

Государственное бюджетное профессиональное  
образовательное учреждение Московской области  
«Физико-технический колледж»

**Отчёт по работе на 1 интенсиве МДК 05.01:**

Работу выполнил:  
Студент группы № ИСП-22:  
Хусаинов Марат

## **Введение**

В данном отчёте представлены результаты аналитической работы с данными по квартирам в Московской области, Новой Москве и Москве.

## **Цель**

Целью работы является сбор и анализ данных для последующего использования, например, в разработке модели прогнозирования.

## **Задачи**

Для достижения поставленной цели были определены следующие задачи:

- I. **Сбор данных:** Используя открытые источники, собрать исчерпывающий набор данных о квартирах.
- II. **Очистка и подготовка данных:** Провести предварительную обработку данных, включая удаление ненужной информации, дополнение недостающих данных, выявление и устранение аномалий.
- III. **Визуализация и анализ:** представить данные в наглядном виде с помощью как минимум двух инструментов визуализации. Выявить взаимосвязи между данными, определить средние показатели и сделать выводы, необходимые для составления отчёта.
- IV. **Дополнительное задание:** Получить данные о самых дешёвых однокомнатных квартирах в Москве

## **Основная часть:**

Для сбора данных использовалась библиотека Cianparser для Python. С ее помощью была собрана информация о 20 тысячах предложений о продаже квартир.

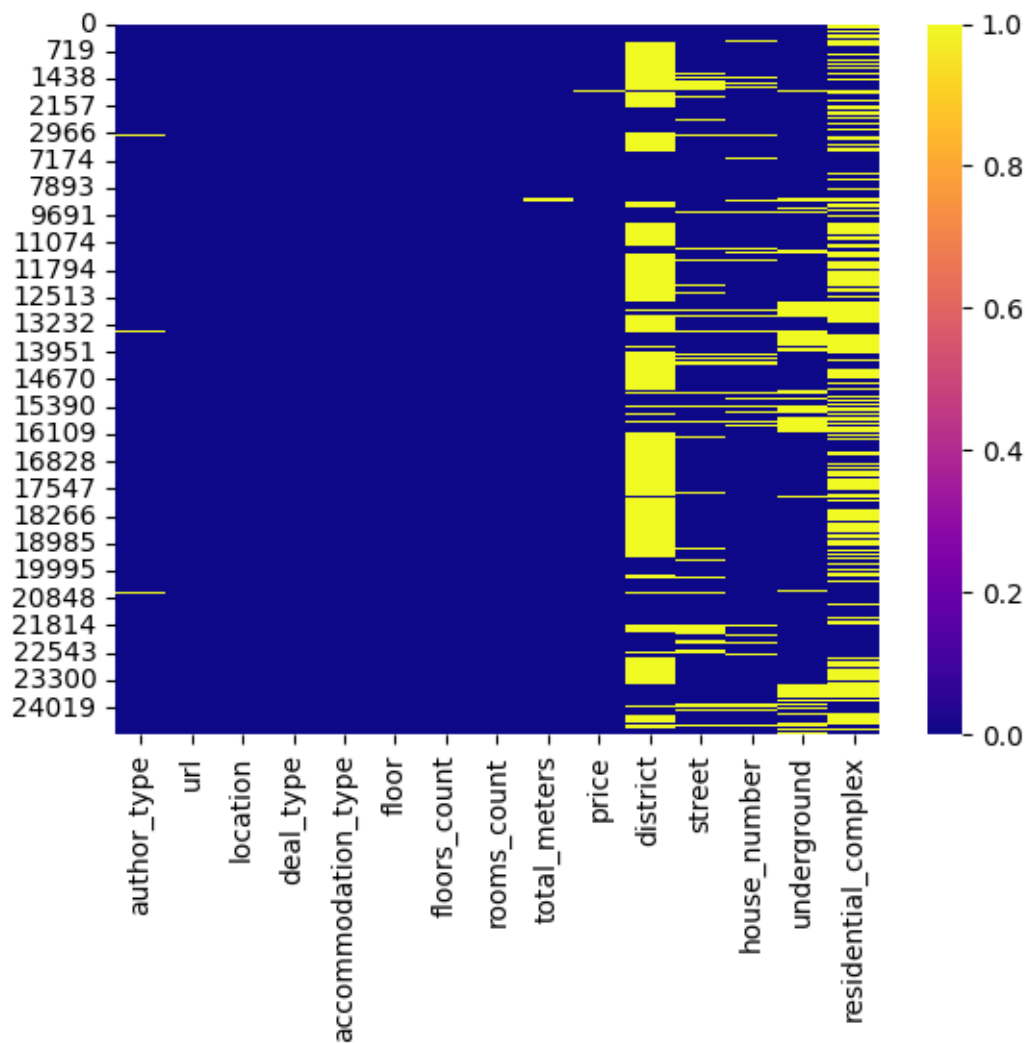
Почти вся остальная работа подготовки данных к анализу была произведена с помощью библиотеки Pandas.

В полученном дата-сете было 16 видов данных:

1. author – имя автора, сразу удалил за ненужностью
2. author\_type – тип автора (собственник, агент)
3. url – ссылка на объявление
4. location - город

5. deal\_type – тип сделки(везде продажа)
6. accommodation\_type – тип помещения(везде квартира)
7. floor - этаж
8. floors\_count – кол-во этажей в здании
9. rooms\_count – кол-во комнат
10. total\_meters - общая площадь
11. price - цена
12. district - район
13. street - улица
14. house\_number – номер дома(удален)
15. underground - ближайшая остановка метро/мцд
16. residential\_complex – жилой комплекс

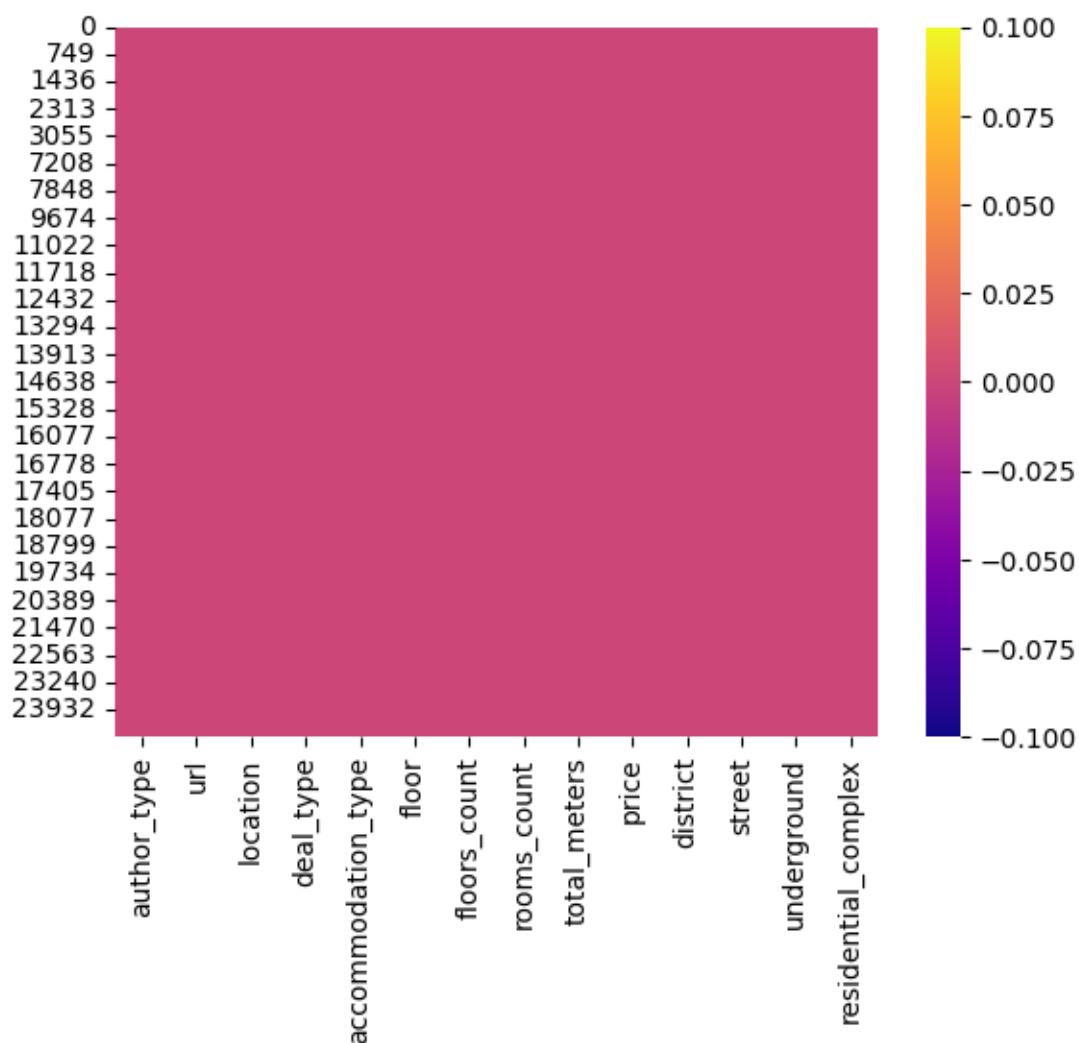
Дальше я приступил к удалению изаполнению пропусков.



Для этого я построил такой график. Из него понятно что больше всего пропусков в колонках «district» и «residential\_complex». Чаше всего они пропущены в строках квартир области. Поэтому пустоты в этой колонке я заполнил значениями из колонки “location”, а пропуски колонки “underground” я заполнил значениями из «district», а пропуски колонки “residential\_complex” я заполнил значениями из «underground».

Так же просто удалил строки с любыми другими пропусками тк их число незначительно.

После этого график стал выглядеть так:



Он показывает что пропусков нет.

Дальше я занялся приведением типов данных, потому что данный всех колонках были типа “object”. Привел к числовым колонки которые должны быть числовыми:

## Приведение типов

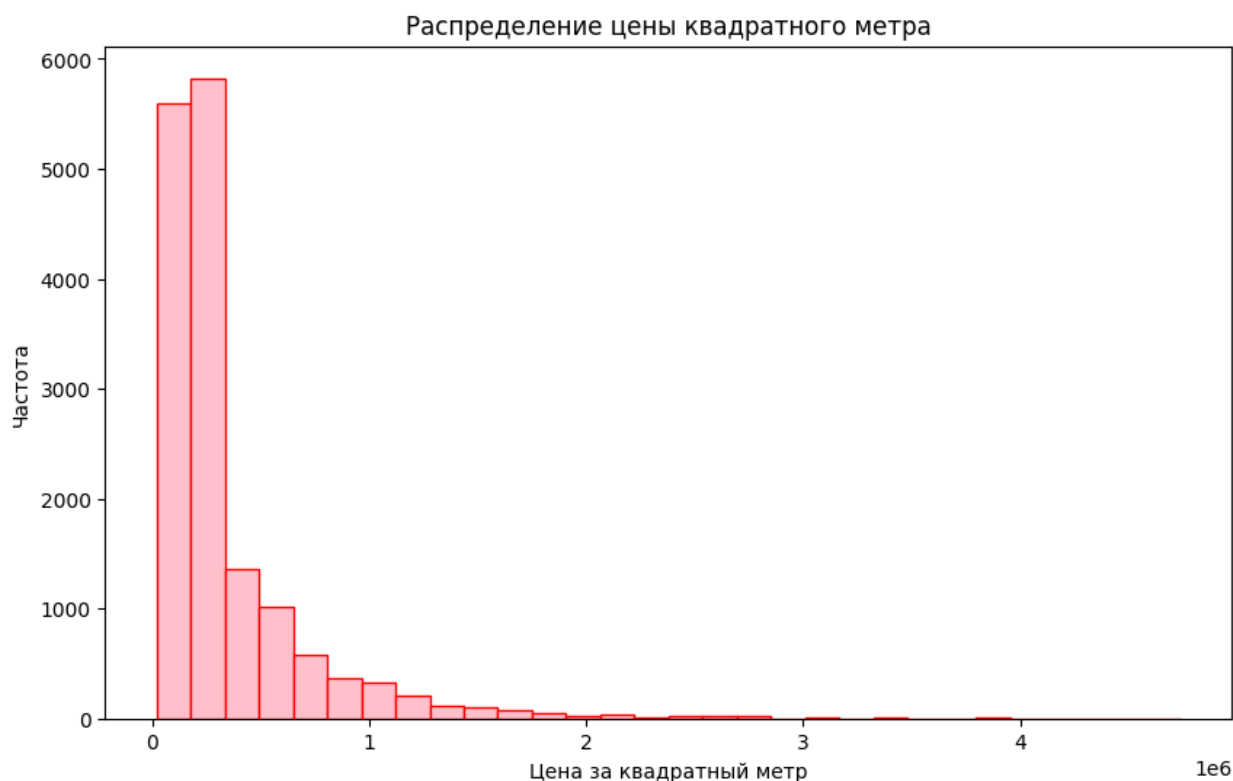
```
In [8]: df['floor'] = pd.to_numeric(df['floor'], errors='coerce')
df['floors_count'] = pd.to_numeric(df['floors_count'], errors='coerce')
df['total_meters'] = pd.to_numeric(df['total_meters'], errors='coerce')
df['price'] = pd.to_numeric(df['price'], errors='coerce')
df['rooms_count'] = pd.to_numeric(df['rooms_count'], errors='coerce')
```

Перевод данных которые должны быть численными

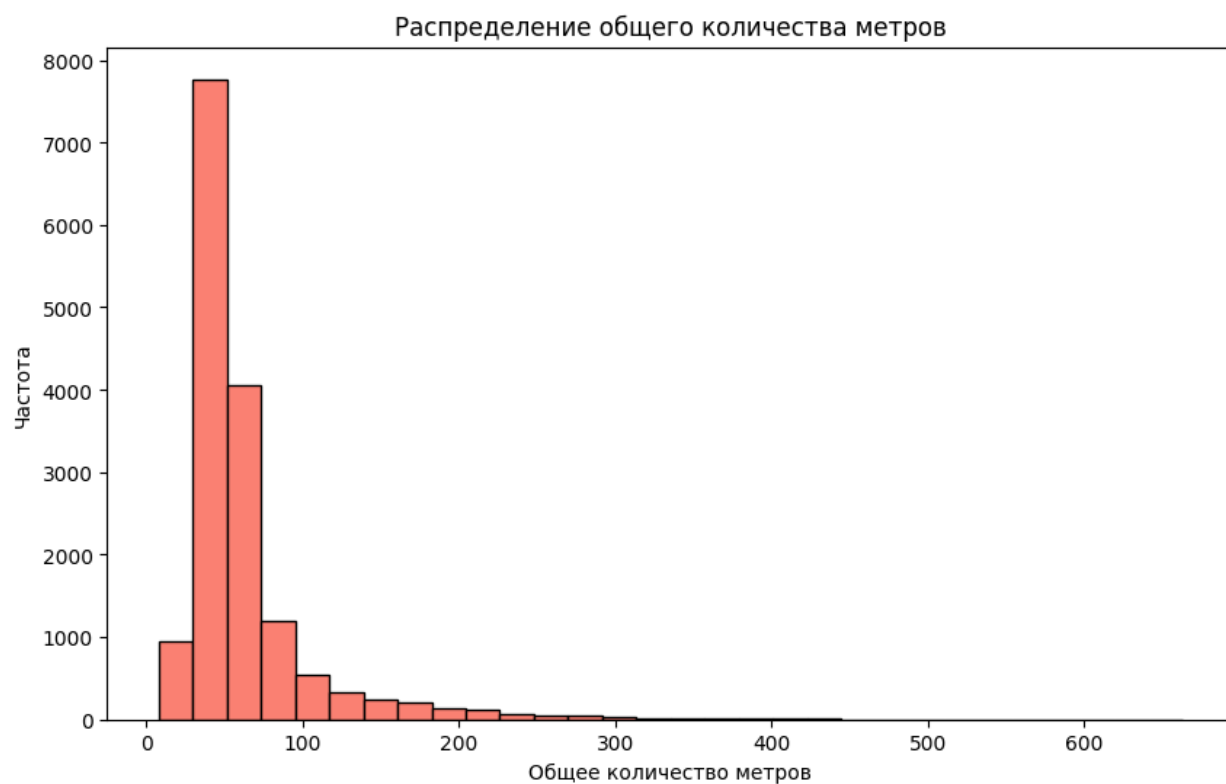
Количество комнат/этажей и цену перевел в числовой тип  
Теперь через вывод строк с максимальным значением по каждому из этих колонок можно искать аномалии. Из было много в колонке с количеством комнат. Поэтому я удалил все строки где комнат больше 10.

И добавил нашу целевую переменную разделив цену на площадь, и получил цену за квадратный метр.

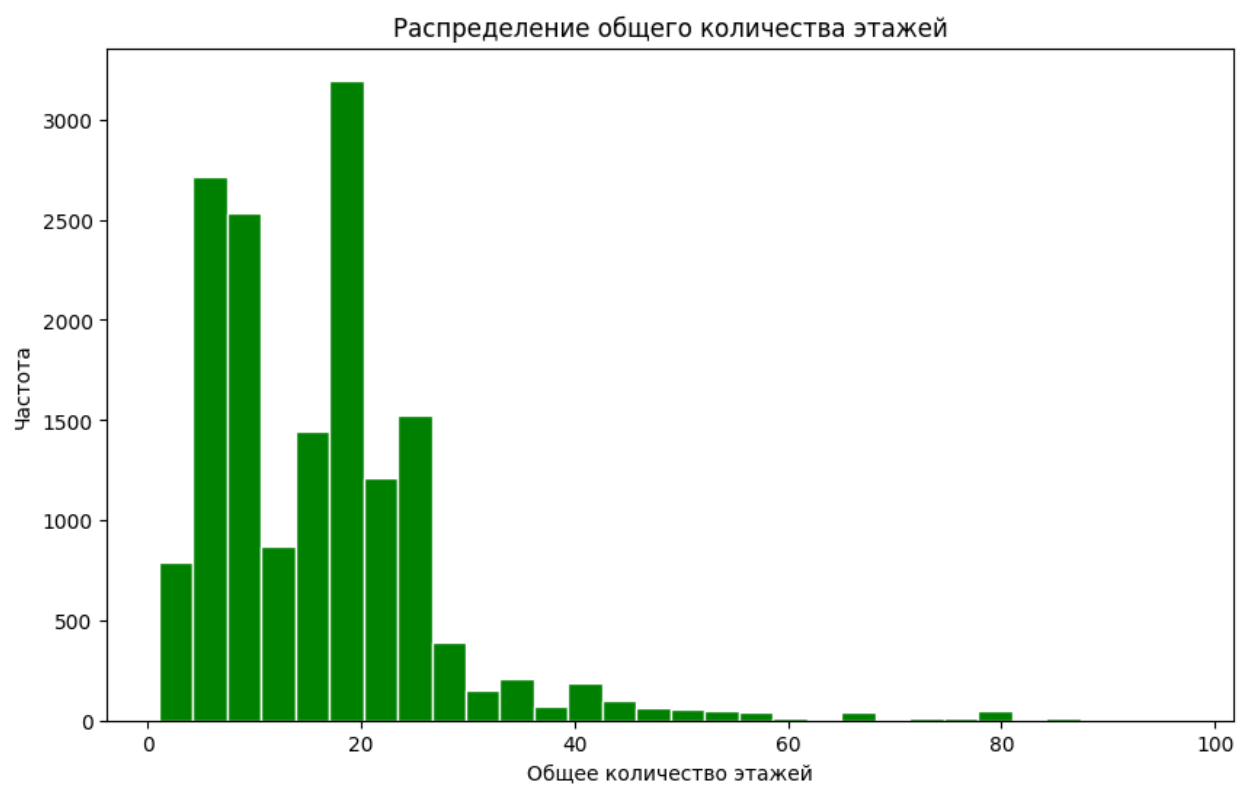
На основании числовых значений уже можно было строить некоторые графики для анализа:



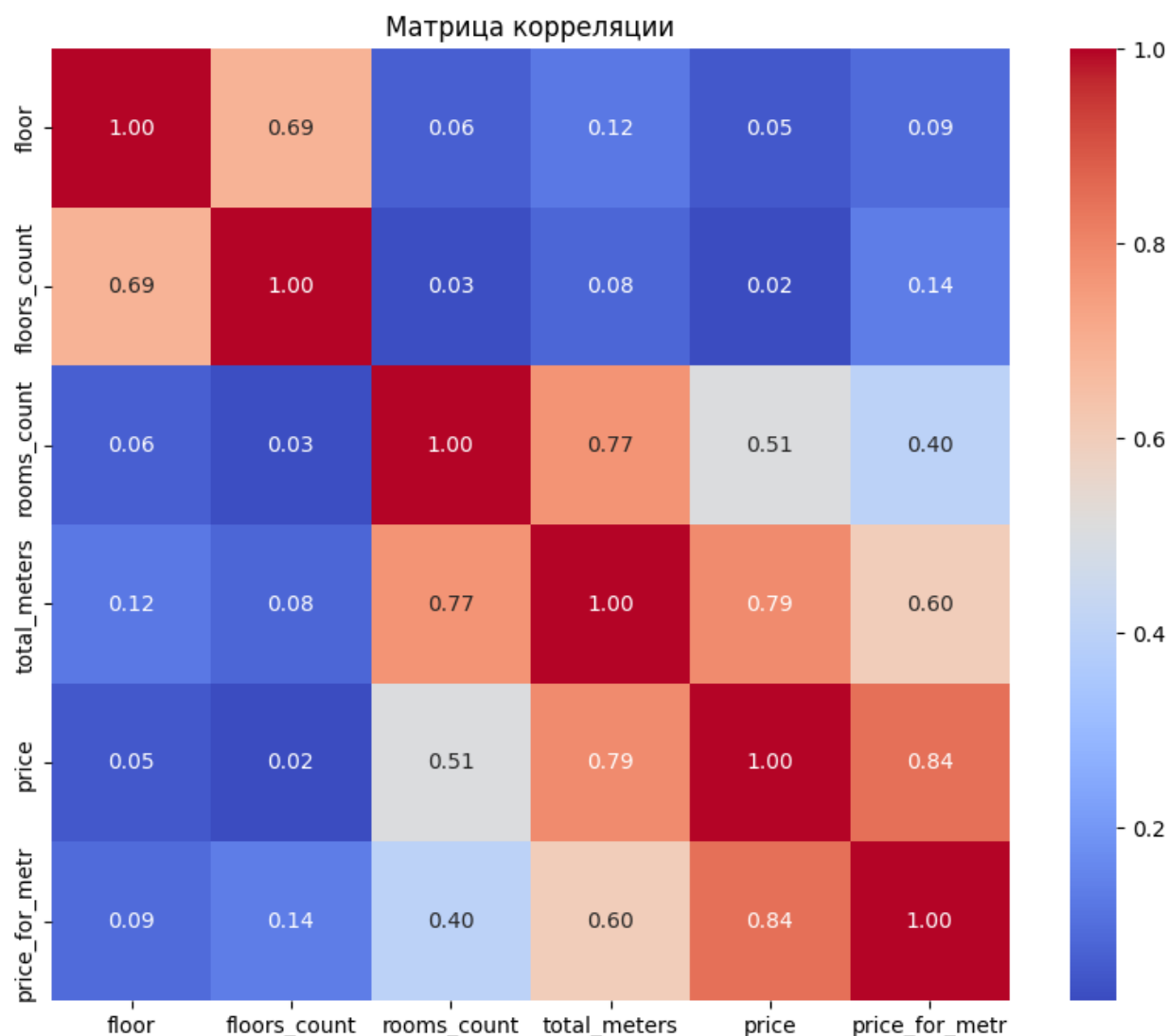
Самая частая цена за М\*2 около 200-300 тыс.



Чаще всего встречаются квартиры площадью 50-70 м\*2.



Чаще всего встречаются 15-20 этажные здания.



Построив такую матрицу корреляции можем понять:

на целевую переменную сильно влияют цена и площадь квартиры, чуть меньше влияет количество комнат. На цену влияют площадь и колво комнат. этаж коррелируют количество этажей. Площадь коррелирует с колвом комнат и ценой

Удалил колонки 'url' и 'deal\_type' за ненадобностью. Приступил к преобразованию категориальных типов с помощью библиотеки sklearn

```

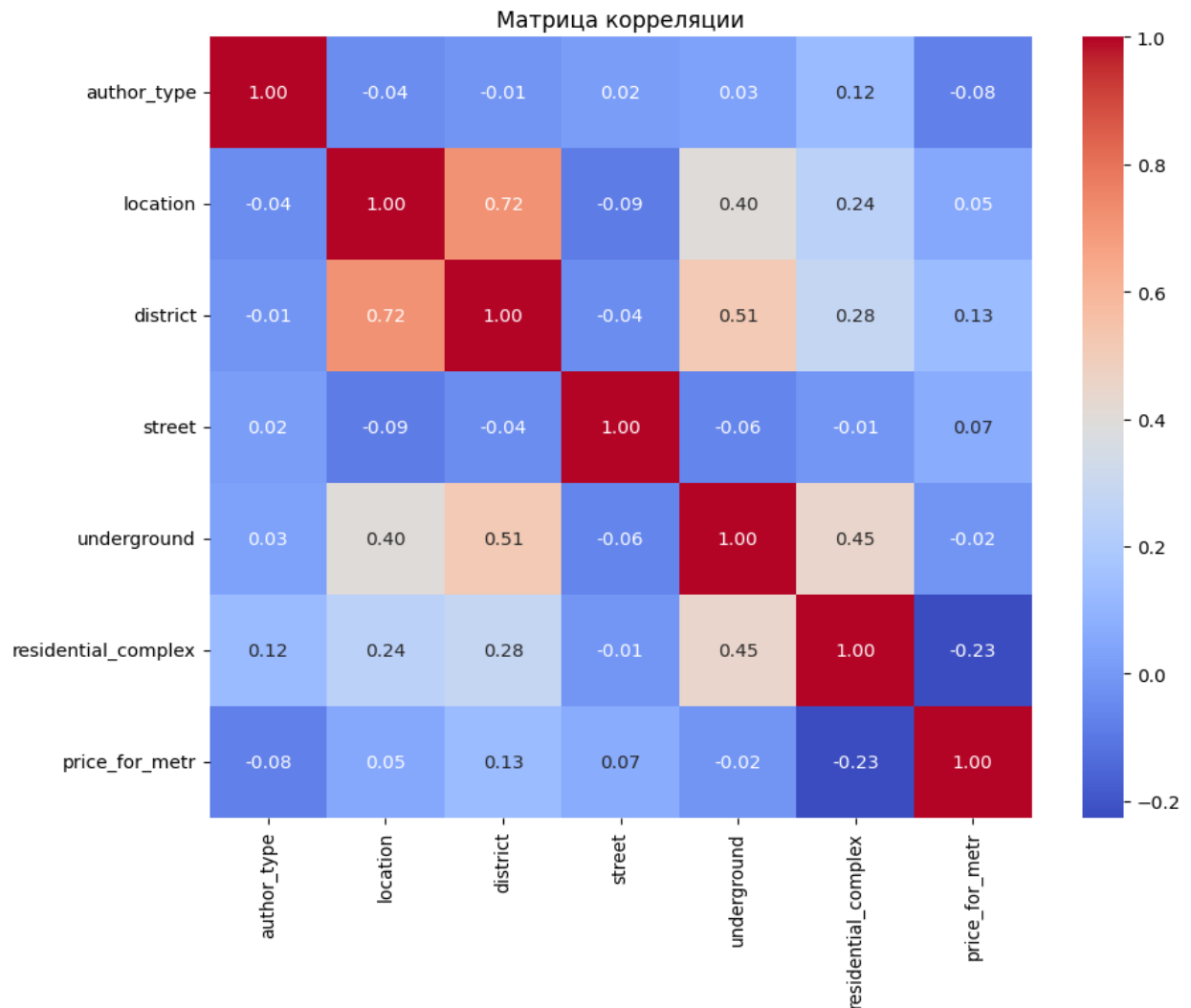
from sklearn.preprocessing import OrdinalEncoder

categorical_columns = df.select_dtypes(include=['object']).columns.tolist()

ordinal_encoder = OrdinalEncoder()
encoded_data = ordinal_encoder.fit_transform(df[categorical_columns])
df[categorical_columns] = encoded_data.astype(int)

```

Преобразовав построил еще одну матрицу корреляции. Она дала понять что расположение в конкретном дата-сете не сильно влияет на цену, что не отражает действительности.



## Дополнительное задание:

С помощью такого кода:

```

1 import cianparser
2
3 moscow_parser = cianparser.CianParser(location="Москва")
4 data = moscow_parser.get_flats(deal_type="rent_long", rooms=(1), with_saving_csv=True, additional_settings={"start_page":1, "end_page":1000,"sort_by":"price_from_min_to_max"})
5
6 print(data[0])

```

Собрал 100 самых дешевых однокомнатных квартир в Москве



Самым дешевым оказался такой вариант:

Только на Цане

**Сдается 1-комн. квартира, 38 м<sup>2</sup>**

Москва, ЗелАО, р-н Крюково, Зеленоград, мкр. 16-й, к1602 [На карте](#)

Зеленоград — Крюково 22 мин.

Фирсановская 14 мин.

Сходня 20 мин.

Пятницкое шоссе 25 км от МКАД

Ленинградское шоссе 26 км от МКАД

🗨

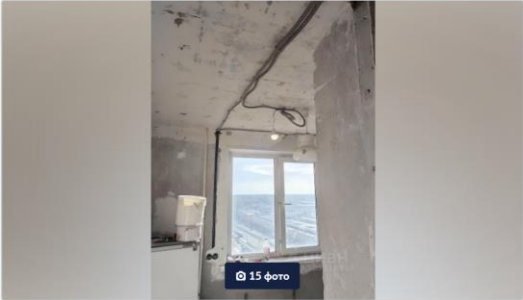
✎

📷
















🖨

🔗

🚩 Пожаловаться



15 фото



Общая площадь

38 м<sup>2</sup>

Жилая площадь

16 м<sup>2</sup>

Площадь кухни

10 м<sup>2</sup>

Этаж

12 из 14

Год постройки

1993

Отзыв о сайте

20 000 Р/мес.

Следить за изменением цены

Предложите свою цену

Оплата ЖКХ \_\_\_\_\_ включена (без счётчиков)

Залог \_\_\_\_\_ 20 000 Р

Комиссия \_\_\_\_\_ 100%

Предоплата \_\_\_\_\_ 1 мес/ц

Срок аренды \_\_\_\_\_ от года

Показать телефон

Написать

Быстро отвечает на сообщения

АГЕНТСТВО НЕДВИЖИМОСТИ  
**ОХОТНЫЙ РЯД** офис Тверская  
Документы проверены

РИЕЛТОР  
**Инга Агент**  
★ 4.4 - 55

**Заключение:** что то сделал, данные подготовлены, анализ провел.  
Данные взяты неполные, что сильно уменьшило точность расчетов.