0.105263

119386 rows × 47 columns

Unsupervised data is ready.

Now let's try to answer the first question:
Can we predict a type of hotel by the feature of booking?
I will use supervised learning with the help of **Naive Bayes Classifier**. This classifier is a statistical classification technique based on Bayes Theorem.
Following the main question, I will try to predict the right type of hotel between the two:

- Resort hotel
- City hotel

In [24]:
```python
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score
```

Separating the target column and all the rest features:

In [25]:
```python
X = hotel_supervised.drop(columns = ['hotel']).copy()
y = hotel_supervised["hotel"].copy()
```

Building two functions that supposed to split the data for **train** and **test** and create the **Naive Bayes Classifier**:

In [26]:
```python
def split_test_train(X, y, test_size):
    try:
        return train_test_split(X, y, test_size=test_size, random_state=0)
    except:
        print("Something got wrong - split_test_train")
```

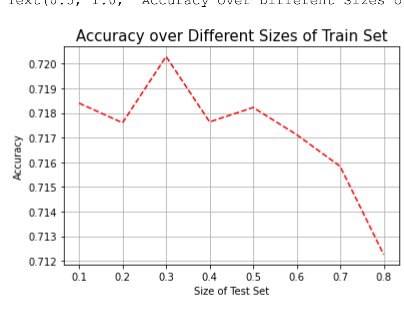
In [27]:
```python
def create_naive_bayes_classifier(X, y):
    try:
        model = GaussianNB()
        model.fit(X, y)
        return model
    except:
        print("Something got wrong - create_naive_bayes_classifier")
```

Before I choose the train and the test part, I want to make a plot that can show me what is the best accuracy based on the partition of the train and test.

In [28]:
```python
ratios = np.arange(0.1, 0.9, 0.1)
accuracy = []
for ratio in ratios:
    X_train, X_test, y_train, y_test = split_test_train(X, y, ratio)
    model = create_naive_bayes_classifier(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy.append(accuracy_score(y_test, y_pred))
```

In [29]:
```python
plt.grid(True)
plt.plot(ratios, accuracy, 'r--')
plt.xlabel('Size of Test Set')
plt.ylabel('Accuracy')
plt.title('Accuracy over Different Sizes of Train Set', fontsize=15)
```

Out[29]: Text(0.5, 1.0, 'Accuracy over Different Sizes of Train Set')



As we can see, The best accuracy I will get is if I will seperate the data to 0.3 test group and 0.7 train group.
Now we are ready to activate the algorithm and analyze the results.

In [30]:
```python
X_train, X_test, y_train, y_test = split_test_train(X, y, 0.3)
```

In [31]:
```python
model = create_naive_bayes_classifier(X_train, y_train)
```

In [32]:
```python
y_pred = model.predict(X_test)
```

In [33]:
```python
accuracy_score(y_test, y_pred)
```

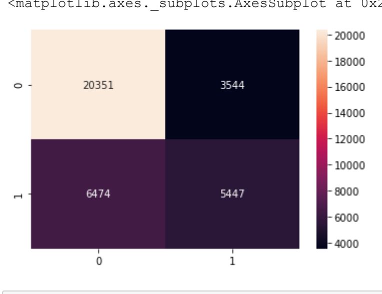
Out[33]: 0.7202926066562431

We received not bad accuracy depends on the starting point which our chances were 50/50.

Let's use a confusion matrix to see if we can find something significant.

In [34]:
```python
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt="d")
```

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x2a1c6ef93a0>



In [35]:
```python
tn, fp, fn, tp = confusion_matrix(y_test, y_pred).ravel()
print("TP:", tp)
print("FN:", fn)
print("FP:", fp)
print("TN:", tn)
```

```
TP: 5447
FN: 6474
FP: 3544
TN: 20351
```

Based of the confusion matrix we can see that it was easier to the algorithm to predict the city hotel(0).

To dig more into the result, I will use two rates:

- **Precision** - this rate using all the points predicted to positive to how many of them are actually positive. In our cases, the positive is the resort hotel.
- **Recall** - this is the ratio of the number correctly predicted positive class (resort hotel) to the total number of positive classes.

In [36]:
```python
precision_score(y_test, y_pred)
```

Out[36]: 0.6058280502724948

From the precision score, we learn that when the algorithm predicts resort hotel, in most of the cases he was right (60.5%).

In [37]:
```python
recall_score(y_test, y_pred)
```

Out[37]: 0.4569247546346783

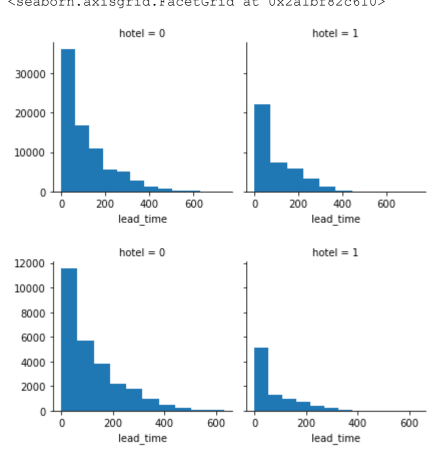
From the recall score we learn that most of the predictions of the actual resort hotel was false (45.6%).

Below are two plots about the difference between the actual data to the predicted one.

In [38]:
```python
predicted_data = X_test.copy()
predicted_data["hotel"] = y_pred
```

In [39]:
```python
graph = sns.FacetGrid(hotel_supervised, col='hotel')
graph.map(plt.hist, 'lead_time', bins=10)
graph = sns.FacetGrid(predicted_data, col='hotel')
graph.map(plt.hist, 'lead_time', bins=10)
```

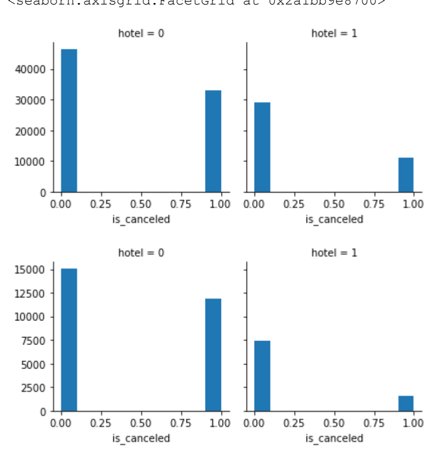
Out[39]: <seaborn.axisgrid.FacetGrid at 0x2a1bf82c610>



We can see that the distribution of the lead time between the actual data to the predicted data is pretty the same.

In [40]:
```python
graph = sns.FacetGrid(hotel_supervised, col='hotel')
graph.map(plt.hist, 'is_canceled', bins=10)
graph = sns.FacetGrid(predicted_data, col='hotel')
graph.map(plt.hist, 'is_canceled', bins=10)
```

Out[40]: <seaborn.axisgrid.FacetGrid at 0x2a1bb9e8700>



The relative gap between the cancellation of the actual data is a bit less than the predicted one.

After I got the result and I analyzed it, I wanted to check if I can optimize the algorithm by keeping the features that clarify the most the prediction (city/resort hotel). Therefore, I am going to use two methods about feature selection and after I will keep only the most important features, I will activate the algorithm one more time to check if the prediction was more accurate.

In [41]:
```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import RFE
```

Method 1 - Tree based Feature Selection and Random Forest Classification

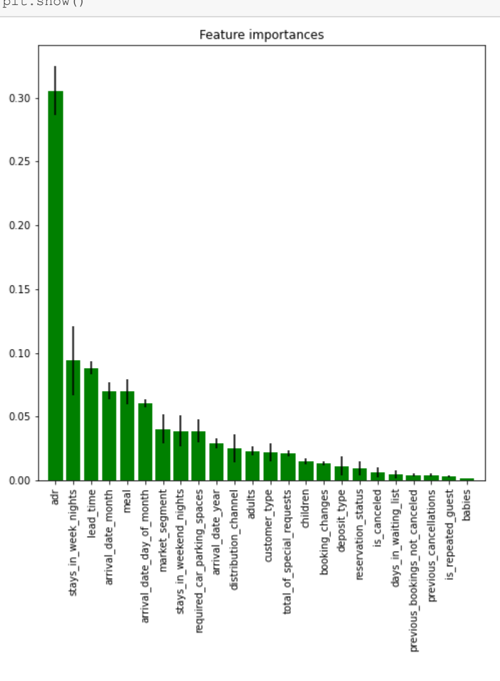
In this method, I will use **Random Forest** algorithm to predict the type of hotel. After that, I will use "feature*importances*" to check the contribution of any feature to the prediction and in the end, I will display it on bar plot.

In [42]:
```python
clf_rf_ = RandomForestClassifier(n_estimators=20)
clr_rf_ = clf_rf_.fit(X_train,y_train)
importances = clf_rf_.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf_rf_.estimators_],axis=0)
indices = np.argsort(importances)[::-1]
```

```
In [43]: print("Feature ranking:")
         for f in range(X_train.shape[1]):
             print("%d. feature %d (%f)" % (f + 1, indices[f], importances[indices[f]]))
```

```
Feature ranking:
1. feature 20 (0.305477)
2. feature 6 (0.093916)
3. feature 1 (0.088147)
4. feature 3 (0.070110)
5. feature 10 (0.069752)
6. feature 4 (0.060717)
7. feature 11 (0.040345)
8. feature 5 (0.038731)
9. feature 21 (0.038652)
10. feature 2 (0.029089)
11. feature 12 (0.025174)
12. feature 7 (0.023088)
13. feature 19 (0.021813)
14. feature 22 (0.021399)
15. feature 8 (0.014975)
16. feature 16 (0.013429)
17. feature 17 (0.011353)
18. feature 23 (0.009392)
19. feature 0 (0.006285)
20. feature 18 (0.004780)
21. feature 15 (0.004295)
22. feature 14 (0.004279)
23. feature 13 (0.003171)
24. feature 9 (0.001630)
```

```
In [44]: plt.figure(1, figsize=(8, 8))
         plt.title("Feature importances")
         plt.bar(range(X_train.shape[1]), importances[indices],
         color="g", yerr=std[indices], align="center")
         plt.xticks(range(X_train.shape[1]), X_train.columns[indices],rotation=90)
         plt.xlim([-1, X_train.shape[1]])
         plt.show()
```



We can learn that the feature "adr" is significantly more clearly to the prediction in comparison to the other features, for instance, this feature almost 3 times bigger (2.877) than the second most important feature.

Method 2 - Recursive feature elimination (RFE) with Random Forest Classification

The algorithm assign weights to each of features. Whose absolute weights are the smallest are pruned from the current set features. That procedure is recursively repeated on the pruned set until the desired number of features.

```
In [45]: clf_rf_ = RandomForestClassifier(n_estimators=20)
         rfe = RFE(estimator=clf_rf_, n_features_to_select=11, step=1)
         rfe = rfe.fit(X_train, y_train)
```

```
In [46]: print('Chosen best 11 feature by RFE:',X_train.columns[rfe.support_])
```

```
Chosen best 11 feature by RFE: Index(['lead_time', 'arrival_date_year', 'arrival_date_month',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'meal', 'market_segment', 'adr',
       'required_car_parking_spaces'],
      dtype='object')
```

This method brought me the same 11 most important features as the first method, therefore I will use only those 11 features and will activate the algorithm once again.

```
In [47]: new_hotel_supervised = hotel_supervised.drop(columns = ['deposit_type', 'children', 'booking_changes',
         'reservation_status', 'is_canceled', 'previous_bookings_not_canceled', 'previous_cancellations','days_i
         n_waiting_list','is_repeated_guest','babies','distribution_channel', 'deposit_type'])
```

```
In [48]: X = new_hotel_supervised.drop(columns = ['hotel']).copy()
         y = new_hotel_supervised["hotel"].copy()
```

```
In [49]: X_train, X_test, y_train, y_test = split_test_train(X, y, 0.3)
```

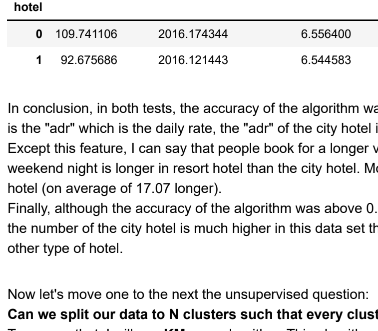
```
In [50]: model = create_naive_bayes_classifier(X_train, y_train)
```

```
In [51]: y_pred = model.predict(X_test)
```

```
In [52]: accuracy_score(y_test, y_pred)
Out[52]: 0.7288921152557516
```

I increased the accuracy by 0.008 points - unfortunately not so significantly.

```
In [53]: sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt="d")
Out[53]: &lt;matplotlib.axes._subplots.AxesSubplot at 0x2a1d0ffd790&gt;
```




```
In [54]: tn, fp, fn, tp = confusion_matrix(y_test, y_pred).ravel()
         print("TP:", tp)
         print("FN:", fn)
         print("FP:", fp)
         print("TN:", tn)
```

```
TP: 3893
FN: 8028
FP: 1682
TN: 22213
```

From the confusion matrix, we can see that the algorithm predicted better the city hotel. However, the prediction of the resulting hotel was less good than before.

```
In [55]: hotel_type = hotel_supervised.groupby('hotel')[['lead_time', 'arrival_date_year', 'arrival_date_month',
                'arrival_date_day_of_month', 'stays_in_weekend_nights',
                'stays_in_week_nights', 'adults', 'meal', 'adr',
                'required_car_parking_spaces', 'total_of_special_requests']].mean()
         hotel_type
```

Out[55]:

	lead_time	arrival_date_year	arrival_date_month	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adu
hotel							
0	109.741106	2016.174344	6.556400	15.787094	0.795187	2.182954	1.850
1	92.675686	2016.121443	6.544583	15.821243	1.189815	3.128732	1.867

In conclusion, in both tests, the accuracy of the algorithm was around 0.72. The feature that has the most clarity about the type of the hotel is the "adr" which is the daily rate, the "adr" of the city hotel is significantly higher than the resort hotel (on average of 10.348 higher). Except this feature, I can say that people book for a longer vacation in resort hotel since the average of the stays in the week and the weekend night is longer in resort hotel than the city hotel. Moreover, the lead time of the booking for the city hotel is longer than the resort hotel (on average of 17.07 longer).
Finally, although the accuracy of the algorithm was above 0.5 (our starting point) the algorithm got difficulties to predict the resort hotel since the number of the city hotel is much higher in this data set than the resort hotel, and the range of almost every feature was very alike to the other type of hotel.

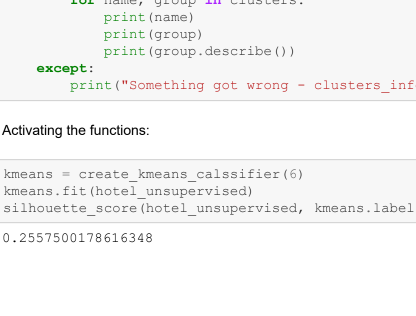
Now let's move one to the next the unsupervised question:
Can we split our data to N clusters such that every cluster is including similar booking features?
To answer that, I will use **KMeans** algorithm. This algorithm is a method of vector quantization, originally from signal processing, that aims to partition **n** observations into **k** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

```
In [56]: from sklearn.cluster import KMeans
         from sklearn.metrics import silhouette_score
```

Firstly, I need to find out how many clusters do I need to divide into each data I used. In order to deal with this dilemma, I will build a function which called 'number_of_clusters()'. This function will return us two graphs. The first graph is called "Sum of square error ", in this graph we will see what is the sum of square error as long as the number of clusters increases. Now comes the question of what is the number of clusters that are "suspicious" to be optimal. For answering this question, I will use the "Elbow Method". In this method, when we see a break in some number clusters, we will take that number as an option about how many clusters should we use in this data set. The second graph I will use is called "Silhouette", the silhouette is a measure about how good our algorithm work, his range is between -1 to 1. As long as the value is close to 1 we can say that the clustering was good. In our graph, I will wish to choose the number of clusters by the best silhouette I got. I need to find the best cluster by a combination of those two graphs, a number of clusters which has a good "Elbow break" and ones that has the best silhouette.

```
In [57]: def number_of_clusters(data_frame):
             try:
                 sum_squared = []
                 silhouette = []
                 for i in range(2, 11):
                     kmeans = KMeans(n_clusters=i, init='k-means++')
                     kmeans.fit(data_frame)
                     sum_squared.append(kmeans.inertia_)
                     silhouette.append(silhouette_score(data_frame, kmeans.labels_))
                 x1 = (range(2, 11))
                 x2 = (range(2, 11))
                 y1 = sum_squared
                 y2 = silhouette
                 plt.subplot(2, 1, 1)
                 plt.plot(x1, y1)
                 plt.title(r'Sum of Squred Error ${R_2}$', fontsize=15)
                 plt.grid()
                 plt.ylabel(r'${R_2}$')
                 plt.subplot(2, 1, 2)
                 plt.plot(x2, y2)
                 plt.title('Silhouette', fontsize=15)
                 plt.xlabel('No. of Clusters')
                 plt.ylabel('Silhouette')
                 plt.grid()
                 plt.show()
             except:
                 print("Something got wrong - number_of_clusters")
```

```
In [58]: number_of_clusters(hotel_unsupervised)
```



As we can see, in the first plot, the sharper elbow break is when I divide the data into 6 clusters. However, when we look at the second plot, the highest silhouette is when I divide the data into 10 clusters although the silhouette score in each number of clusters is not high. Therefore, I will divide the data into 6 clusters.

Below is two functions:

- **create_kmeans_classifier** - creating KMeans algorithm.
- **clusters_information** - display each cluster and his statistic details.

```
In [59]: def create_kmeans_calssifier(k):
             try:
                 return KMeans(n_clusters=k, init='k-means++')
             except:
                 print("Something got wrong - create_kmeans_calssifier")
```

```
In [60]: def clusters_information(data_frame):
             try:
                 clusters = data_frame.groupby("label")
                 for name, group in clusters:
                     print(name)
                     print(group)
                     print(group.describe())
             except:
                 print("Something got wrong - clusters_information")
```

Activating the functions:

```
In [61]: kmeans = create_kmeans_calssifier(6)
         kmeans.fit(hotel_unsupervised)
         silhouette_score(hotel_unsupervised, kmeans.labels_)
Out[61]: 0.2557500178616348
```

```
mean                 0.000053                 0.000426
std                  0.007295                 0.020629
min                  0.000000                 0.000000
25%                  0.000000                 0.000000
50%                  0.000000                 0.000000
75%                  0.000000                 0.000000
max                  1.000000                 1.000000

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
count            18792.000000         18792.000000             18792.000000   
mean                 0.085834             0.010909                 0.834291   
std                  0.280127             0.103877                 0.371829   
min                  0.000000             0.000000                 0.000000   
25%                  0.000000             0.000000                 1.000000   
50%                  0.000000             0.000000                 1.000000   
75%                  0.000000             0.000000                 1.000000   
max                  1.000000             1.000000                 1.000000   

       customer_type_Transient-Party  reservation_status_Canceled  \
count                   18792.000000                      18792.0   
mean                        0.068966                          0.0   
std                         0.253402                          0.0   
min                         0.000000                          0.0   
25%                         0.000000                          0.0   
50%                         0.000000                          0.0   
75%                         0.000000                          0.0   
max                         1.000000                          0.0   

       reservation_status_Check-Out  reservation_status_No-Show    label
count                  18792.000000                18792.000000  18792.0
mean                       0.996275                    0.003725      1.0
std                        0.060921                    0.060921      0.0
min                        0.000000                    0.000000      1.0
25%                        1.000000                    0.000000      1.0
50%                        1.000000                    0.000000      1.0
75%                        1.000000                    0.000000      1.0
max                        1.000000                    1.000000      1.0

[8 rows x 48 columns]
2
        is_canceled  lead_time  arrival_date_year  arrival_date_month  \
47              0.0   0.153324                0.0            0.545455   
50              0.0   0.153324                0.0            0.545455   
313             0.0   0.122117                0.0            0.545455   
317             0.0   0.122117                0.0            0.545455   
347             0.0   0.156038                0.0            0.545455   
...             ...        ...                ...                 ...   
119343          0.0   0.085482                1.0            0.636364   
119344          0.0   0.085482                1.0            0.636364   
119348          0.0   0.085482                1.0            0.636364   
119349          0.0   0.085482                1.0            0.636364   
119350          0.0   0.085482                1.0            0.636364   

        arrival_date_day_of_month  stays_in_weekend_nights  \
47                       0.033333                 0.105263   
50                       0.033333                 0.105263   
313                      0.366667                 0.105263   
317                      0.366667                 0.105263   
347                      0.400000                 0.105263   
...                           ...                      ...   
119343                   1.000000                 0.000000   
119344                   1.000000                 0.000000   
119348                   1.000000                 0.000000   
119349                   1.000000                 0.000000   
119350                   1.000000                 0.000000   

        stays_in_week_nights    adults  children  babies  ...  \
47                      0.10  0.036364       0.0     0.0  ...   
50                      0.10  0.036364       0.0     0.0  ...   
313                     0.10  0.036364       0.0     0.0  ...   
317                     0.10  0.036364       0.0     0.1  ...   
347                     0.10  0.036364       0.0     0.1  ...   
...                      ...       ...       ...     ...  ...   
119343                  0.06  0.054545       0.0     0.0  ...   
119344                  0.06  0.054545       0.0     0.0  ...   
119348                  0.06  0.054545       0.0     0.0  ...   
119349                  0.06  0.054545       0.0     0.0  ...   
119350                  0.06  0.054545       0.0     0.0  ...   

        deposit_type_Non Refund  deposit_type_Refundable  \
47                          0.0                      0.0   
50                          0.0                      0.0   
313                         0.0                      0.0   
317                         0.0                      0.0   
347                         0.0                      0.0   
...                         ...                      ...   
119343                      0.0                      0.0   
119344                      0.0                      0.0   
119348                      0.0                      0.0   
119349                      0.0                      0.0   
119350                      0.0                      0.0   

        customer_type_Contract  customer_type_Group  customer_type_Transient  \
47                         0.0                  0.0                      0.0   
50                         0.0                  0.0                      0.0   
313                        0.0                  0.0                      0.0   
317                        0.0                  0.0                      0.0   
347                        0.0                  0.0                      0.0   
...                        ...                  ...                      ...   
119343                     0.0                  0.0                      0.0   
119344                     0.0                  0.0                      0.0   
119348                     0.0                  0.0                      0.0   
119349                     0.0                  0.0                      0.0   
119350                     0.0                  0.0                      0.0   

        customer_type_Transient-Party  reservation_status_Canceled  \
47                                1.0                          0.0   
50                                1.0                          0.0   
313                               1.0                          0.0   
317                               1.0                          0.0   
347                               1.0                          0.0   
...                               ...                          ...   
119343                            1.0                          0.0   
119344                            1.0                          0.0   
119348                            1.0                          0.0   
119349                            1.0                          0.0   
119350                            1.0                          0.0   

        reservation_status_Check-Out  reservation_status_No-Show  label  
47                               1.0                         0.0      2  
50                               1.0                         0.0      2  
313                              1.0                         0.0      2  
317                              1.0                         0.0      2  
347                              1.0                         0.0      2  
...                              ...                         ...    ...  
119343                           1.0                         0.0      2  
119344                           1.0                         0.0      2  
119348                           1.0                         0.0      2  
119349                           1.0                         0.0      2  
119350                           1.0                         0.0      2  

[21331 rows x 48 columns]
        is_canceled     lead_time  arrival_date_year  arrival_date_month  \
count  21331.000000  21331.000000       21331.000000        21331.000000   
mean       0.224040      0.186384           0.461394            0.526285   
std        0.416958      0.163469           0.361954            0.282928   
min        0.000000      0.000000           0.000000            0.000000   
25%        0.000000      0.050204           0.000000            0.272727   
50%        0.000000      0.138399           0.500000            0.545455   
75%        0.000000      0.289009           0.500000            0.727273   
max        1.000000      0.735414           1.000000            1.000000   

       arrival_date_day_of_month  stays_in_weekend_nights  \
count               21331.000000             21331.000000   
mean                    0.498944                 0.042352   
std                     0.284473                 0.048352   
min                     0.000000                 0.000000   
25%                     0.266667                 0.000000   
50%                     0.500000                 0.052632   
75%                     0.733333                 0.105263   
max                     1.000000                 0.736842   

       stays_in_week_nights        adults      children        babies  ...  \
count          21331.000000  21331.000000  21331.000000  21331.000000  ...   
mean               0.044954      0.031076      0.001814      0.000263  ...   
std                0.032545      0.009231      0.016433      0.007742  ...   
min                0.000000      0.000000      0.000000      0.000000  ...   
25%                0.020000      0.018182      0.000000      0.000000  ...   
50%                0.040000      0.036364      0.000000      0.000000  ...   
75%                0.060000      0.036364      0.000000      0.000000  ...   
max                0.700000      0.072727      0.300000      0.900000  ...   

       deposit_type_Non Refund  deposit_type_Refundable  \
count             21331.000000             21331.000000   
mean                  0.004172                 0.005954   
std                   0.064460                 0.076932   
min                   0.000000                 0.000000   
25%                   0.000000                 0.000000   
50%                   0.000000                 0.000000   
75%                   0.000000                 0.000000   
max                   1.000000                 1.000000   

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
count            21331.000000         21331.000000                  21331.0   
mean                 0.004454             0.005626                      0.0   
std                  0.066588             0.074795                      0.0   
min                  0.000000             0.000000                      0.0   
25%                  0.000000             0.000000                      0.0   
50%                  0.000000             0.000000                      0.0   
75%                  0.000000             0.000000                      0.0   
max                  1.000000             1.000000                      0.0   

       customer_type_Transient-Party  reservation_status_Canceled  \
count                   21331.000000                 21331.000000   
mean                        0.989921                     0.214992   
std                         0.099890                     0.410827   
min                         0.000000                     0.000000   
25%                         1.000000                     0.000000   
50%                         1.000000                     0.000000   
75%                         1.000000                     0.000000   
max                         1.000000                     1.000000   

       reservation_status_Check-Out  reservation_status_No-Show    label
count                  21331.000000                21331.000000  21331.0
mean                       0.775960                    0.009048      2.0
std                        0.416958                    0.094691      0.0
min                        0.000000                    0.000000      2.0
25%                        1.000000                    0.000000      2.0
50%                        1.000000                    0.000000      2.0
75%                        1.000000                    0.000000      2.0
max                        1.000000                    1.000000      2.0

[8 rows x 48 columns]
3
        is_canceled  lead_time  arrival_date_year  arrival_date_month  \
40060           0.0   0.008141                0.0            0.545455   
40085           0.0   0.081411                0.0            0.545455   
40104           0.0   0.093623                0.0            0.545455   
40113           0.0   0.073270                0.0            0.545455   
40119           0.0   0.069199                0.0            0.545455   
...             ...        ...                ...                 ...   
119381          0.0   0.031208                1.0            0.636364   
119382          0.0   0.138399                1.0            0.636364   
119383          0.0   0.046133                1.0            0.636364   
119384          0.0   0.147897                1.0            0.636364   
119385          0.0   0.278155                1.0            0.636364   

        arrival_date_day_of_month  stays_in_weekend_nights  \
40060                    0.000000                 0.000000   
40085                    0.066667                 0.105263   
40104                    0.200000                 0.000000   
40113                    0.200000                 0.105263   
40119                    0.233333                 0.052632   
...                           ...                      ...   
119381                   0.966667                 0.105263   
119382                   1.000000                 0.105263   
119383                   1.000000                 0.105263   
119384                   1.000000                 0.105263   
119385                   0.933333                 0.105263   

        stays_in_week_nights    adults  children  babies  ...  \
40060                   0.04  0.018182       0.0     0.0  ...   
40085                   0.10  0.036364       0.0     0.0  ...   
40104                   0.04  0.036364       0.1     0.0  ...   
40113                   0.16  0.036364       0.0     0.0  ...   
40119                   0.08  0.036364       0.0     0.0  ...   
...                      ...       ...       ...     ...  ...   
119381                  0.10  0.036364       0.0     0.0  ...   
119382                  0.10  0.054545       0.0     0.0  ...   
119383                  0.10  0.036364       0.0     0.0  ...   
119384                  0.10  0.036364       0.0     0.0  ...   
119385                  0.14  0.036364       0.0     0.0  ...   

        deposit_type_Non Refund  deposit_type_Refundable  \
40060                       0.0                      0.0   
40085                       0.0                      0.0   
40104                       0.0                      0.0   
40113                       0.0                      0.0   
40119                       0.0                      0.0   
...                         ...                      ...   
119381                      0.0                      0.0   
119382                      0.0                      0.0   
119383                      0.0                      0.0   
119384                      0.0                      0.0   
119385                      0.0                      0.0   

        customer_type_Contract  customer_type_Group  customer_type_Transient  \
40060                      0.0                  0.0                      1.0   
40085                      0.0                  0.0                      1.0   
40104                      0.0                  0.0                      1.0   
40113                      0.0                  0.0                      1.0   
40119                      0.0                  0.0                      1.0   
...                        ...                  ...                      ...   
119381                     0.0                  0.0                      1.0   
119382                     0.0                  0.0                      1.0   
119383                     0.0                  0.0                      1.0   
119384                     0.0                  0.0                      1.0   
119385                     0.0                  0.0                      1.0   

        customer_type_Transient-Party  reservation_status_Canceled  \
40060                             0.0                          0.0   
40085                             0.0                          0.0   
40104                             0.0                          0.0   
40113                             0.0                          0.0   
40119                             0.0                          0.0   
...                               ...                          ...   
119381                            0.0                          0.0   
119382                            0.0                          0.0   
119383                            0.0                          0.0   
119384                            0.0                          0.0   
119385                            0.0                          0.0   

        reservation_status_Check-Out  reservation_status_No-Show  label  
40060                            1.0                         0.0      3  
40085                            1.0                         0.0      3  
40104                            1.0                         0.0      3  
40113                            1.0                         0.0      3  
40119                            1.0                         0.0      3  
...                              ...                         ...    ...  
119381                           1.0                         0.0      3  
119382                           1.0                         0.0      3  
119383                           1.0                         0.0      3  
119384                           1.0                         0.0      3  
119385                           1.0                         0.0      3  

[28723 rows x 48 columns]
       is_canceled     lead_time  arrival_date_year  arrival_date_month  \
count      28723.0  28723.000000       28723.000000        28723.000000   
mean           0.0      0.089827           0.638165            0.495207   
std            0.0      0.093541           0.321347            0.281916   
min            0.0      0.000000           0.000000            0.000000   
25%            0.0      0.014925           0.500000            0.272727   
50%            0.0      0.056988           0.500000            0.454545   
75%            0.0      0.138399           1.000000            0.727273   
max            0.0      0.582090           1.000000            1.000000   

       arrival_date_day_of_month  stays_in_weekend_nights  \
count               28723.000000             28723.000000   
mean                    0.487402                 0.045601   
std                     0.294238                 0.045759   
min                     0.000000                 0.000000   
25%                     0.233333                 0.000000   
50%                     0.500000                 0.052632   
75%                     0.733333                 0.105263   
max                     1.000000                 0.842105   

       stays_in_week_nights        adults      children        babies  ...  \
count          28723.000000  28723.000000  28723.000000  28723.000000  ...   
mean               0.044183      0.034442      0.011639      0.000630  ...   
std                0.029269      0.009554      0.041287      0.009765  ...   
min                0.000000      0.000000      0.000000      0.000000  ...   
25%                0.020000      0.036364      0.000000      0.000000  ...   
50%                0.040000      0.036364      0.000000      0.000000  ...   
75%                0.060000      0.036364      0.000000      0.000000  ...   
max                0.820000      0.072727      0.300000      1.000000  ...   

       deposit_type_Non Refund  deposit_type_Refundable  \
count                  28723.0                  28723.0   
mean                       0.0                      0.0   
std                        0.0                      0.0   
min                        0.0                      0.0   
25%                        0.0                      0.0   
50%                        0.0                      0.0   
75%                        0.0                      0.0   
max                        0.0                      0.0   

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
count            28723.000000         28723.000000             28723.000000   
mean                 0.038332             0.003516                 0.953173   
std                  0.191999             0.059195                 0.211271   
min                  0.000000             0.000000                 0.000000   
25%                  0.000000             0.000000                 1.000000   
50%                  0.000000             0.000000                 1.000000   
75%                  0.000000             0.000000                 1.000000   
max                  1.000000             1.000000                 1.000000   

       customer_type_Transient-Party  reservation_status_Canceled  \
count                   28723.000000                      28723.0   
mean                        0.004979                          0.0   
std                         0.070384                          0.0   
min                         0.000000                          0.0   
25%                         0.000000                          0.0   
50%                         0.000000                          0.0   
75%                         0.000000                          0.0   
max                         1.000000                          0.0   

       reservation_status_Check-Out  reservation_status_No-Show    label
count                       28723.0                     28723.0  28723.0
mean                            1.0                         0.0      3.0
std                             0.0                         0.0      0.0
min                             1.0                         0.0      3.0
25%                             1.0                         0.0      3.0
50%                             1.0                         0.0      3.0
75%                             1.0                         0.0      3.0
max                             1.0                         0.0      3.0

[8 rows x 48 columns]
4
       is_canceled  lead_time  arrival_date_year  arrival_date_month  \
2133           1.0   0.118046                0.0            0.818182   
2134           1.0   0.118046                0.0            0.818182   
2135           1.0   0.118046                0.0            0.818182   
2136           1.0   0.118046                0.0            0.818182   
2141           1.0   0.118046                0.0            0.818182   
...            ...        ...                ...                 ...   
82926          1.0   0.240163                0.5            0.363636   
82927          1.0   0.240163                0.5            0.363636   
82928          1.0   0.240163                0.5            0.363636   
82929          1.0   0.240163                0.5            0.363636   
85207          1.0   0.412483                0.5            1.000000   

       arrival_date_day_of_month  stays_in_weekend_nights  \
2133                    0.066667                 0.000000   
2134                    0.066667                 0.000000   
2135                    0.066667                 0.000000   
2136                    0.066667                 0.000000   
2141                    0.066667                 0.000000   
...                          ...                      ...   
82926                   0.233333                 0.105263   
82927                   0.233333                 0.105263   
82928                   0.233333                 0.105263   
82929                   0.233333                 0.105263   
85207                   0.833333                 0.052632   

       stays_in_week_nights    adults  children  babies  ...  \
2133                   0.02  0.036364       0.0     0.0  ...   
2134                   0.02  0.036364       0.0     0.0  ...   
2135                   0.02  0.036364       0.0     0.0  ...   
2136                   0.02  0.036364       0.0     0.0  ...   
2141                   0.02  0.036364       0.0     0.0  ...   
...                     ...       ...       ...     ...  ...   
82926                  0.00  0.036364       0.0     0.0  ...   
82927                  0.00  0.036364       0.0     0.0  ...   
82928                  0.00  0.036364       0.0     0.0  ...   
82929                  0.00  0.036364       0.0     0.0  ...   
85207                  0.04  0.018182       0.0     0.0  ...   

       deposit_type_Non Refund  deposit_type_Refundable  \
2133                       1.0                      0.0   
2134                       1.0                      0.0   
2135                       1.0                      0.0   
2136                       1.0                      0.0   
2141                       1.0                      0.0   
...                        ...                      ...   
82926                      1.0                      0.0   
82927                      1.0                      0.0   
82928                      1.0                      0.0   
82929                      1.0                      0.0   
85207                      1.0                      0.0   

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
2133                      0.0                  0.0                      1.0   
2134                      0.0                  0.0                      1.0   
2135                      0.0                  0.0                      1.0   
2136                      0.0                  0.0                      1.0   
2141                      0.0                  0.0                      1.0   
...                       ...                  ...                      ...   
82926                     0.0                  0.0                      1.0   
82927                     0.0                  0.0                      1.0   
82928                     0.0                  0.0                      1.0   
82929                     0.0                  0.0                      1.0   
85207                     0.0                  0.0                      0.0   

       customer_type_Transient-Party  reservation_status_Canceled  \
2133                             0.0                          1.0   
2134                             0.0                          1.0   
2135                             0.0                          1.0   
2136                             0.0                          1.0   
2141                             0.0                          1.0   
...                              ...                          ...   
82926                            0.0                          1.0   
82927                            0.0                          1.0   
82928                            0.0                          1.0   
82929                            0.0                          1.0   
85207                            1.0                          1.0   

       reservation_status_Check-Out  reservation_status_No-Show  label  
2133                            0.0                         0.0      4  
2134                            0.0                         0.0      4  
2135                            0.0                         0.0      4  
2136                            0.0                         0.0      4  
2141                            0.0                         0.0      4  
...                             ...                         ...    ...  
82926                           0.0                         0.0      4  
82927                           0.0                         0.0      4  
82928                           0.0                         0.0      4  
82929                           0.0                         0.0      4  
85207                           0.0                         0.0      4  

[14703 rows x 48 columns]
       is_canceled     lead_time  arrival_date_year  arrival_date_month  \
count      14703.0  14703.000000       14703.000000        14703.000000   
mean           1.0      0.290561           0.508604            0.512419   
std            0.0      0.179673           0.368569            0.276918   
min            1.0      0.000000           0.000000            0.000000   
25%            1.0      0.141113           0.000000            0.272727   
50%            1.0      0.252374           0.500000            0.454545   
75%            1.0      0.412483           1.000000            0.727273   
max            1.0      0.853460           1.000000            1.000000   

       arrival_date_day_of_month  stays_in_weekend_nights  \
count               14703.000000             14703.000000   
mean                    0.486975                 0.032507   
std                     0.288724                 0.043046   
min                     0.000000                 0.000000   
25%                     0.233333                 0.000000   
50%                     0.500000                 0.000000   
75%                     0.733333                 0.052632   
max                     1.000000                 0.368421   

       stays_in_week_nights        adults      children   babies  ...  \
count          14703.000000  14703.000000  14703.000000  14703.0  ...   
mean               0.041802      0.033001      0.000027      0.0  ...   
std                0.022936      0.007113      0.002020      0.0  ...   
min                0.000000      0.018182      0.000000      0.0  ...   
25%                0.020000      0.036364      0.000000      0.0  ...   
50%                0.040000      0.036364      0.000000      0.0  ...   
75%                0.060000      0.036364      0.000000      0.0  ...   
max                0.320000      0.054545      0.200000      0.0  ...   

       deposit_type_Non Refund  deposit_type_Refundable  \
count             14703.000000             14703.000000   
mean                  0.983881                 0.000408   
std                   0.125938                 0.020198   
min                   0.000000                 0.000000   
25%                   1.000000                 0.000000   
50%                   1.000000                 0.000000   
75%                   1.000000                 0.000000   
max                   1.000000                 1.000000   

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
count            14703.000000              14703.0             14703.000000   
mean                 0.048085                  0.0                 0.880841   
std                  0.213954                  0.0                 0.323987   
min                  0.000000                  0.0                 0.000000   
25%                  0.000000                  0.0                 1.000000   
50%                  0.000000                  0.0                 1.000000   
75%                  0.000000                  0.0                 1.000000   
max                  1.000000                  0.0                 1.000000   

       customer_type_Transient-Party  reservation_status_Canceled  \
count                   14703.000000                 14703.000000   
mean                        0.071074                     0.998844   
std                         0.256957                     0.033985   
min                         0.000000                     0.000000   
25%                         0.000000                     1.000000   
50%                         0.000000                     1.000000   
75%                         0.000000                     1.000000   
max                         1.000000                     1.000000   

       reservation_status_Check-Out  reservation_status_No-Show    label
count                       14703.0                14703.000000  14703.0
mean                            0.0                    0.001156      4.0
std                             0.0                    0.033985      0.0
min                             0.0                    0.000000      4.0
25%                             0.0                    0.000000      4.0
50%                             0.0                    0.000000      4.0
75%                             0.0                    0.000000      4.0
max                             0.0                    1.000000      4.0

[8 rows x 48 columns]
5
        is_canceled  lead_time  arrival_date_year  arrival_date_month  \
0               0.0   0.464043                0.0            0.545455   
1               0.0   1.000000                0.0            0.545455   
2               0.0   0.009498                0.0            0.545455   
6               0.0   0.000000                0.0            0.545455   
7               0.0   0.012212                0.0            0.545455   
...             ...        ...                ...                 ...   
119354          0.0   0.183175                1.0            0.636364   
119361          0.0   0.073270                1.0            0.636364   
119364          0.0   0.354138                1.0            0.636364   
119365          0.0   0.280868                1.0            0.636364   
119377          0.0   0.255088                1.0            0.636364   

        arrival_date_day_of_month  stays_in_weekend_nights  \
0                        0.000000                 0.000000   
1                        0.000000                 0.000000   
2                        0.000000                 0.000000   
6                        0.000000                 0.000000   
7                        0.000000                 0.000000   
...                           ...                      ...   
119354                   0.966667                 0.052632   
119361                   0.866667                 0.157895   
```

```
119364                 0.966667                 0.052632
119365                 0.800000                 0.157895
119377                 1.000000                 0.105263

        stays_in_week_nights    adults  children  babies  ...  \
0                       0.00  0.036364       0.0     0.0  ...
1                       0.00  0.036364       0.0     0.0  ...
2                       0.02  0.018182       0.0     0.0  ...
6                       0.04  0.036364       0.0     0.0  ...
7                       0.04  0.036364       0.0     0.0  ...
...                      ...       ...       ...     ...  ...
119354                  0.08  0.036364       0.0     0.0  ...
119361                  0.10  0.018182       0.0     0.0  ...
119364                  0.08  0.036364       0.0     0.0  ...
119365                  0.14  0.036364       0.0     0.0  ...
119377                  0.06  0.036364       0.0     0.0  ...

        deposit_type_Non Refund  deposit_type_Refundable  \
0                           0.0                      0.0
1                           0.0                      0.0
2                           0.0                      0.0
6                           0.0                      0.0
7                           0.0                      0.0
...                         ...                      ...
119354                      0.0                      0.0
119361                      0.0                      0.0
119364                      0.0                      0.0
119365                      0.0                      0.0
119377                      0.0                      0.0

        customer_type_Contract  customer_type_Group  customer_type_Transient  \
0                          0.0                  0.0                      1.0
1                          0.0                  0.0                      1.0
2                          0.0                  0.0                      1.0
6                          0.0                  0.0                      1.0
7                          0.0                  0.0                      1.0
...                        ...                  ...                      ...
119354                     0.0                  0.0                      1.0
119361                     0.0                  0.0                      1.0
119364                     0.0                  0.0                      1.0
119365                     0.0                  0.0                      1.0
119377                     0.0                  0.0                      1.0

        customer_type_Transient-Party  reservation_status_Canceled  \
0                                 0.0                          0.0
1                                 0.0                          0.0
2                                 0.0                          0.0
6                                 0.0                          0.0
7                                 0.0                          0.0
...                               ...                          ...
119354                            0.0                          0.0
119361                            0.0                          0.0
119364                            0.0                          0.0
119365                            0.0                          0.0
119377                            0.0                          0.0

        reservation_status_Check-Out  reservation_status_No-Show  label
0                                1.0                         0.0      5
1                                1.0                         0.0      5
2                                1.0                         0.0      5
6                                1.0                         0.0      5
7                                1.0                         0.0      5
...                              ...                         ...    ...
119354                           1.0                         0.0      5
119361                           1.0                         0.0      5
119364                           1.0                         0.0      5
119365                           1.0                         0.0      5
119377                           1.0                         0.0      5

[13115 rows x 48 columns]
       is_canceled     lead_time  arrival_date_year  arrival_date_month  \
count  13115.00000  13115.000000       13115.000000        13115.000000
mean       0.14838      0.065632           0.583568            0.495228
std        0.35549      0.093518           0.360882            0.292798
min        0.00000      0.000000           0.000000            0.000000
25%        0.00000      0.001357           0.500000            0.272727
50%        0.00000      0.018996           0.500000            0.545455
75%        0.00000      0.097693           1.000000            0.727273
max        1.00000      1.000000           1.000000            1.000000

       arrival_date_day_of_month  stays_in_weekend_nights  \
count               13115.000000             13115.000000
mean                    0.497385                 0.043670
std                     0.295283                 0.053497
min                     0.000000                 0.000000
25%                     0.233333                 0.000000
50%                     0.500000                 0.052632
75%                     0.766667                 0.105263
max                     1.000000                 1.000000

       stays_in_week_nights        adults      children        babies  ...  \
count          13115.000000  13115.000000  13115.000000  13115.000000  ...
mean               0.046027      0.033725      0.017339      0.002356  ...
std                0.041191      0.017458      0.050192      0.015565  ...
min                0.000000      0.000000      0.000000      0.000000  ...
25%                0.020000      0.036364      0.000000      0.000000  ...
50%                0.040000      0.036364      0.000000      0.000000  ...
75%                0.060000      0.036364      0.000000      0.000000  ...
max                1.000000      1.000000      0.300000      0.200000  ...

       deposit_type_Non Refund  deposit_type_Refundable  \
count             13115.000000             13115.000000
mean                  0.000229                 0.000229
std                   0.015123                 0.015123
min                   0.000000                 0.000000
25%                   0.000000                 0.000000
50%                   0.000000                 0.000000
75%                   0.000000                 0.000000
max                   1.000000                 1.000000

       customer_type_Contract  customer_type_Group  customer_type_Transient  \
count            13115.000000         13115.000000             13115.000000
mean                 0.001220             0.008616                 0.905757
std                  0.034908             0.092426                 0.292178
min                  0.000000             0.000000                 0.000000
25%                  0.000000             0.000000                 1.000000
50%                  0.000000             0.000000                 1.000000
75%                  0.000000             0.000000                 1.000000
max                  1.000000             1.000000                 1.000000

       customer_type_Transient-Party  reservation_status_Canceled  \
count                   13115.000000                 13115.000000
mean                        0.084407                     0.131224
std                         0.278008                     0.337658
min                         0.000000                     0.000000
25%                         0.000000                     0.000000
50%                         0.000000                     0.000000
75%                         0.000000                     0.000000
max                         1.000000                     1.000000

       reservation_status_Check-Out  reservation_status_No-Show    label
count                   13115.00000                13115.000000  13115.0
mean                        0.85162                    0.017156      5.0
std                         0.35549                    0.129857      0.0
min                         0.00000                    0.000000      5.0
25%                         1.00000                    0.000000      5.0
50%                         1.00000                    0.000000      5.0
75%                         1.00000                    0.000000      5.0
max                         1.00000                    1.000000      5.0

[8 rows x 48 columns]
```

As we can see here, the silhouette score that we got is very low. An explanation for this is because the values of the features are very similar that the algorithm had difficulties to handle those samples and to separate them into groups that have the same behavior.