**Московский государственный технический университет им. Н.Э. Баумана**

Кафедра «Системы обработки информации и управления»

**Рубежный контроль №1**

**по дисциплине**

**«Методы машинного обучения»**

Выполнил:

Хотин П.Ю.

ИУ5-24М

Москва, 2020 год

import numpy as np  
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt

data = pd.read\_csv('sample\_data/hotel\_bookings.csv', sep=",")  
print(data.shape)

(119390, 32)

import os

os.getcwd()

'/content'

total\_count = data.shape[0]  
cat\_cols = []  
for col in data.columns:  
 # Количество пустых значений   
 temp\_null\_count = data[data[col].isnull()].shape[0]  
 dt = str(data[col].dtype)  
 if temp\_null\_count>0 and (dt=='object'):  
 cat\_cols.append(col)  
 temp\_perc = round((temp\_null\_count / total\_count) \* 100.0, 2)  
 print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp\_null\_count, temp\_perc))

Колонка country. Тип данных object. Количество пустых значений 488, 0.41%.

print('Всего строк: {}'.format(total\_count))

Всего строк: 119390

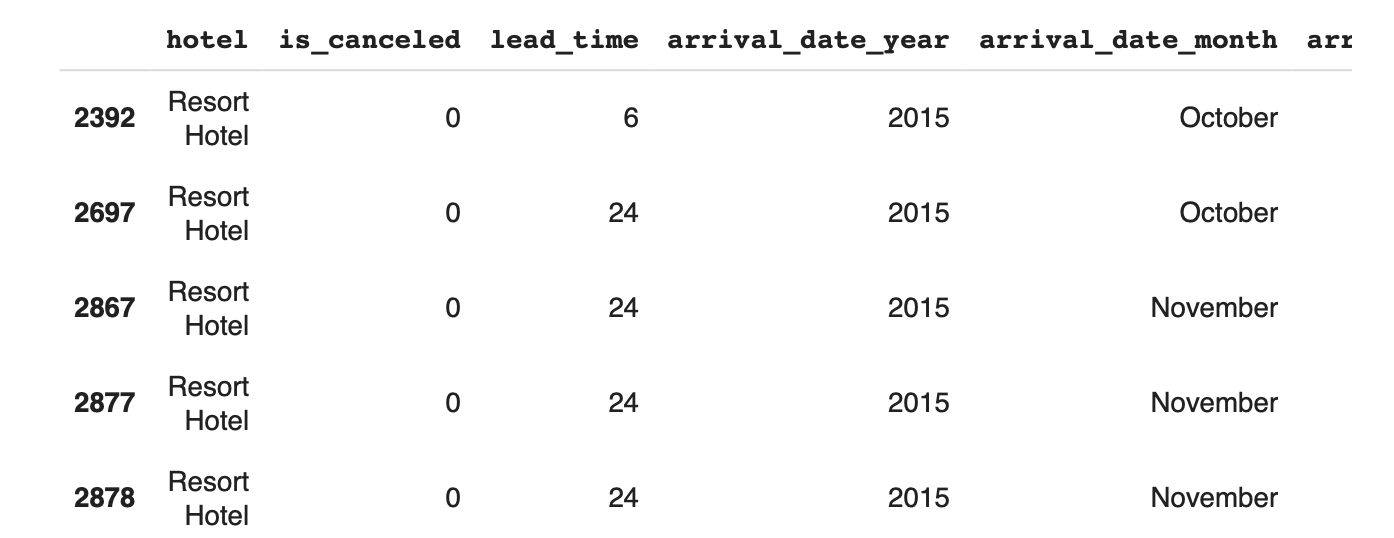
# Удаление колонок, содержащих пустые значения  
data\_new\_1 = data.dropna(axis=1, how='any')  
(data.shape, data\_new\_1.shape)

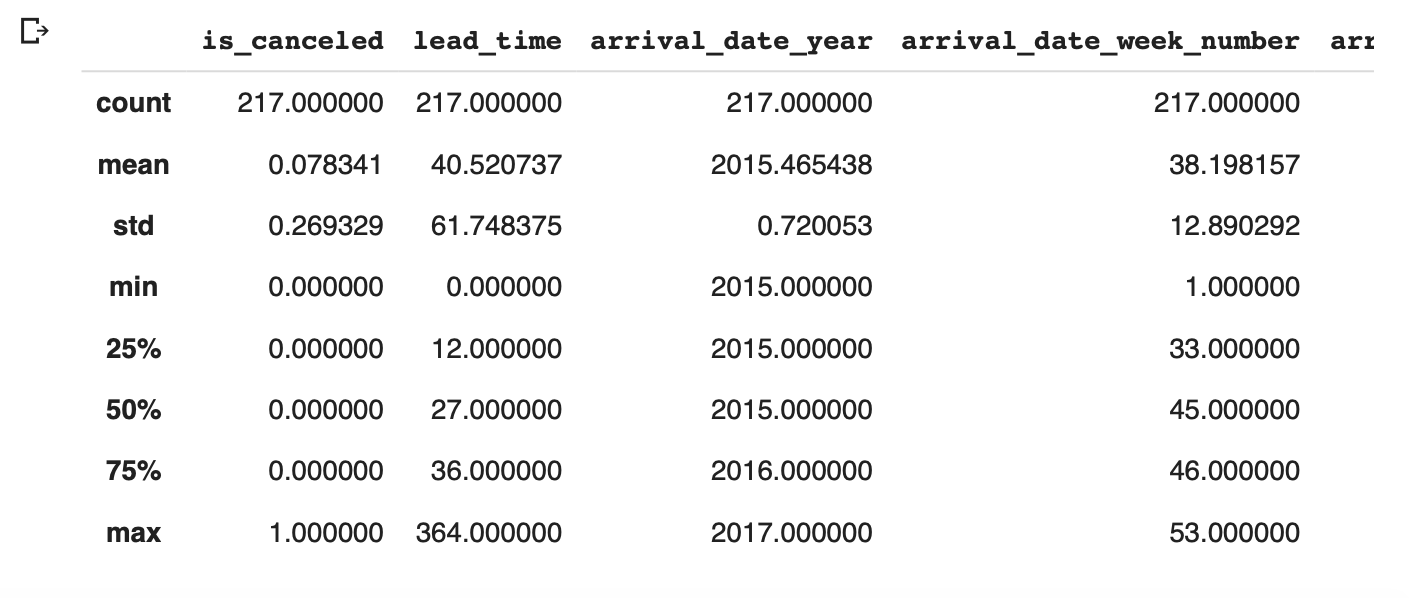
((119390, 32), (119390, 28))

Удалим строки, содержащие null значения

data\_new\_2 = data.dropna(axis=0, how='any')  
(data.shape, data\_new\_2.shape)

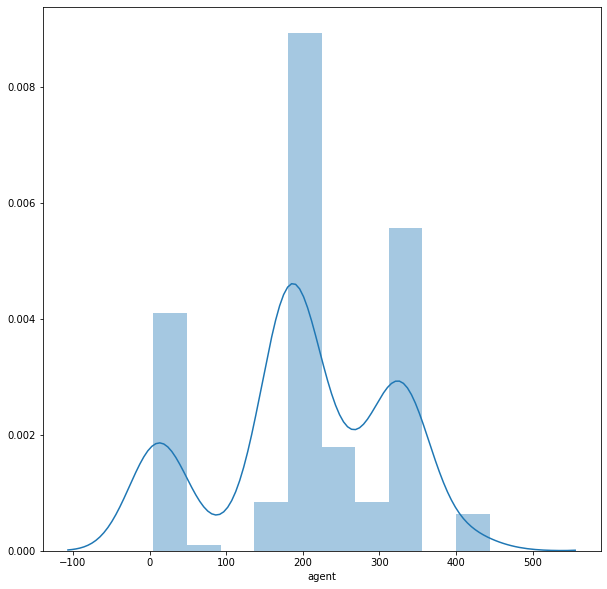
((119390, 32), (217, 32))

data\_new\_2.head()

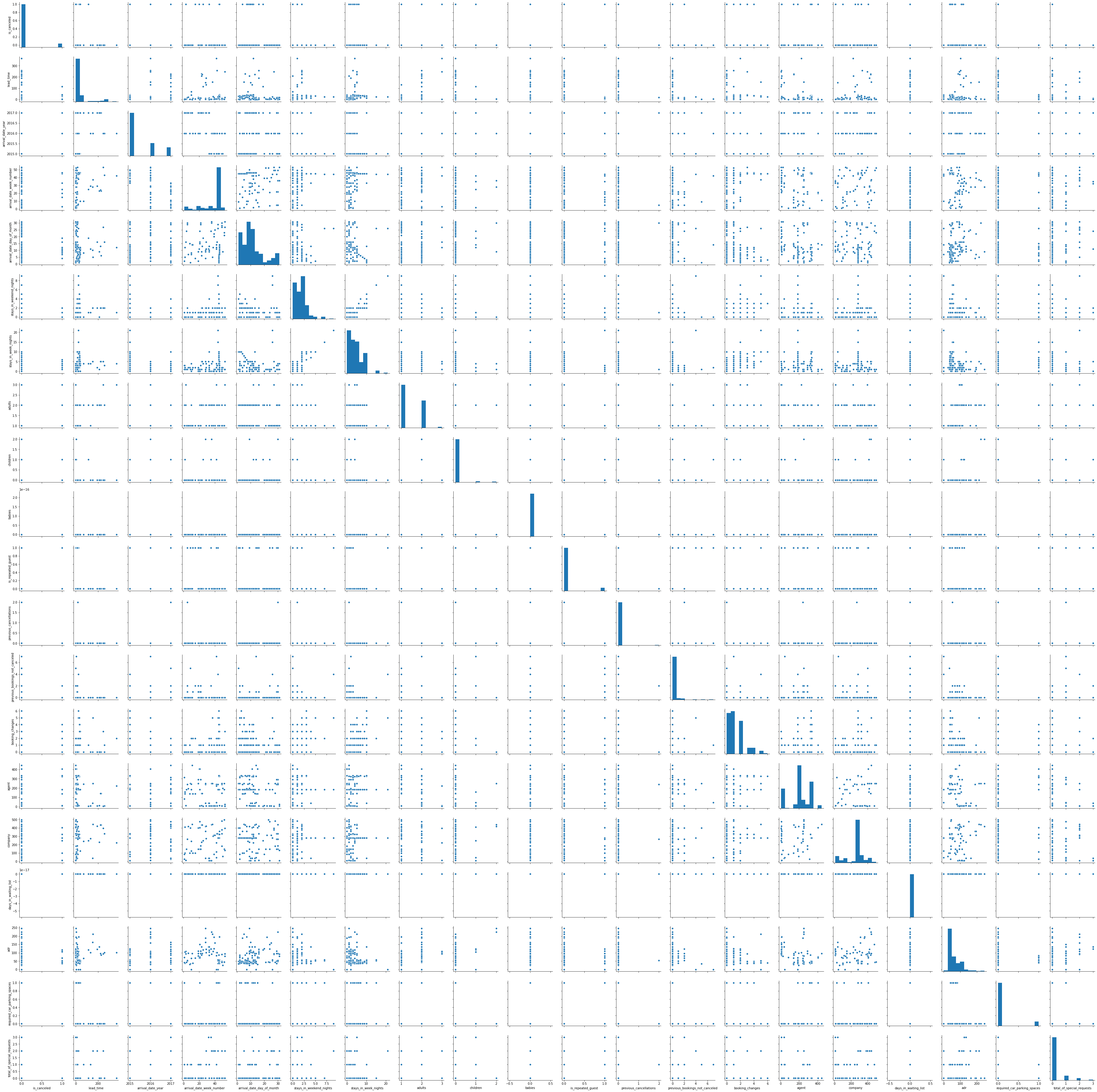
data\_new\_2.describe()

Оценим плотность вероятности распределения данных

fig, ax = plt.subplots(figsize=(10,10))   
sns.distplot(data\_new\_2['agent'])

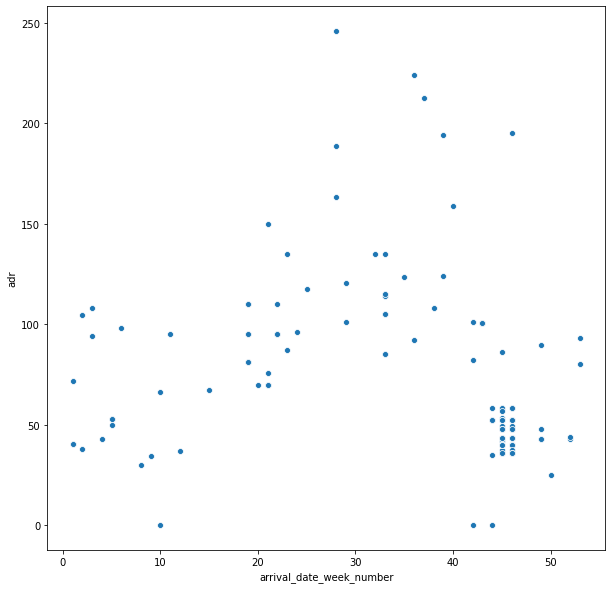
<matplotlib.axes.\_subplots.AxesSubplot at 0x7f8469858240>

sns.pairplot(data\_new\_2)

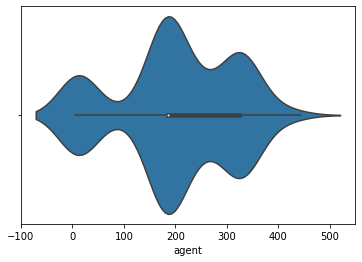
<seaborn.axisgrid.PairGrid at 0x7f84736635c0>

Находим почти линейную зависимость между значениями двух колонок: arrival\_date\_week\_number и adr

fig, ax = plt.subplots(figsize=(10,10))   
sns.scatterplot(ax=ax, x='arrival\_date\_week\_number', y='adr', data=data\_new\_2)

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f8469608da0>

sns.violinplot(x=data\_new\_2['agent'])

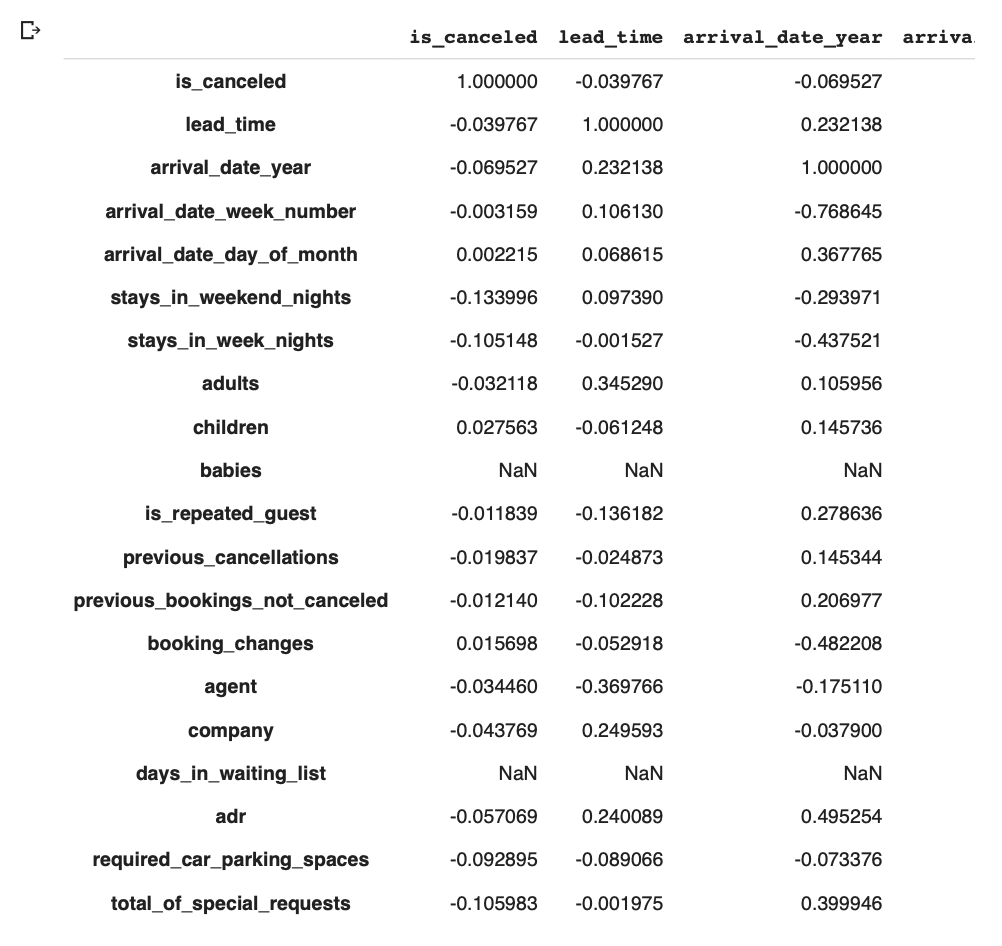
<matplotlib.axes.\_subplots.AxesSubplot at 0x7f846951efd0>

Из приведенных графиков видно, что violinplot действительно показывает распределение плотности.

**Корреляционный анализ**

Построим корреляционную матрицу

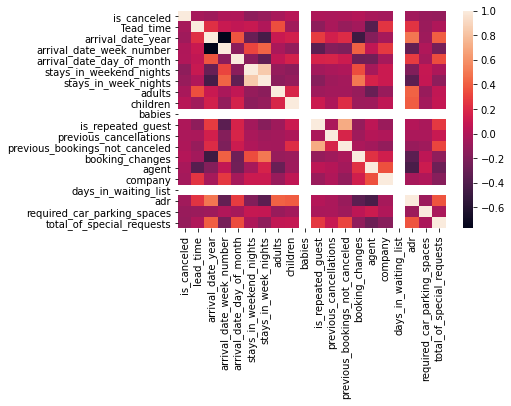
data\_new\_2.corr()



Также построим матрицу корреляций по Пирсону  
Так как значений довольно много, выберем матрицу без подписания числовых значений

sns.heatmap(data\_new\_2.corr())

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f846799fa20>



В примере тепловая карта помогает определить сильную корреляцию, например, между признаками *stays\_in\_week\_nights* и *stays\_in\_weekeend\_nights*, следовательно только один из этих признаков можно включать в модель.