**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: Исследование набора данных

|  |  |  |
| --- | --- | --- |
| Студент гр. |  | Юрин А.А. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Выбор набора данных для дальнейшего исследования алгоритмов кластеризации и классификации. Получение практических навыков работы с инструментами анализа и визуализации данных.

**Постановка задачи.**

1. Создать Jupyter Notebook
2. Выбрать исследуемый набор данных из предложенных источников
3. Описать выбранный набор данных, а именно:
   1. Предметную область, источник данных, характер данных;
   2. Атрибуты, их тип, что они обозначают;
   3. Задачу анализу.
4. Для каждого атрибута:
   1. Определить среднее значение, СКО;
   2. Построить гистограмму распределения значений, определить наличие выбросов;
   3. Определить наличие пропущенных значений и их количество;
   4. Предложить вариант обработки пропущенных значений.
5. Определить:
   1. Высоко коррелированные атрибуты, характер корреляции;
   2. Не коррелированные атрибуты;
   3. Графики рассеивания.
6. Проанализировать полученные результаты.

**Выполнение работы.**

Для исследования был выбран синтетический набор данных для задач прогнозирования и классификации. Проанализированы признаки, которые содержит набор (см. приложение А).

Далее прочитаны данные, обработаны категориальные типы признаков type\_of\_meal\_plan, room\_type\_reserved, market\_segment\_type, booking\_status.

С помощью средств pandas.df.describe() и seaborn.histplot() вычислены основные параметры признаков и построены гистограммы распределения соответственно. Из гистограмм видно, что набор данных не содержит выбросов и пропусков.

Применив seaborn.heatmap() были выделены коррелированные атрибуты и выдвинуты основные предположения для предобработки, а именно:

1. Существует умеренная положительная корреляция между повторными гостями и количеством предыдущих бронирований, которые не были отменены. Это может свидетельствовать о лояльности таких гостей;
2. Существует умеренная положительная корреляция между количеством предыдущих отмен и количеством предыдущих бронирований, которые не были отменены. Это может означать, что гости, которые часто бронируют, также могут часто отменять бронирования.;
3. Семьи с детьми предпочитают бронировать более дорогие номера, так как им нужно больше мест.;
4. Определенные типы номеров стоят дороже.

Можно заметить, что самый не популярный тип номера под номером 2, она же является самой дешёвой. Люди предпочитают комнаты в среднем и высоком сегменте (типы 3-6). Самая высокая популярность у типа 5. Также можно заметить выбросы — это может указывать на то, что в каждый тип номера могут быть включены услуги, например наличие парковочного места (рис. 1).

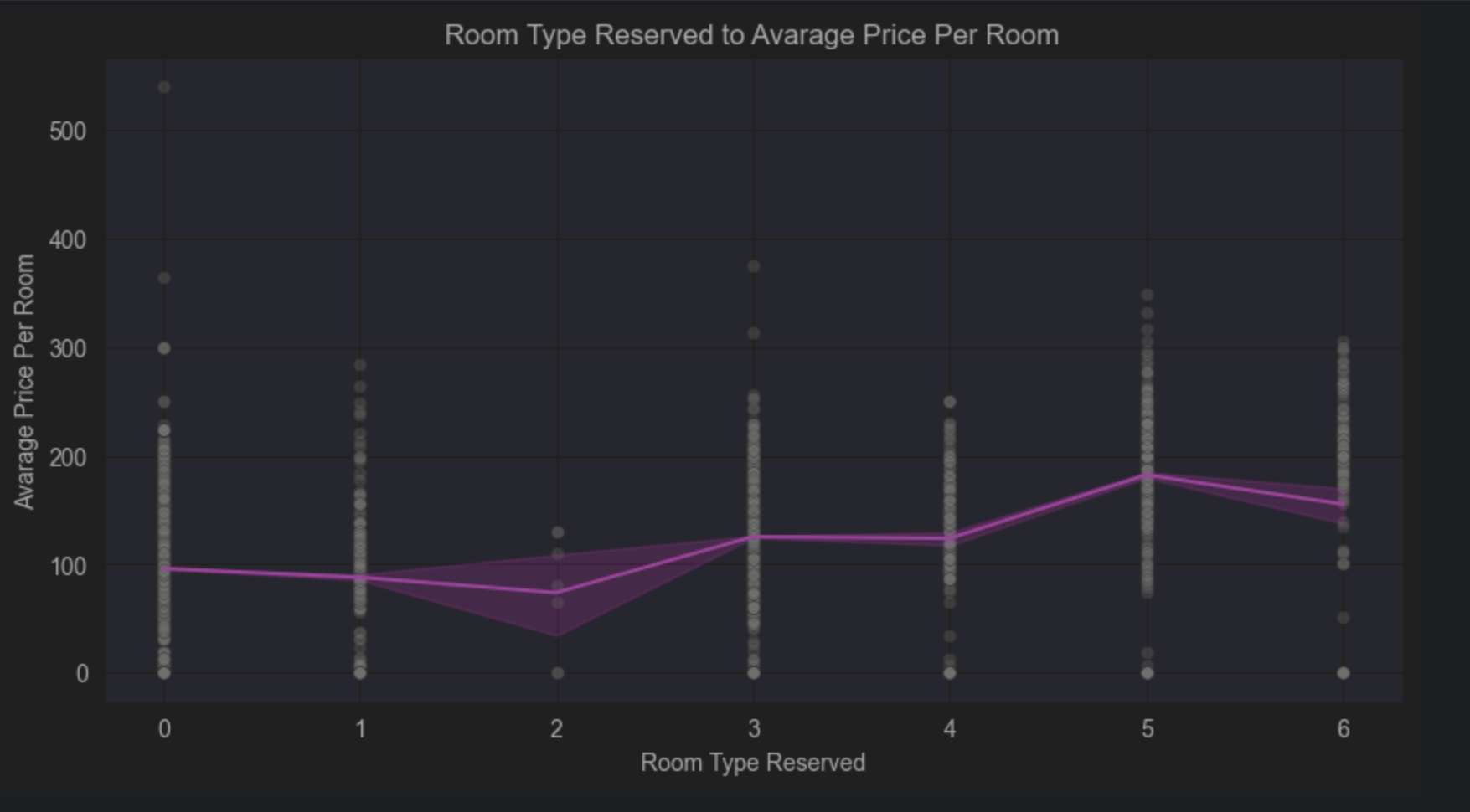


Рисунок 1. Зависимость Типа комнаты от Средней стоимости

Средняя стоимость проживания зависит от сезона. Можно увидеть, что в июне, июле, и также в декабре средняя цена выше (рис. 2).



Рисунок 2. Зависимость между Месяцем прибытия и Средней стоимости

Самая низкая средняя цена отмечена для сегмента 1, после чего цена постепенно увеличивается для сегментов 2, 3 и 4. Это означает, что самым популярным в бронировании будет сегмент авиации и “доп.”. (рис. 3).

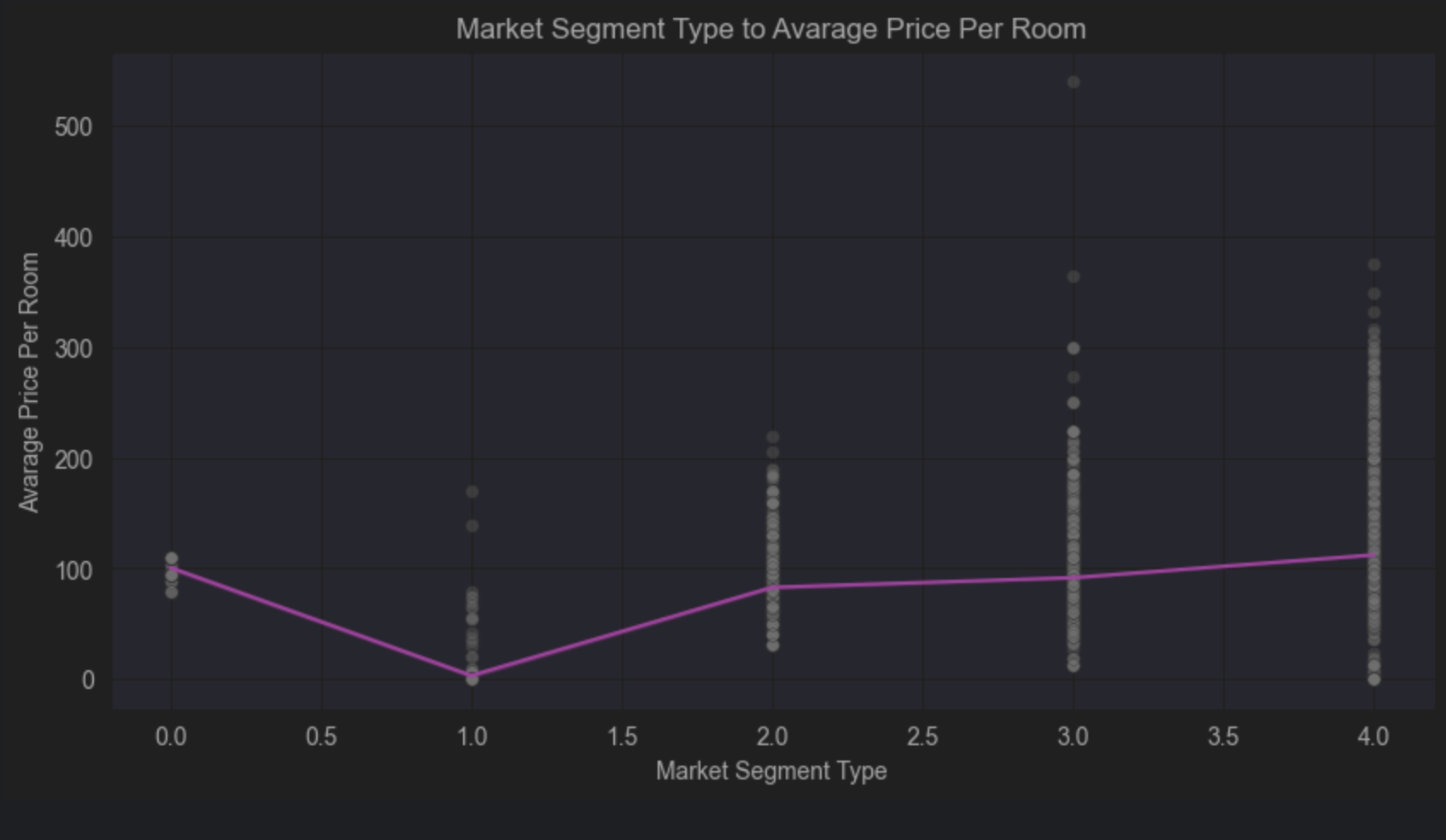


Рисунок 3. Зависимость сегмента рынка от Средней стоимости

**Выводы.**

В ходе работы были выбран набор данных для исследования. Получены практические навыки работы с анализом и визуализацией данных с помощью инструментов numpy, pandas, matplotlib, seaborn.

Приложение А

**Атрибуты набора данных**

