Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа №1 по дисциплине**

**«Методы машинного обучения»**

**По теме «Создание истории о данных»**

**ИСПОЛНИТЕЛЬ:**

Позняк А.А.

Группа ИУ5-22M

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2023 г.

Москва 2023

Оглавление

[**Задание:** 2](#_Toc137065394)

Задание

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.

На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.

Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.

Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.

История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Сформировать отчет и разместить его в своем репозитории на github.

Текст программы

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv('Hotel\_Reservations.csv')

data.head()

data.shape #Размер датасета - 36275 строк, 19 колонок

data.info()

data.isnull().sum()

data['booking\_status'].value\_counts()

# Основные статистические характеристки набора данных

data.describe()

data['booking\_status'].unique()

plt.figure(figsize=(15,10))

sns.heatmap(data.corr(), cmap = "RdPu", annot=True, linewidth=3)

plt.figure(figsize=(15, 10))

sns.countplot(x="room\_type\_reserved", data=data, order = data['room\_type\_reserved'].value\_counts().index)

plt.xticks(rotation=90)

test = data[['arrival\_year', 'no\_of\_adults', 'no\_of\_children', 'repeated\_guest' ]]

data\_by\_year = test.groupby(by = 'arrival\_year').sum()

data\_by\_year

data\_by\_year=data\_by\_year.apply(lambda x : x.astype("int"))

data\_by\_year.plot.line(figsize=(10,10), grid="on");

plt.ylabel("Number of guests");

test = data[['arrival\_month', 'no\_of\_adults', 'no\_of\_children', 'repeated\_guest' ]]

data\_by\_month = test.groupby(by = 'arrival\_month').sum()

data\_by\_month

data\_by\_month=data\_by\_month.apply(lambda x : x.astype("int"))

data\_by\_month.plot.line(figsize=(10,10), grid="on");

plt.ylabel("Number of guests");

data = pd.DataFrame([data['avg\_price\_per\_room']]).T

q = data.quantile(0.90)

data = data[data < q]

plt.figure(figsize=(12,8))

colors = sns.color\_palette("Set3", len(data))

ax = sns.boxplot(data=data, orient='h', palette=colors)

ax.set\_title(label='Distribution of Room', fontsize=20)

plt.show()

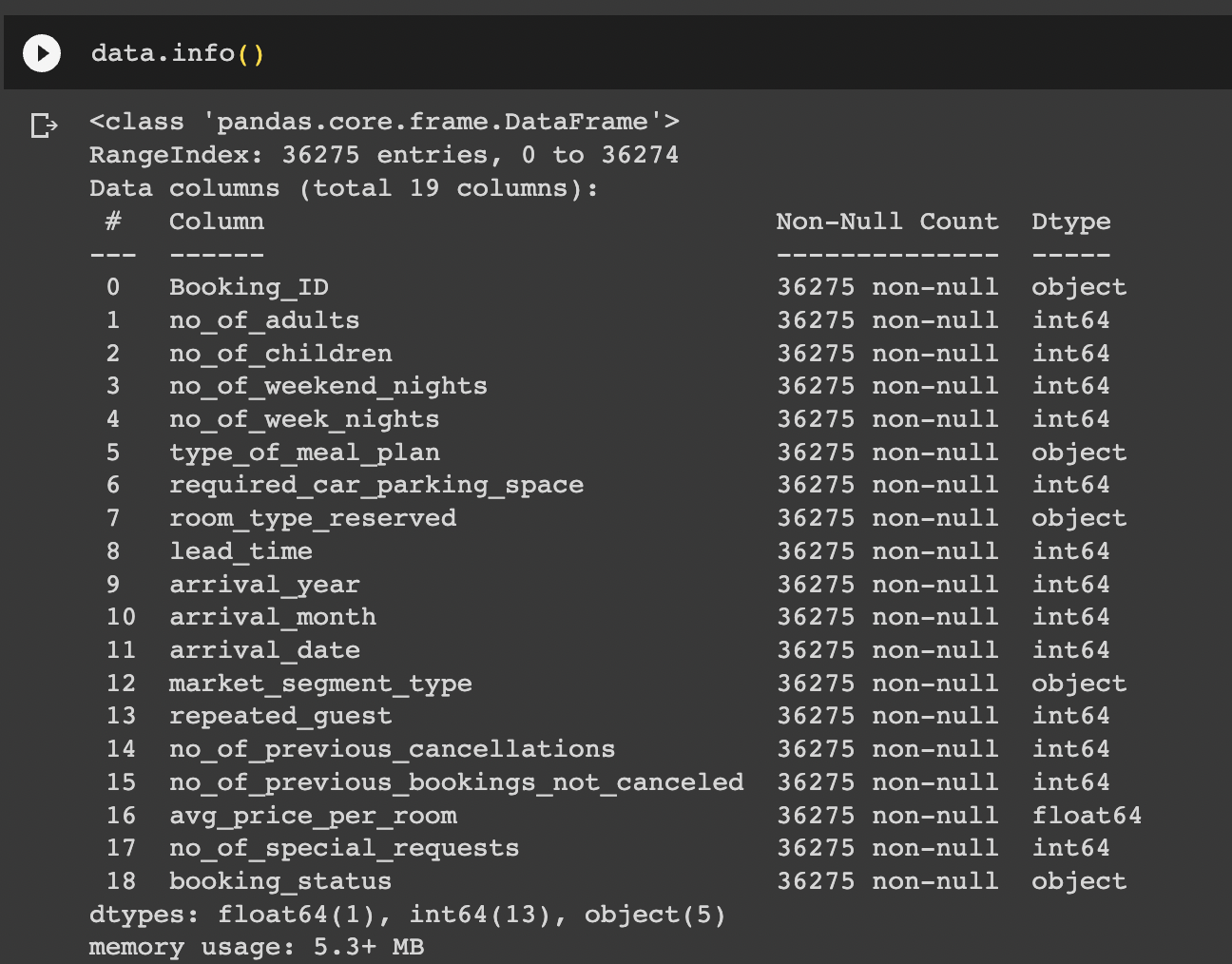
colors = sns.color\_palette("Set1", 2)

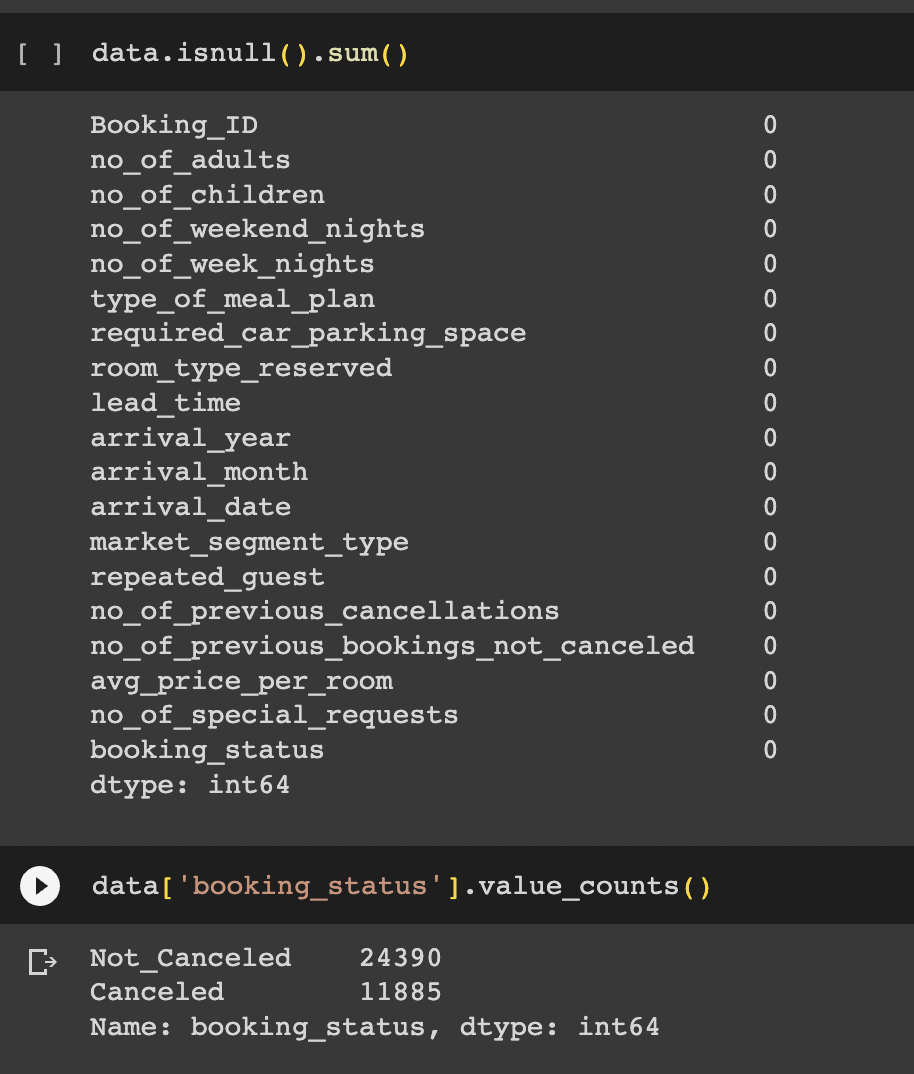
Экранные формы с примерами выполнения программы

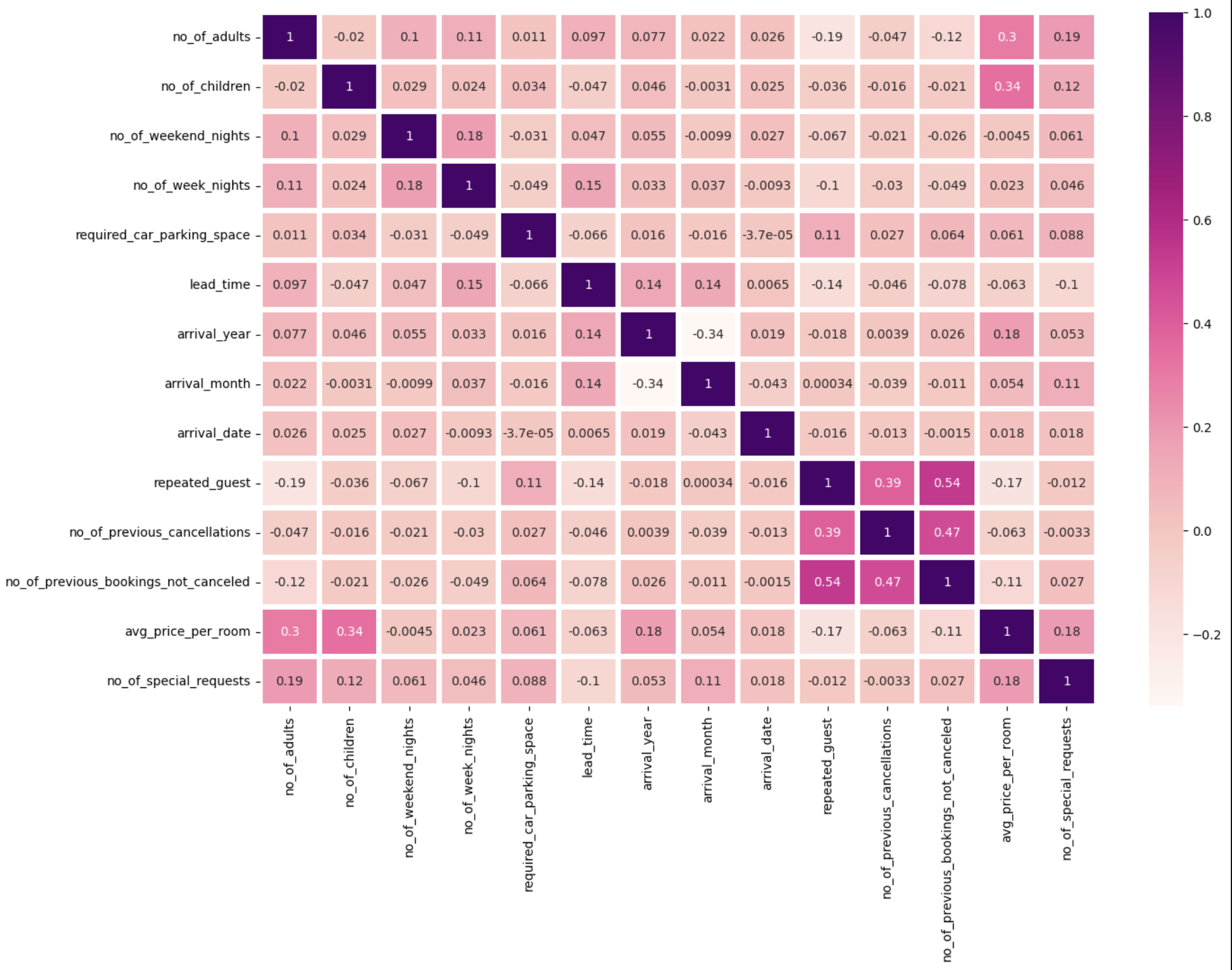
Вывод части даатсета:



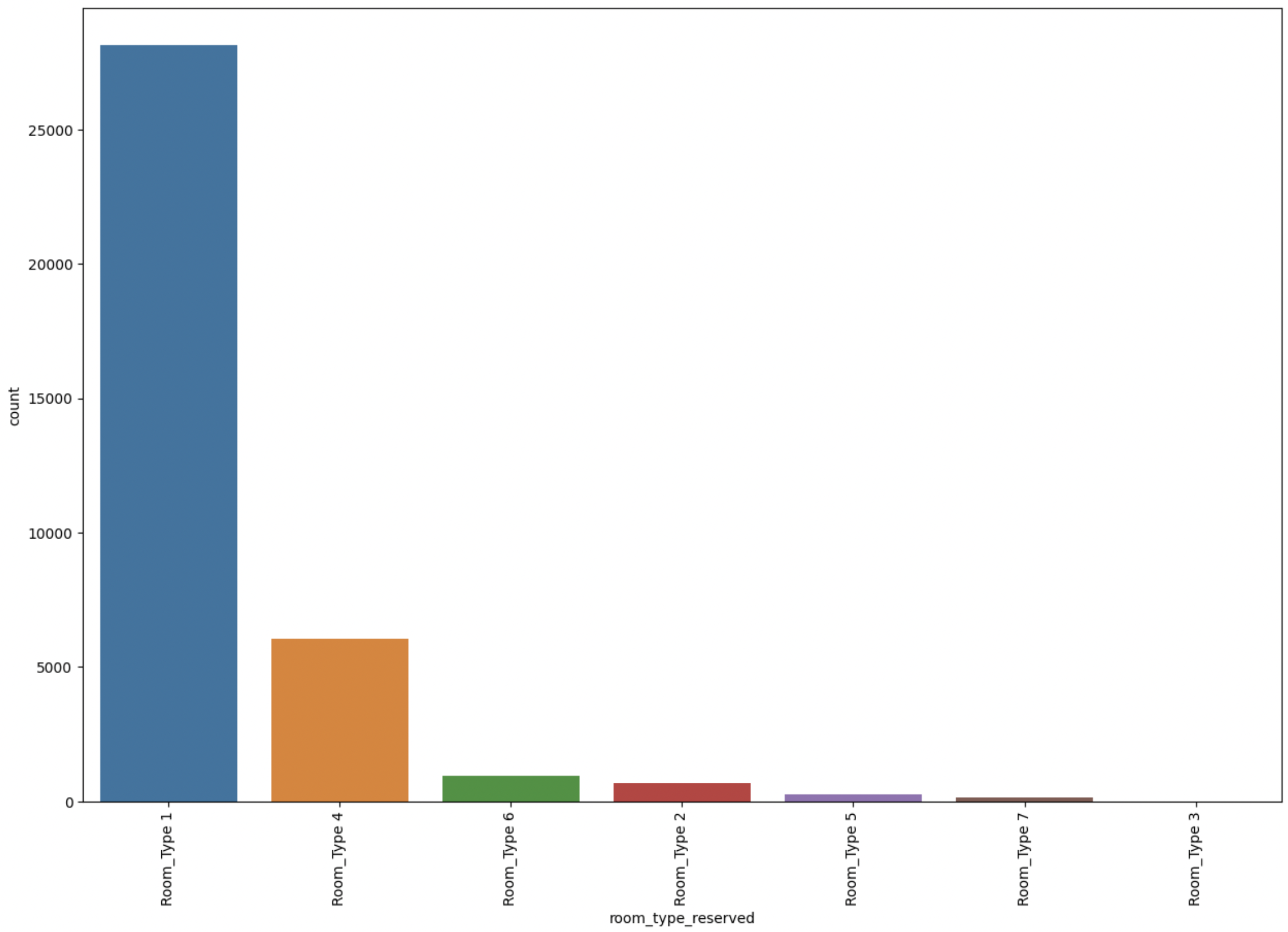
Информация о датасете:



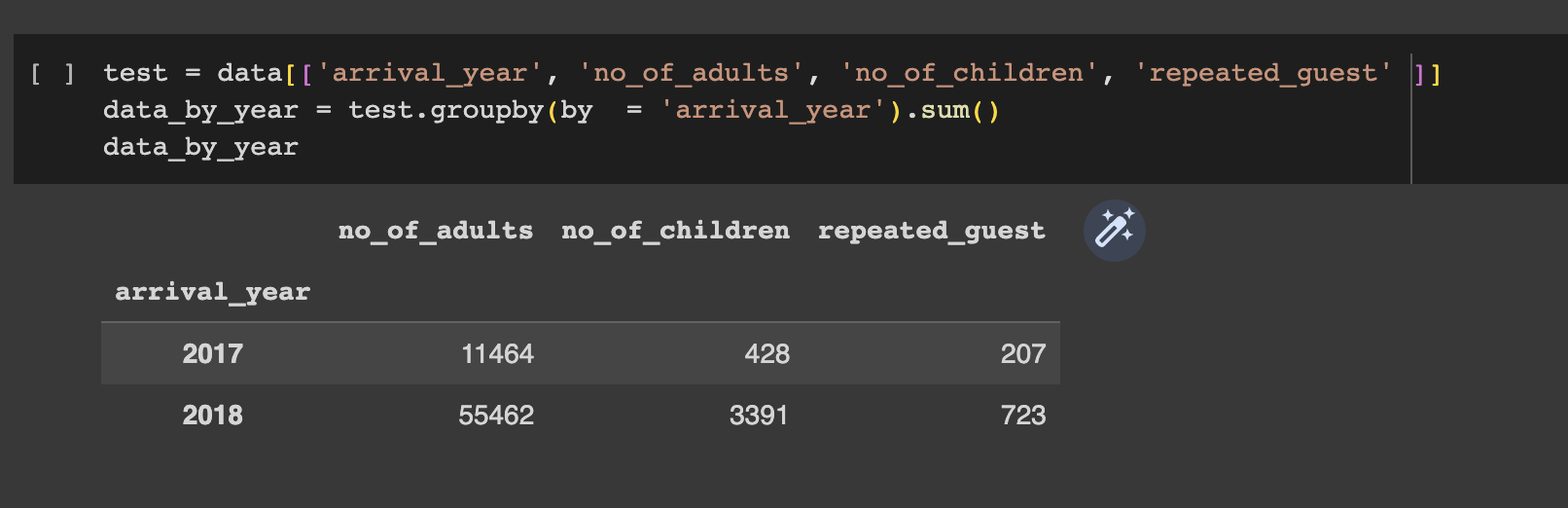
 Построим тепловую карту:

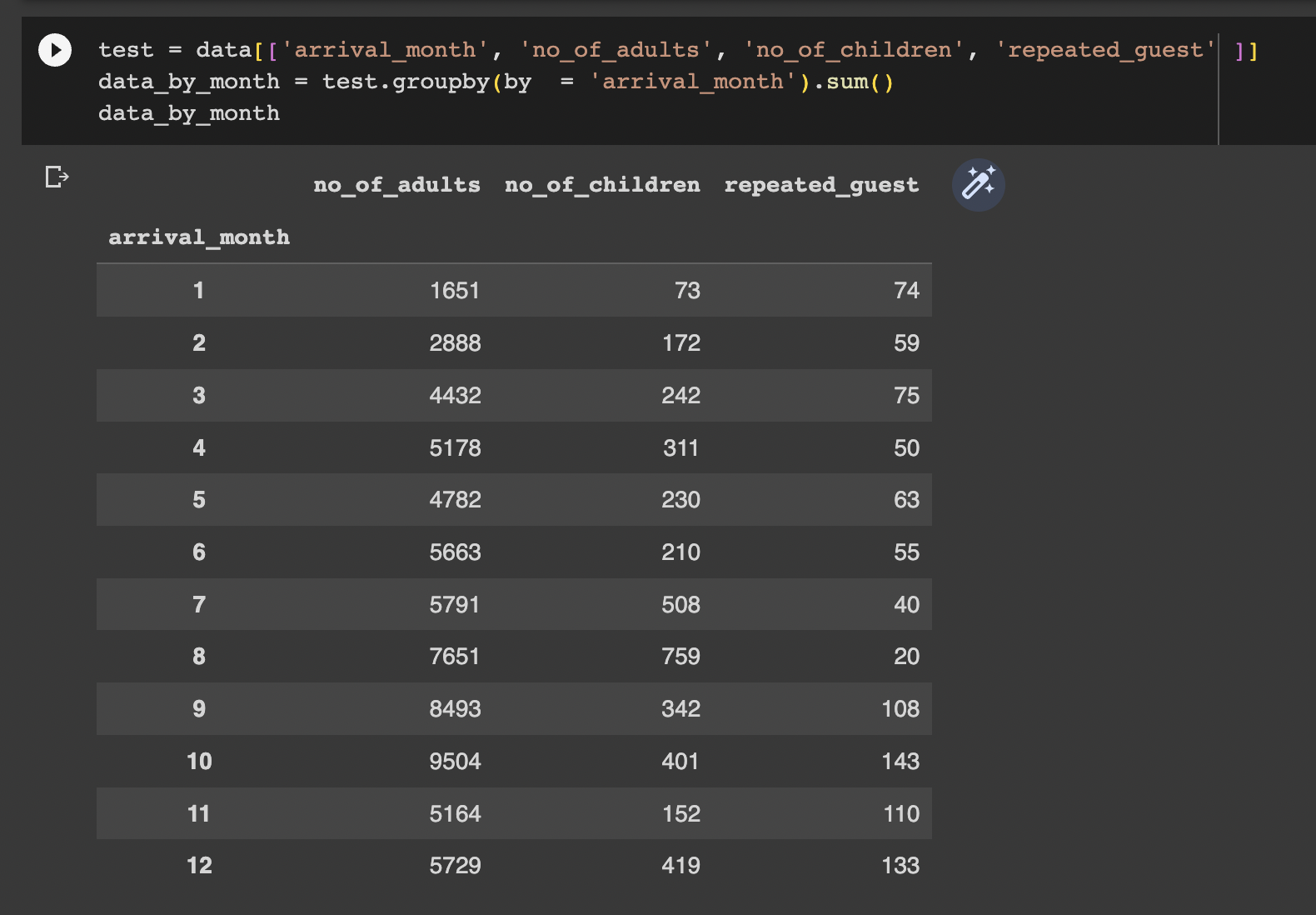
Наблюдаем корреляцию между количеством повторяющихся гостей и каличеством отмененных и неотмененных бронирований, что логично. Также видна зависимость между колчиеством гостей и ценой за номер, причем средняя цена больше зависит от количества детей, чем от количества взрослых.

Построим гистограмму:

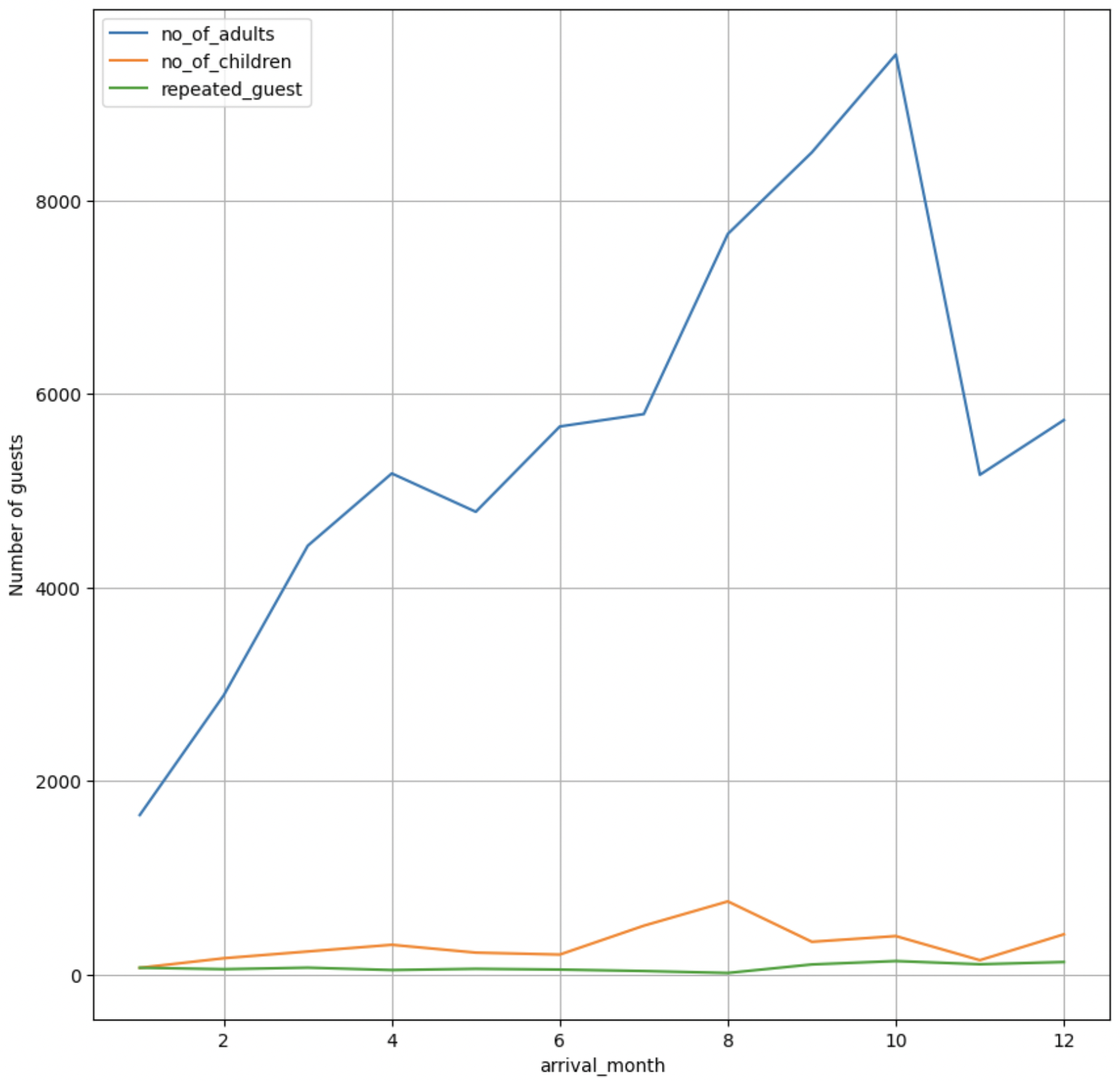


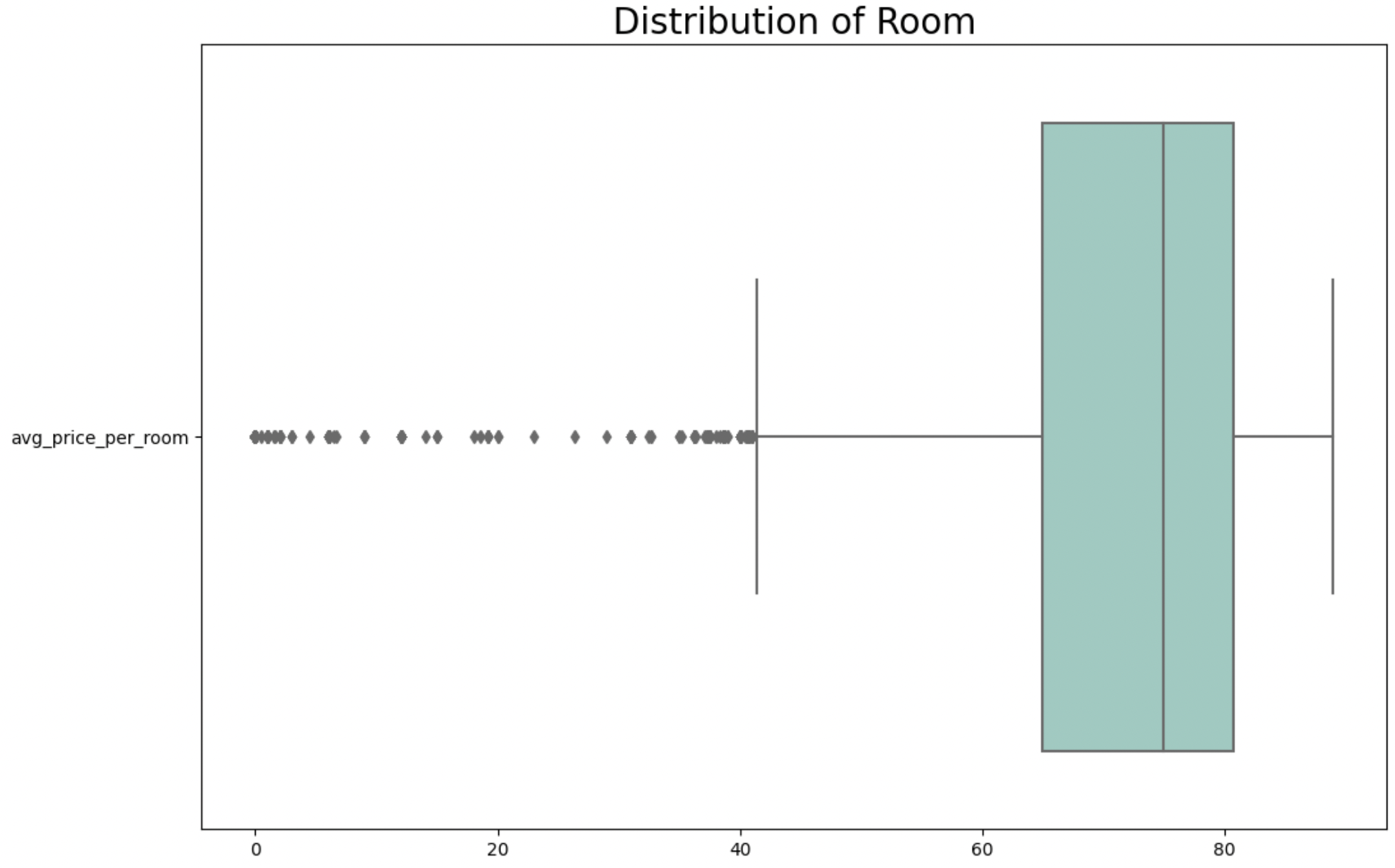
Наиболее популярным типом номеров являются комнаты первого типа, а наименее популярны - комнаты третьего типа.





Исходя из графика ниже, можно сказать, что у детей наиболее популярное время для отдыха – август, а у взрослых – октябрь.



Большинство цен больше 40 долларов за номер:

Более дорогие номера чаще отменяют, при этом очень дешевые номера тоже часто отменяют:

