Государственное бюджетное профессиональное образовательное учреждение Московской области

«Физико-технический колледж»

**МДК.05.01 ИСП**

**Кейс компании «Самолет» (big data)**

**«Модель оценки цены квартиры на вторичном рынке по Московскому региону: Москва, Новая Москва, Московская область»**

**1 интенсив**

Выполнил: студент группы ИСП23В Фёдоров Андрей

2024

**Содержание**

1. Цель
2. Cписок параметров, значительно влияющих на цену квадратного метра жилой площади
3. Парсинг данных
4. Подготовка данных для анализа
5. Исследовательский Анализ Данных (EDA)
6. Рефлексия

**Цель**

Собрать данные и провести разведочный исследовательский анализ данных (EDA) для построения модели, которая будет оценивать цену квадратного метра недвижимости в Московском регионе (Москва, Новая Москва, Московская область).

**Cписок параметров, значительно влияющих на цену квадратного метра жилой площади**

При анализе рынка недвижимости, можно выделить несколько ключевых параметров. Эти параметры могут варьироваться в зависимости от конкретного региона. Вот список основных факторов:  
  
1. ***Местоположение***  
- Географическая расположенность: Цены на недвижимость сильно зависят от расположения (центр города, спальные районы, пригороды).  
- Инфраструктура: Наличие рядом школ, больниц, магазинов, общественного транспорта и развлекательных объектов.  
- Экологическая обстановка: Наличие парков, зеленых зон и чистота окружающей среды.  
  
2. ***Характеристики объекта***  
- Площадь и планировка: Общая площадь квартиры, количество комнат и удобство планировки.  
- Состояние и возраст здания: Новые дома обычно стоят дороже, чем старые, требующие ремонта.  
  
3. ***Рыночные факторы***  
- Спрос и предложение: Уровень спроса на жилье в конкретном районе и количество доступных объектов.  
- Конкуренция: Наличие аналогичных объектов в районе, которые могут влиять на цены.  
  
4. ***Экономические факторы***  
- Экономическая ситуация: Общий уровень экономического роста, уровень инфляции, безработицы и т.д.  
- Процентные ставки по ипотечным кредитам.  
  
 5. ***Социальные факторы***- Качество жизни: Уровень преступности, безопасность, доступность образовательных и медицинских учреждений.  
  
Учитывая все перечисленные факторы, можно получить более полное представление о том, что влияет на цену квадратного метра жилой площади.

**Парсинг данных**

Для выполнения задачи использовался фреймворк cianparser, который позволяет извлекать информацию о квартирах, доступных на продажу на площадке Cian.  
  
***Цели парсинга***  
Основной целью парсинга было получение информации о квартирах с указанными параметрами, такими как:  
- Местоположение (Москва и близлежащие города)  
- Количество комнат (от 1 до 5)  
- Количество страниц для извлечения данных, зависящее от местоположения и количества комнат.

***Код***

  
***Пояснение работы кода***  
  
1. Импорт необходимых библиотек:



В начале кода импортируются библиотеки cianparser, pandas, а также модули для работы со временем и генерации случайных чисел.  
  
2. Определение списка локаций:



В переменной locations задаются населенные пункты, для которых будет выполняться парсинг. В данном случае это Москва и несколько прилегающих городов.

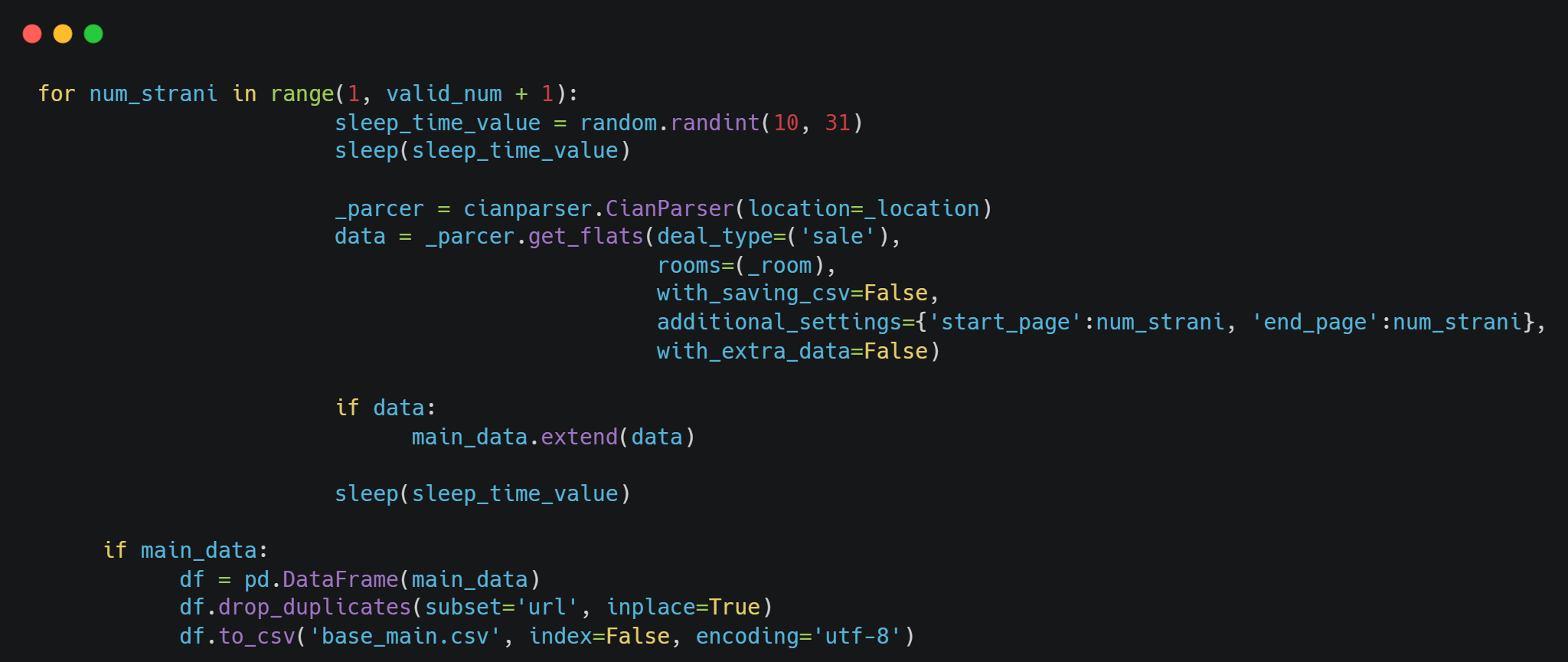
3. Сбор данных:



Сначала инициализируется пустой список main\_data, который будет хранить все собранные данные.

Далее, с помощью вложенных циклов, происходит перебор местоположений, количества комнат и страниц для парсинга. Для каждого местоположения устанавливается максимальное количество страниц (valid\_num), доступных для парсинга, в зависимости от города и количества комнат.

4. Запрос данных и Обработка полученных данных:

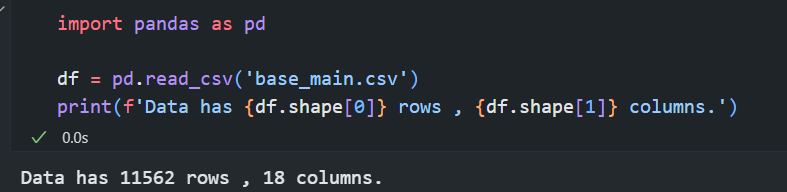


Для каждого запроса создается экземпляр CianParser для заданного местоположения и выполняется запрос на получение данных о квартирах с заданными параметрами.

Запросы выполняются с задержкой от 10 до 31 секунды, чтобы минимизировать риск блокировки со стороны сервера.

Если данные были успешно получены, они добавляются в список main\_data.

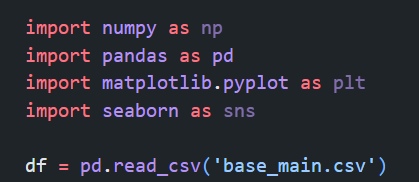
После завершения всех запросов создается DataFrame из собранных данных, дубликаты удаляются с помощью метода drop\_duplicates(), и окончательные данные сохраняются в файл base\_main.csv.  
  
В результате выполнения данного кода была собрана информация о квартирах, доступных на продажу в Москве и прилегающих районах.



**Подготовка данных для анализа**

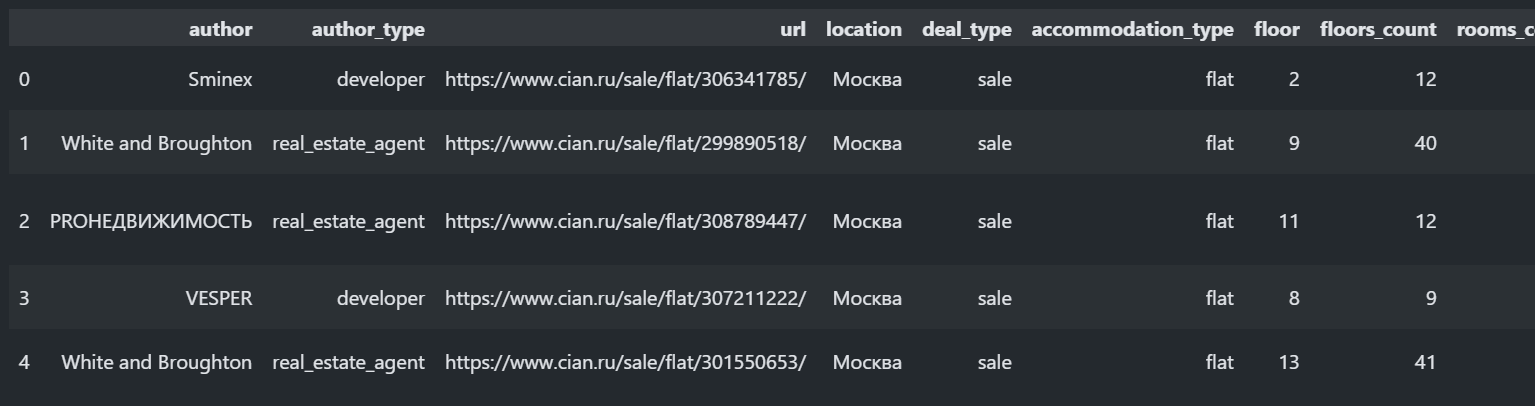
Подготовка данных является ключевым этапом в любом проекте анализа данных, так как качество данных напрямую влияет на результаты анализа.

1. Импорт данных

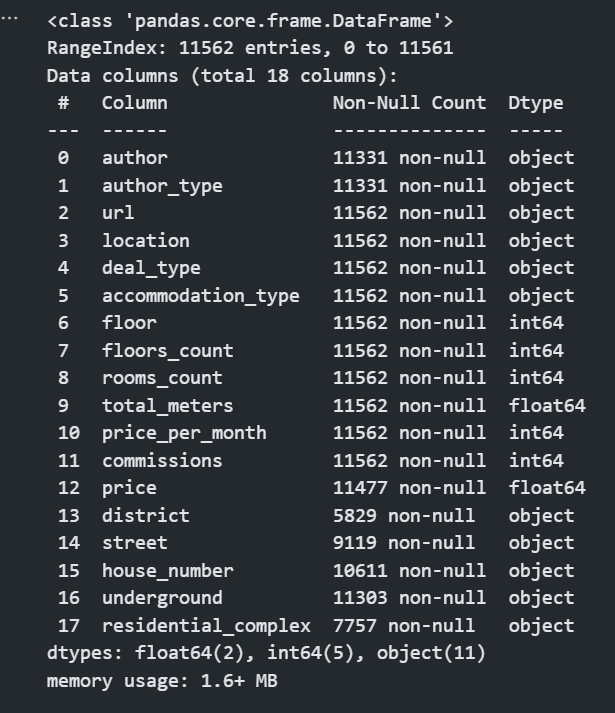


  В первую очередь, данные были импортированы в среду анализа с использованием библиотеки Pandas. Данные представляют собой таблицу, содержащую информацию о различных характеристиках объектов недвижимости.

2. Обзор данных



  После импорта данных был выполнен первичный обзор, включающий:  
   - Проверку структуры DataFrame.  
   - Определение типов данных для каждого столбца.  
   - Подсчет ненулевых значений для выявления отсутствующих данных.



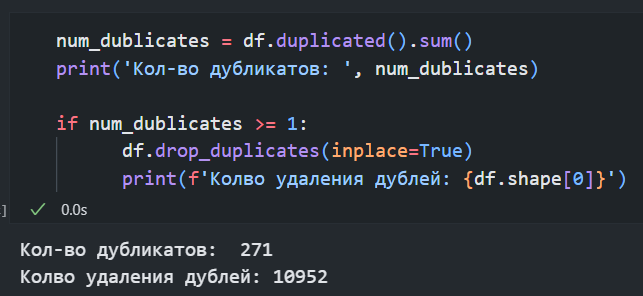
3. Обработка отсутствующих значений, Преобразование типов данных,

Добавление нового столбца, который расчитывает цену за м2

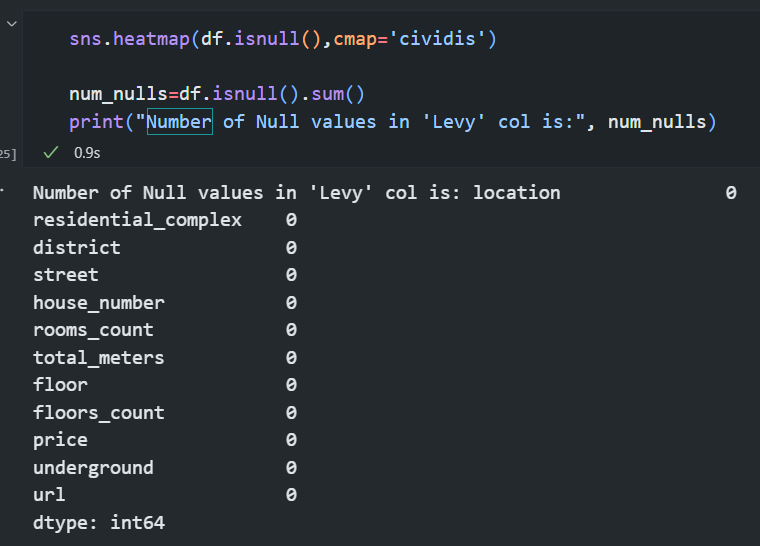
В процессе анализа были выявлены столбцы, содержащие NaN (отсутствующие) значения:

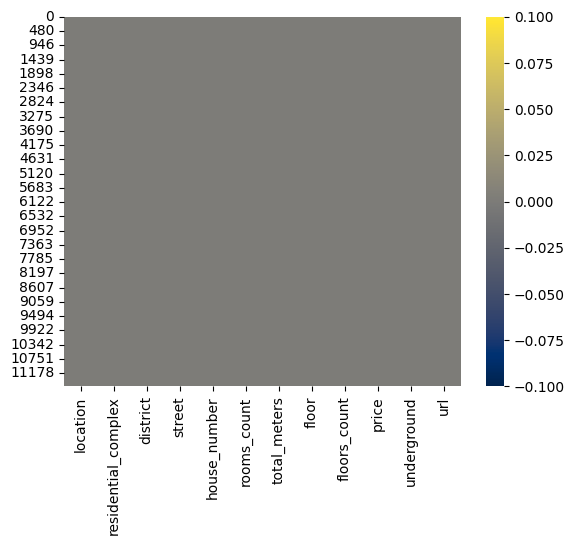


5. Проверка на наличие дубликатов

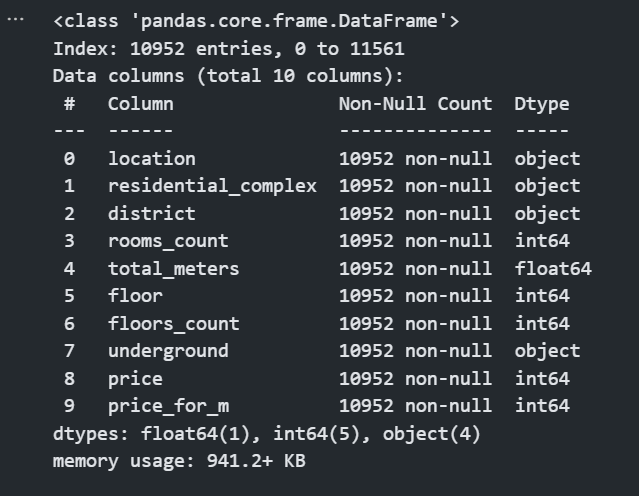


6. Проверка на наличии Null





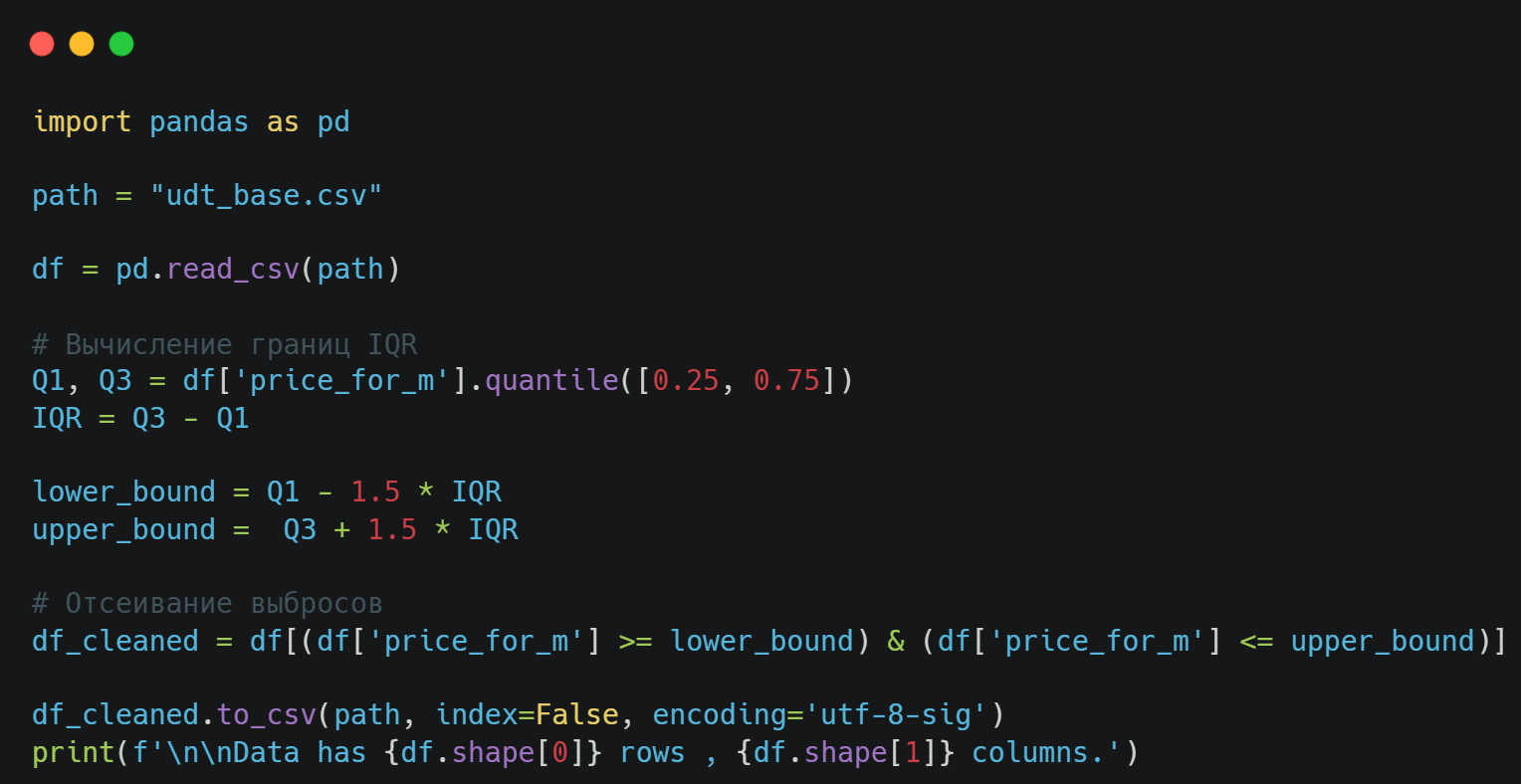
По завершении этапов подготовки данных был получен очищенный и структурированный DataFrame, готовый для дальнейшего анализа. Все столбцы были проверены на наличие ненулевых значений, а типы данных были скорректированы для обеспечения корректности при анализе.





**Исследовательский Анализ Данных (EDA)**

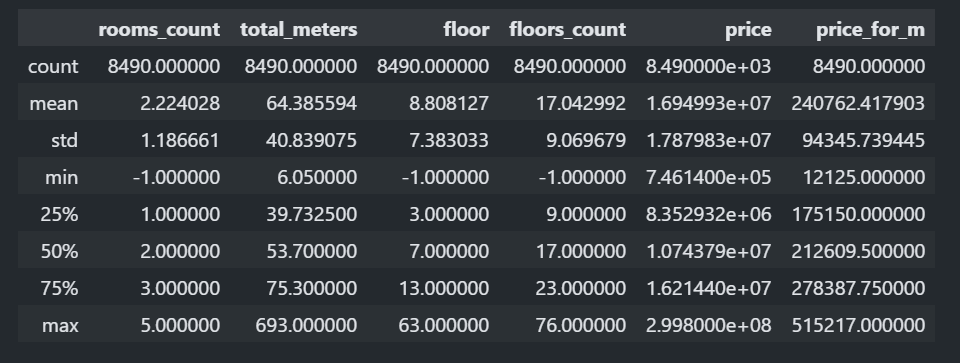
***Поиск выбросов***



До поиска строк было 10к +, после 9242

***Информация о базе***





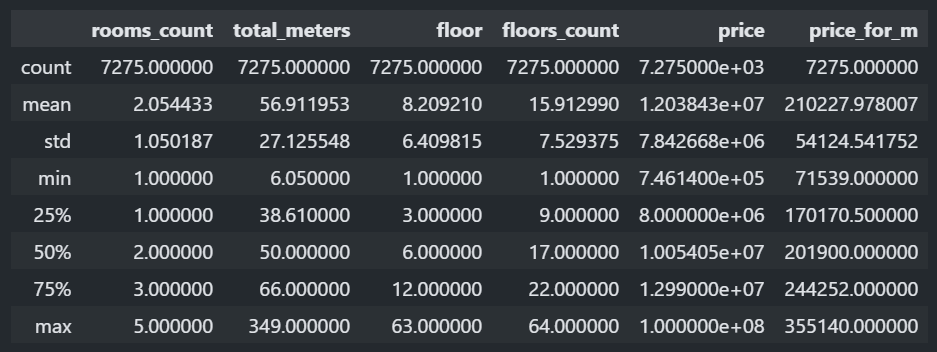
Во время первого прогона, получил аномалии в поле min, в столбцах **rooms\_count, floor, floors\_count, price**.

Удалял руками через поиск в данных

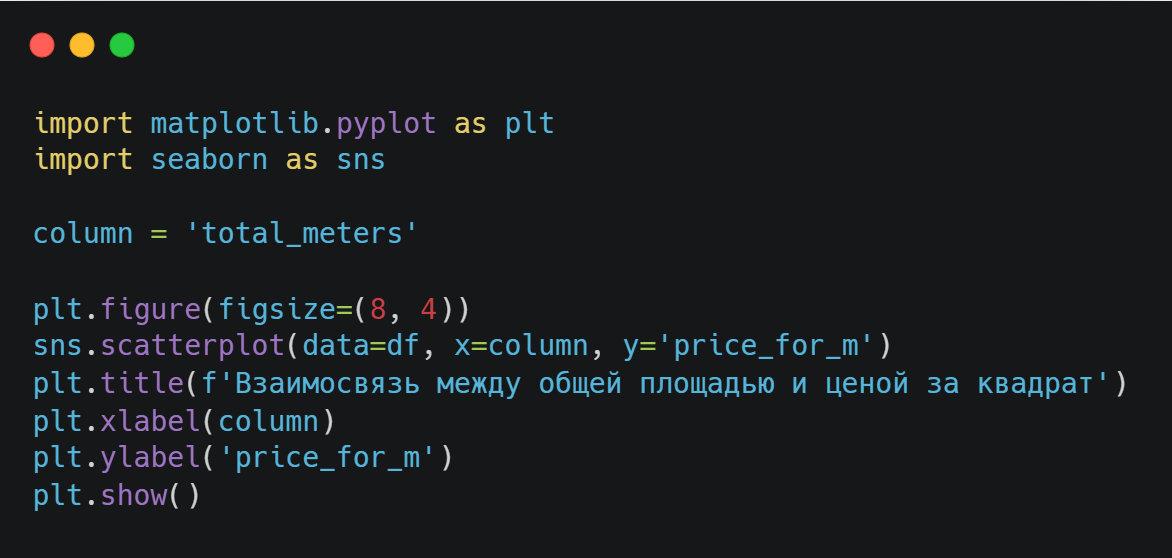
Также в столбце total\_meters в поле max слишком большое знач для квар.

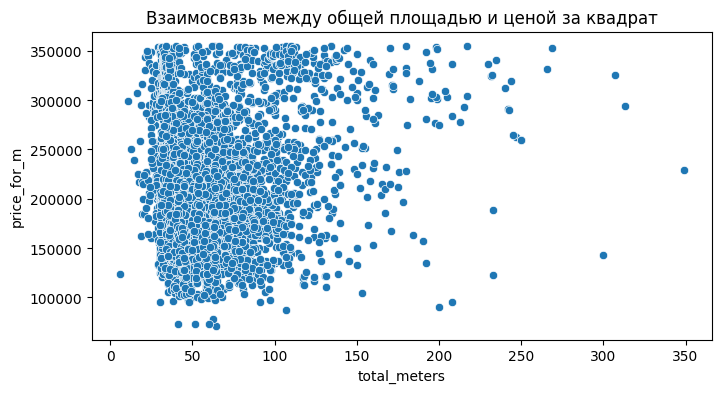
Удалял руками через поиск в данных

После чистки, получаю такие данные

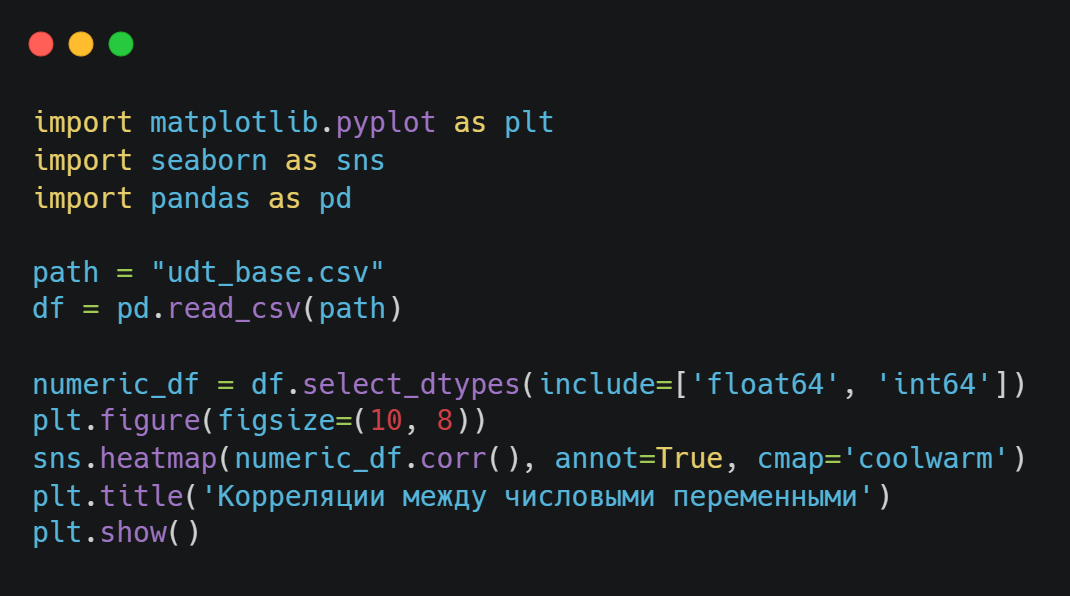


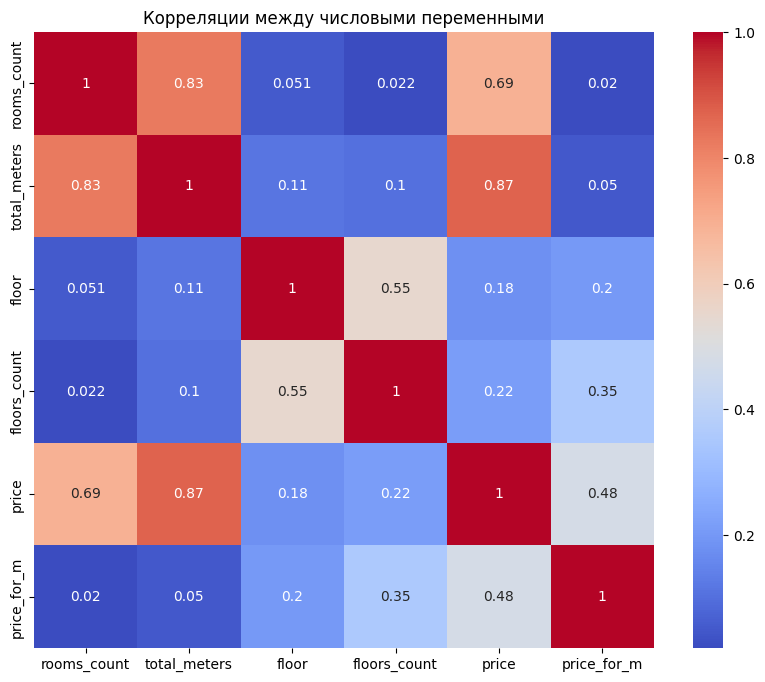
Взаимосвязи



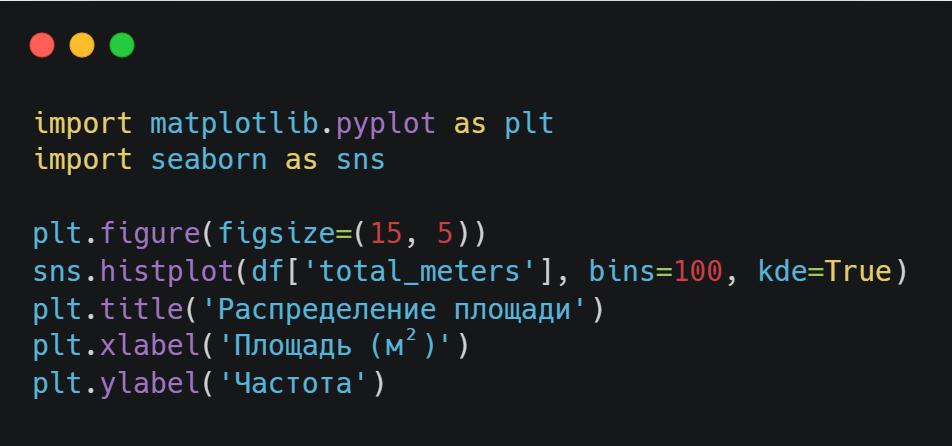


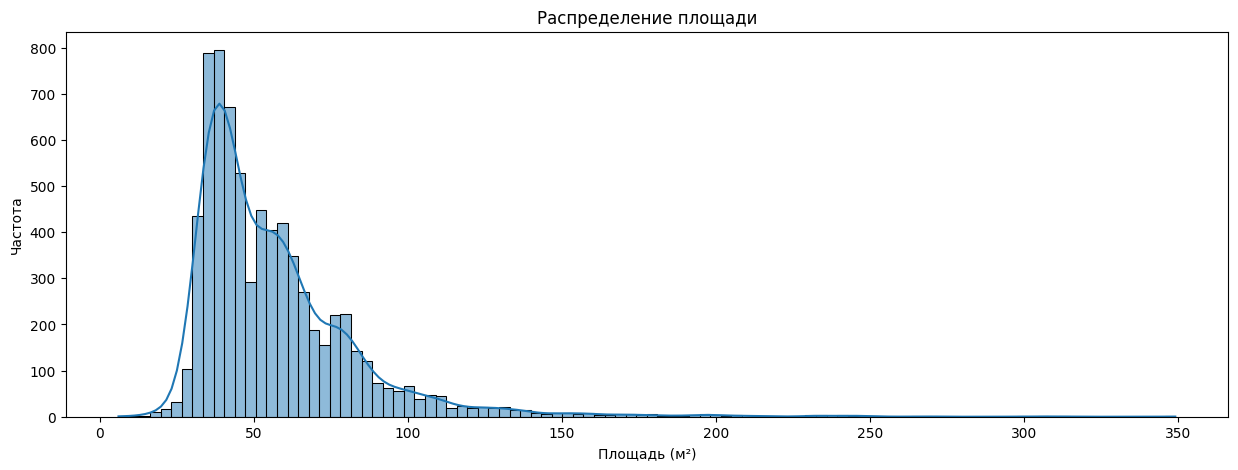
Корреляция





Графики





**Рефлексия**