МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

|  |  |
| --- | --- |
| **Домашнее задание** | |
|  | |
| по дисциплине | Практикум по программированию |

|  |  |  |  |
| --- | --- | --- | --- |
| Студента(ки) | Сагалбаева Дамира Амангельдыевича | | |
|  | фамилия, имя, отчество полностью | | |
| Курс | 2 | Группа | ФИТ-222 |
| Направление | 02.03.02. Фундаментальная информатика и информационные технологии | | |
|  | код, наименование | | |
| Руководитель | ст.преподаватель | | |
|  | должность, ученая степень, звание | | |
| Саматов А. П. | | |
|  | фамилия, инициалы, дата, подпись | | |
| Выполнил |  | | |
|  | дата, подпись студента(ки) | | |

|  |  |
| --- | --- |
| Итоговый рейтинг |  |

Омск 2023

[ВВЕДЕНИЕ 3](#_heading=h.gjdgxs)

[1.Поиск и загрузка данных 4](#_heading=h.30j0zll)

[2.1 Гистограмма распределения числового признака 5](#_heading=h.1fob9te)

[2.2 Диаграмма «ящик с усами» числового признака 6](#_heading=h.3znysh7)

[2.3 Круговая диаграмма номинативного признака 6](#_heading=h.2et92p0)

[2.4 Тепловая карта 7](#_heading=h.tyjcwt)

[2.5 Диаграмма countplot с группировкой по двум номинативным признакам 8](#_heading=h.3dy6vkm)

[3 Предварительная обработка данных 9](#_heading=h.1t3h5sf)

[ЗАКЛЮЧЕНИЕ 11](#_heading=h.2s8eyo1)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 12](#_heading=h.17dp8vu)

ВВЕДЕНИЕ

Объемы данных сегодня стали настолько обширными, что ручной анализ становится трудозатратным. Тем не менее, этот анализ остается важным для принятия решений, создания статистических отчетов и построения моделей машинного обучения. В процессе курса использовались следующие библиотеки Python:

1. **NumPy**: Это библиотека с открытым исходным кодом, предоставляющая поддержку многомерных массивов (включая матрицы) и высокоуровневых математических функций, специально разработанная для работы с многомерными данными.

2. **Matplotlib**: Эта библиотека предназначена для визуализации данных и позволяет построить как двумерные, так и трехмерные графики. Она играет важную роль в наглядном представлении информации.

3. **SymPy**: Данная библиотека представляет собой инструмент для символьных вычислений в Python. Она полезна при решении задач компьютерной алгебры и предоставляет функциональность для символьных выражений.

4. **SciPy**: Это открытая библиотека, предназначенная для выполнения научных и инженерных расчетов. Она предоставляет множество функций для решения различных задач, связанных с наукой и инженерией.

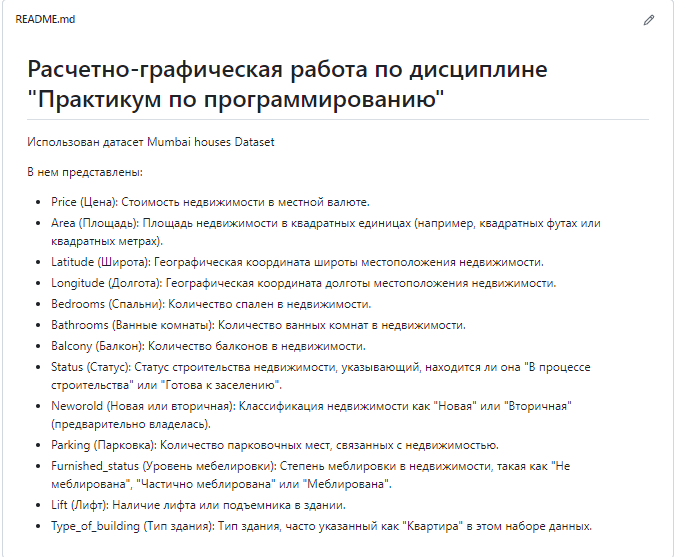
5. **Pandas**: Эта библиотека предоставляет удобные структуры данных и операции для обработки и анализа данных. Pandas часто используется для манипуляций с числовыми таблицами и временными рядами.

6. **Seaborn**: Эта библиотека создана для создания статистических графиков на Python и является дополнением к Matplotlib. Она интегрируется тесно с pandas, обеспечивая удобные средства визуализации статистики.

Эти библиотеки обеспечивают обширные возможности по обработке, анализу и визуализации данных, а также позволяют строить статистику на их основе.

1.Поиск и загрузка данных

Использован датасет Mumbai\_house, специализирующемся на исследовании данных и машинном обучении.



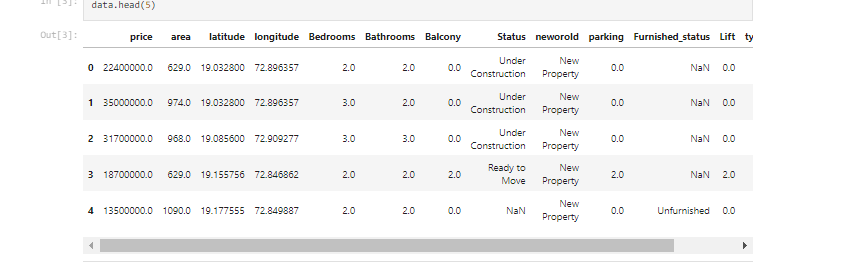
**Рисунок 1 – файл README.md**

Датасет был загружен в ноутбук командой read\_csv() библиотеки pandas.



**Рисунок 2 – загрузка датасета**

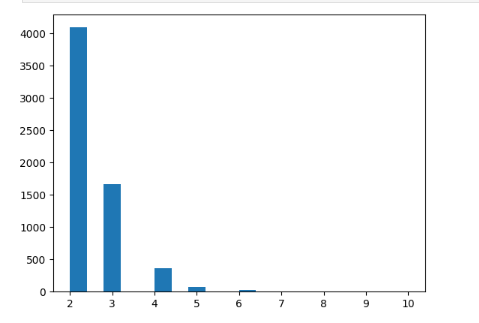
Данный датасет состоит из 6255 строк и 13 столбцов.



**Рисунок 3 – небольшая часть датасета, выведенного в виде таблицы**

2.1 Гистограмма распределения числового признака

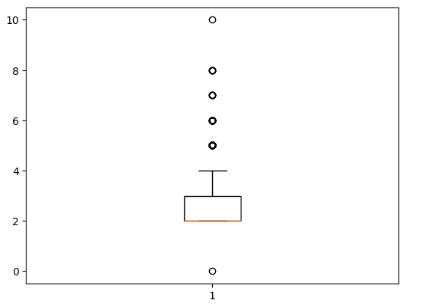
Гистограмма — способ представления табличных данных в графическом виде — в виде столбчатой диаграммы. Количественные соотношения некоторого показателя представлены в виде прямоугольников, площади которых пропорциональны. На гистограмме видно количество спальных комнат.

****

**Рисунок 4 – гистограмма столбца Bedrooms**

2.2 Диаграмма «ящик с усами» числового признака

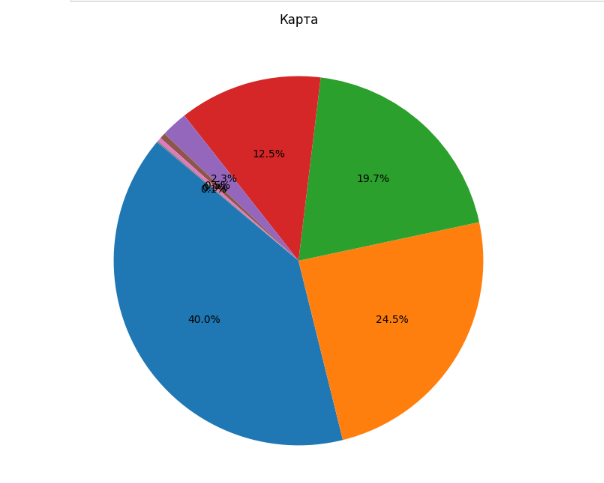
Диаграмма «ящик с усами» — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. 7 Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

****

**Рисунок 5 – Диаграмма «ящик с усами» столбца Bathrooms**

2.3 Круговая диаграмма номинативного признака

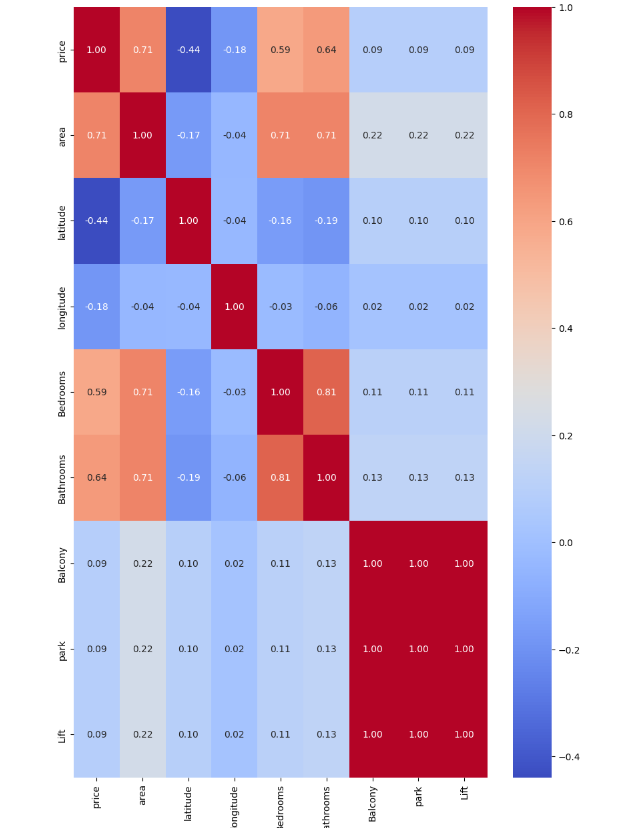
Круговая диаграмма — это круговая статистическая диаграмма, которая разделена на срезы, чтобы проиллюстрировать числовую пропорцию. На круговой диаграмме длина дуги каждого среза пропорциональна величине, 8 которую он представляет. На данной круговой диаграмме видно, балконов чаще всего в квартирах одна штука.



**Рисунок 6 – Круговая диаграмма**

2.4 Тепловая карта

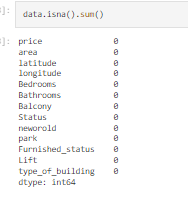
Тепловая карта — графическое представление данных, где индивидуальные значения в таблице отображаются при помощи цвета. На тепловой карте данного датасета можно выявить несколько особенностей, что корреляции между данными есть.



**Рисунок 7 –фрагмент тепловой карты датасета**

3 Предварительная обработка данных

Данные содержат пропуски, и для их заполнения будет использоваться подход с использованием среднего значения и моды, поскольку исходные данные содержат пропущенные значения.



**Рисунок 9 – Проверка на наличие пропусков в таблице**

Были созданы одна переменная:

1. «columns\_to\_replace»: содержит названия колонок, которые имеют пропуски.



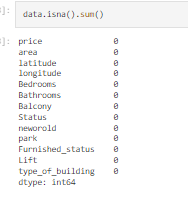
**Рисунок 10 – Создание переменных**

Далее были заменены пропущенные значения на “unknown”:



**Рисунок 11 – Замена пропущенных значений**

Результат:



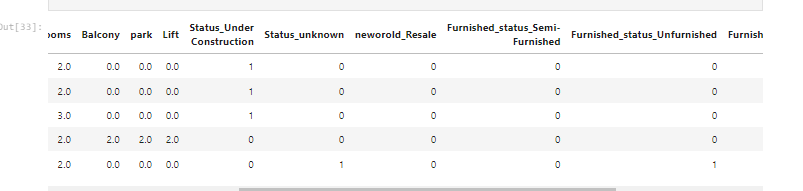
**Рисунок 12 – Результат обработки пропущенных значений**

Также было применено one-hot кодирование, то есть преобразование категориальных переменных в численные путем создания столбцов под каждую категорию и заполнения их значениями 0 и 1 в зависимости от категории каждой строчки.



**Рисунок 13 – Горячее кодирование**

Результат:



**Рисунок 14 – Горячее кодирование**

Обработанные данные были сохранены в формате .csv в той же директории, что и изначальный датасет.



**Рисунок 11 –Экспорт датасета**

ЗАКЛЮЧЕНИЕ

В ходе практики были освоены и применены ключевые библиотеки Python: matplotlib, seaborn, pandas и numpy. Эти инструменты являются фундаментальными для работы с данными и визуализации результатов.

Matplotlib предоставляет возможность создавать разнообразные графики и диаграммы, сделав данные более доступными и интерпретируемыми. Seaborn расширяет возможности визуализации, позволяя создавать более сложные графики, такие как тепловые карты и распределения данных.

Pandas облегчает работу с данными в формате таблицы, обеспечивая удобство анализа и обработки данных. А библиотека NumPy предоставляет мощные средства для математических операций с массивами данных.

Применение этих библиотек позволило успешно решить различные задачи по анализу данных и визуализации результатов. Создание графиков, диаграмм, тепловых карт и распределений способствовало лучшему пониманию структуры данных и выявлению закономерностей.

В общем, использование данных библиотек значительно ускоряет процесс анализа данных, повышает точность выводов и является неотъемлемым инструментом для специалистов, занимающихся анализом данных в среде Python.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. <https://numpy.org/doc/stable/reference/generated/numpy.matrix.html> (дата обращения: 04.11.23).
2. <https://seaborn.pydata.org/installing.html>(дата обращения: 04.10.23).
3. <https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html> (дата обращения: 04.11.23).
4. <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.tight_layout.html> (дата обращения: 04.11.23).