Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Отчет по лабораторной работе №1

«Разведочный анализ данных»

по дисциплине «Технологии машинного обучения»

Выполнил:
студент группы ИУ5Ц-84Б
Папин А.В.
подпись, дата

Проверил:
к.т.н., доц., Ю.Е. Гапанюк
подпись, дата

2024 г.

СОДЕРЖАНИЕ ОТЧЕТА

1. Цель лабораторной работы

Изучение различных методов визуализация данных.

2. Описание задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь https://github.com/ugapanyuk/courses_current/wiki/DSLIST.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/ds/sklearn_datasets.ipynb.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Средства и способы визуализации данных можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/wiki/VISUAL.

В качестве опорного примера для выполнения лабораторной работы можно использовать пример - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/eda/eda_visualization.ipynb.

Дополнительно примеры решения задач, содержащие визуализацию, можно посмотреть в репозитории курса mlcourse.ai - https://github.com/Yorko/mlcourse.ai/wiki/Individual-projects-and-tutorials-(in-Russian)

3. Основные характеристики датасета

Название датасета: Hotels Dataset (Датасет отелей)
Ссылка: https://www.kaggle.com/datasets/raj713335/tbo-hotels-dataset/data

О датасетах

Этот датасет содержит информацию о 1 000 000+ отелях из различных стран и регионов, таких как их тарифы, отзывы, удобства, местоположение и звездный рейтинг. Данные были собраны из различных источников, таких как веб-сайты отелей, онлайн-агентства по бронированию и платформы отзывов. Датасет может использоваться для различных целей, таких как:

- Исследовательский анализ данных для понимания характеристик и распределения отелей по разным рынкам и сегментам.
- Анализ настроений для извлечения идей из отзывов и рейтингов гостей отеля и выявленидаций для предоставления персонализированных предложений по бронированию отелей на основе предпочтений и поведения пользователя.
- Прогнозирование цен для оценки оптимальных тарифов для отелей на основе спроса, сезонности и конкуренции.
- Классификация для определения типа и категории отелей на основе их характеристик и атрибутов.

Структура данных

Датасет состоит из 16 столбцов и 1 000 000+ строк, где каждая строка представляет собой отель. Столбцы включают в себя:
countyCode: Код страны, к которой принадлежит отель.
countyName: Название страны, к которой принадлежит отель.
cityCode: Код города, где расположен отель.
cityName: Город, где расположен отель.
HotelCode: Уникальный идентификатор каждого отеля.

hotel_name: Название отеля.

HotelRating: Звездный рейтинг отеля от 1 до 5.

Address: Адрес отеля.

Attractions: Достопримечательности рядом с отелем.

Description: Подробное описание отеля.

FaxNumber: Номер факса отеля.

HotelFacilities: Доступные в отеле удобства.

Map: Местоположение отеля в формате GPS (широта и долгота).

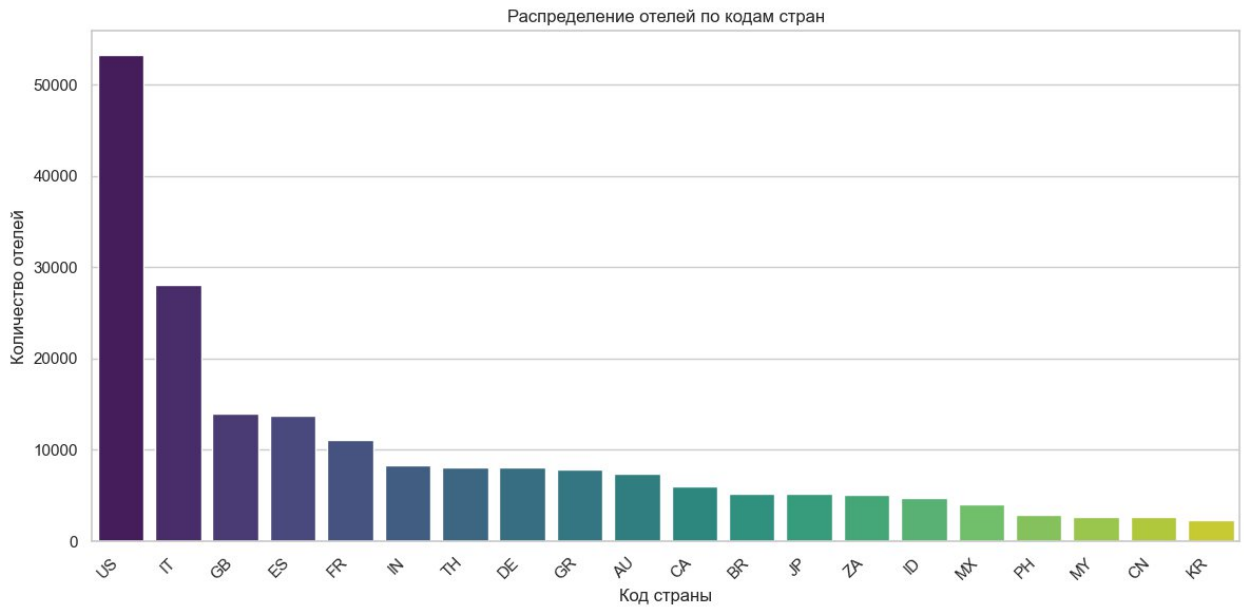
PhoneNumber: Телефонный номер отеля.

PinCode: Почтовый индекс адреса отеля.

HotelWebsiteUrl: Веб-ссылка для бронирования отеля.

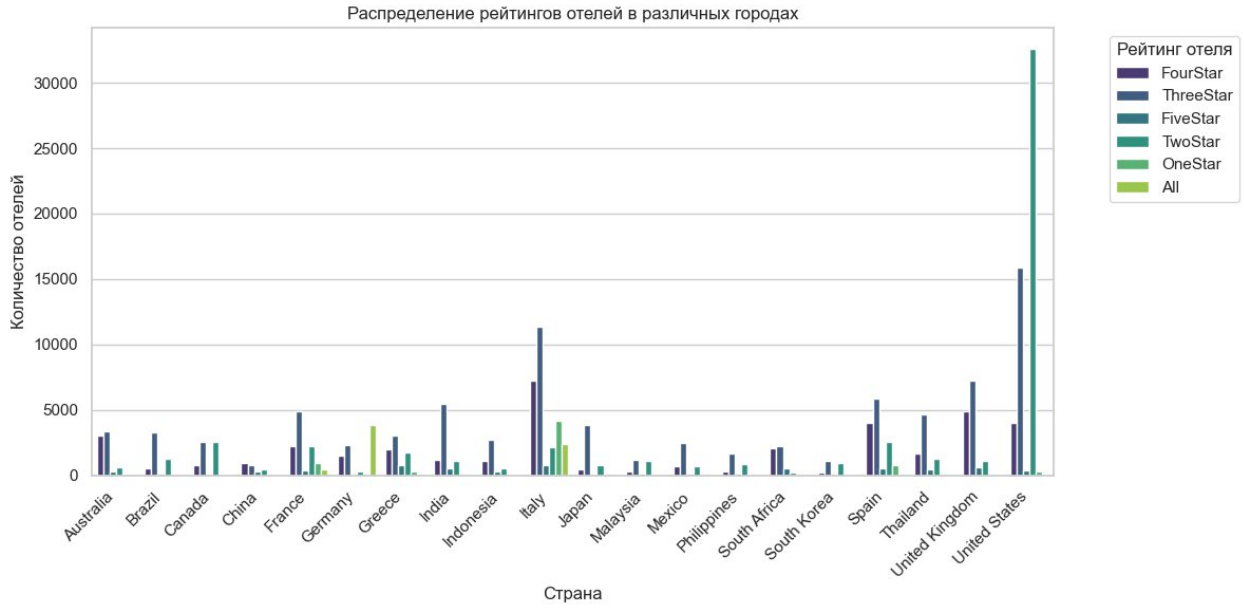
4. Визуальное исследование датасета

4.1. Распределение отелей по кодам стран



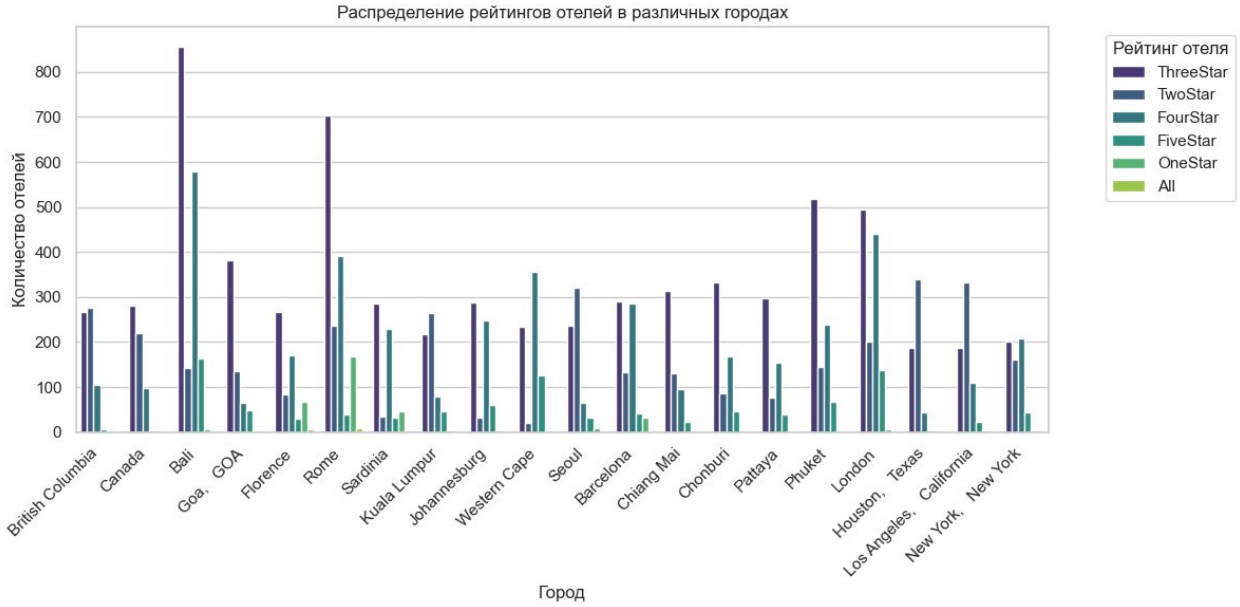
Как и видим, что много отелей в Америке и в Италии

4.2. Распределение рейтингов отелей в различных странах



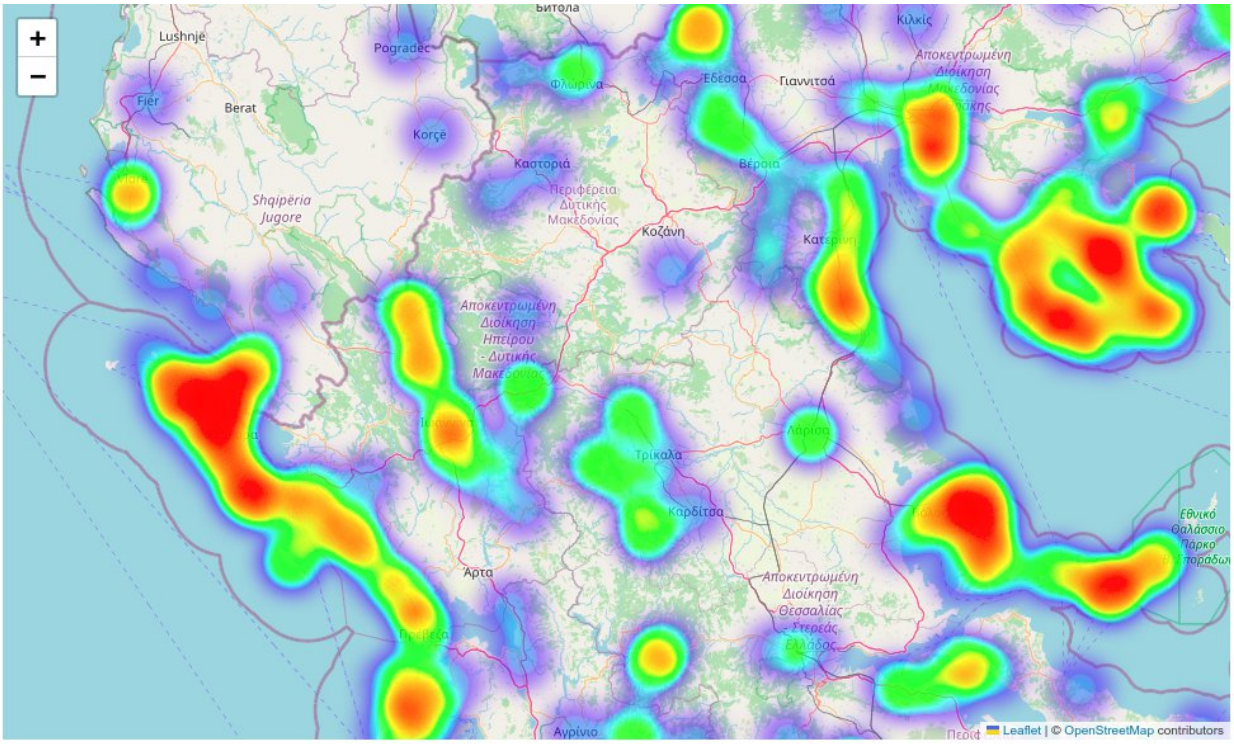
Как и видим, что в Америке слишком много отелей, у которых рейтинг составляет 2 звездочек. Самым лучшим можно дать Италию, которая сумела сохранить высокий рейтинг

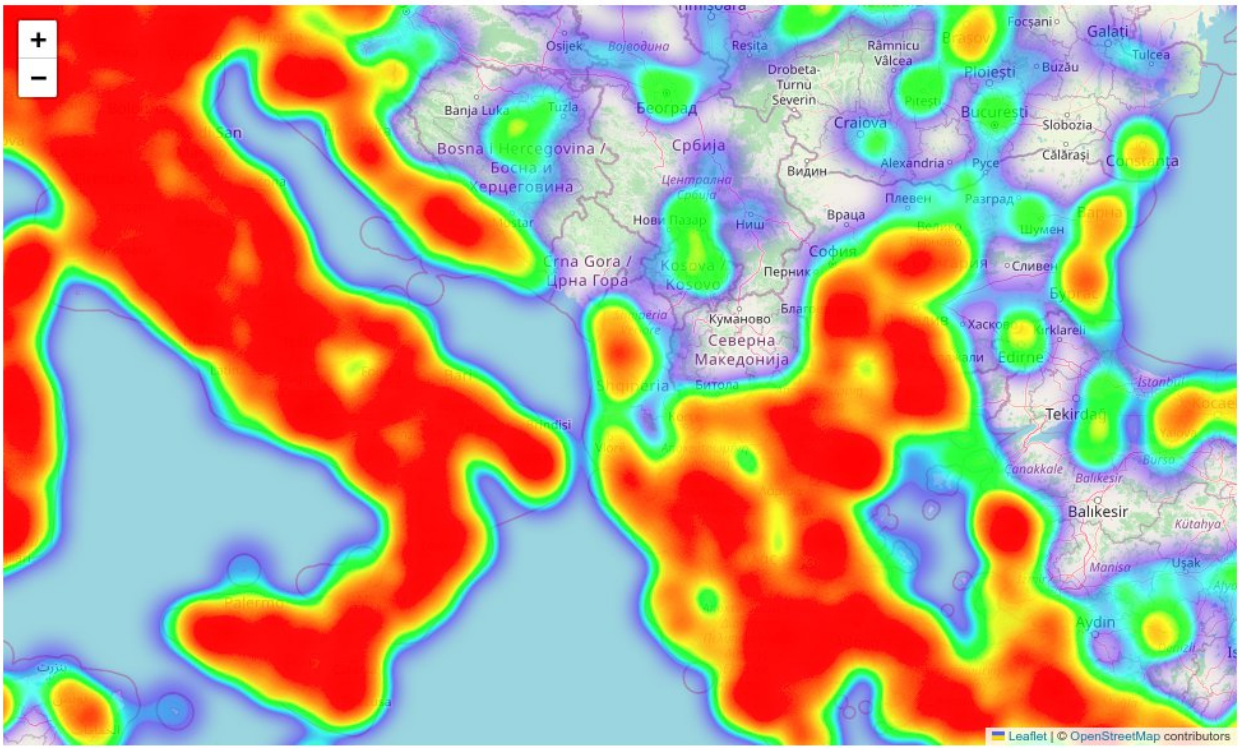
4.3.Распределение рейтингов отелей в различных городах



Как и видим, что отель в Бали и Риме оказался лушчим по сравнению с остальными. Пукхет и Лондон тоже не уступает место и занимает неплохие места

4.4.Визуализация концентрация отелей по географической карте





5. Информация о корреляции признаков

Интересно можно заметить, чем выше значение широты (latitude), тем ниже рейтинга, но не так критично. А если присмотреться, чем выше значение долготы (longitude), тем выше рейтинга. Это можно сделать отсылку на Американского континента, т.к. обычно они находятся в отритательных широтах и полотижельных долготях.

6. Итог

6.1. Анализ данных

- Много отелей можно увидеть в Америке, Италии и другие европейских регионах
- Лучше всего заселиться у Италии, Таиланда или Великобритани, т.к. у них высокие рейтинги
- Самым лучшим городом для отдыха будет: Бали, Рим, Пхукет и Лондом

- По географической карте сразу видно, что много отелей в Еропейских и Американских регонах