Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Аналитический отчет**

**Модель оценки квартиры на рынке продажи**

**в Москве, Новой Москве и Московской области**

Работу выполнил:

Студент группы ИСП-21

Мисюк Вячеслав

Долгопрудный, 2024

**Введение: д**анный аналитический отчет посвящен исследованию факторов, влияющих на стоимость квартир. Для проведения анализа была разработана модель на языке программирования Python с использованием библиотек Pandas и Matplotlib. Модель позволила проанализировать обширный набор данных о квартирах, выявив ключевые факторы, определяющие их стоимость.

**Цель:** Целью данного аналитического отчета является изучение и выявление ключевых факторов, влияющих на стоимость квартир, с использованием модели, разработанной на языке программирования Python с библиотеками Pandas и Matplotlib.

**Задачи:**

**1)** Используя открытые источники, такие как Циан, проанализировать и создать базу данных с характеристиками квартир

**2)** Очистить лишние и «грязные» данные, количество которых недостаточно для анализа, либо они являются не нужными

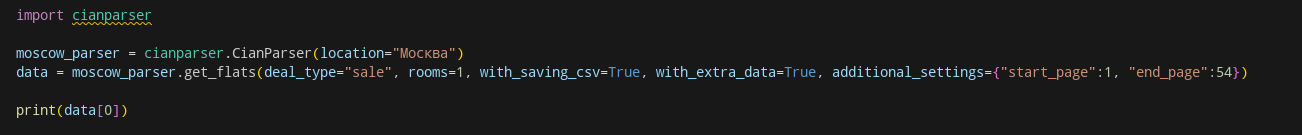
**3)** Визуализировать готовые данные в виде гистограмм и графиков

**Основная часть**

В качестве отправной точки для нашего анализа была выбрана платформа Циан - один из крупнейших онлайн-сервисов по поиску недвижимости в России. Изначально, для сбора данных о квартирах, необходимых для анализа, было решено применить метод парсинга. Парсинг - это процесс автоматического извлечения данных с веб-сайтов. Для этого был разработан специальный скрипт на языке Python, который загружал информацию о квартирах, доступную на Циан.

Для парсинга Циана была использована библиотека CianParser

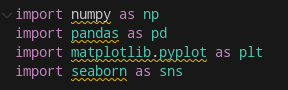
Код, который был использован для парсинга данных:



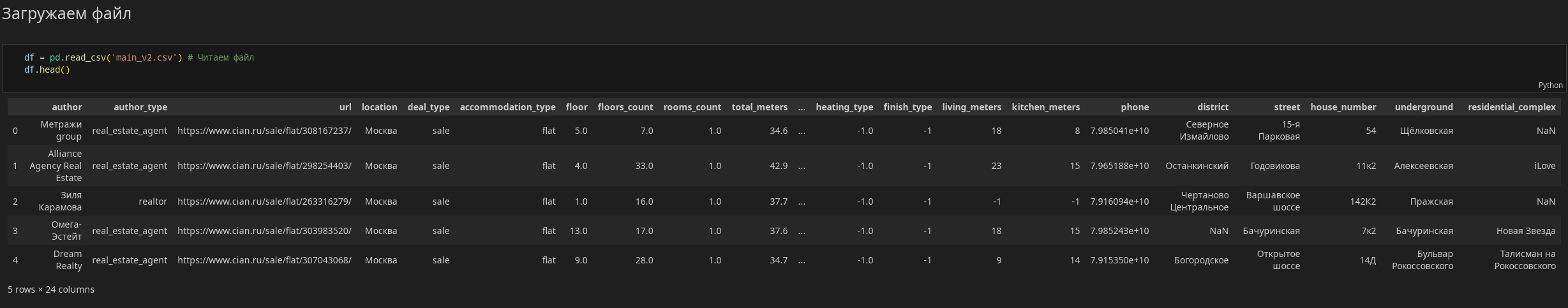
В результате парсинга был готов файл размером в ~9500 строк, где содержались такие значения, как: общая площадь, цена, этаж, район, метро и другие.

**Теперь перейдем к python:**

Для начала работы нам необходимо импортировать нужные библиотеки; импортируем numpy, pandas, matplotlib, seaborn:

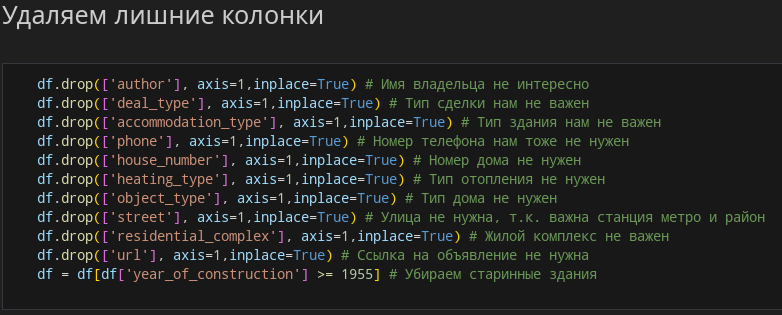


Загружаем файл в программу и выводим первые пять строк, чтобы убедиться, что всё загрузилось и работает верно:

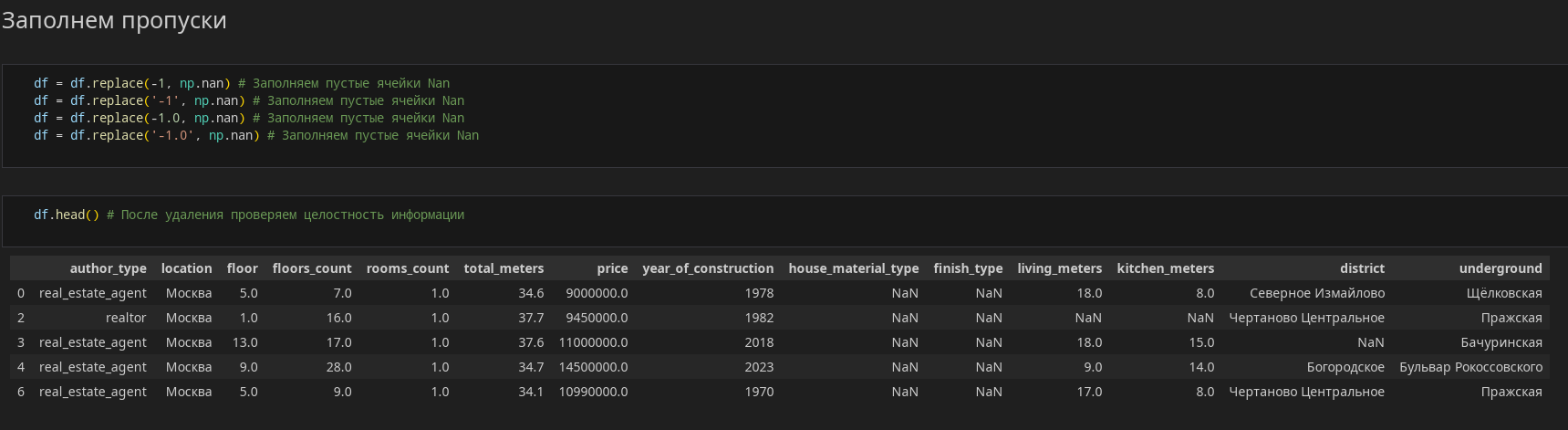


В результате парсинга было добыто много лишней информации, такие как номер дома, имя автора, ссылка на объявление, номер телефона…

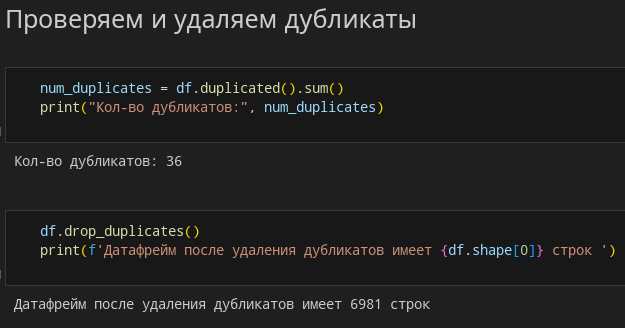
Такие данные нужно удалить и удаляем колонки с этими лишними данными, а также сразу отсортируем наш файл, чтобы в нем остались здания только 1950 года постройки и младше



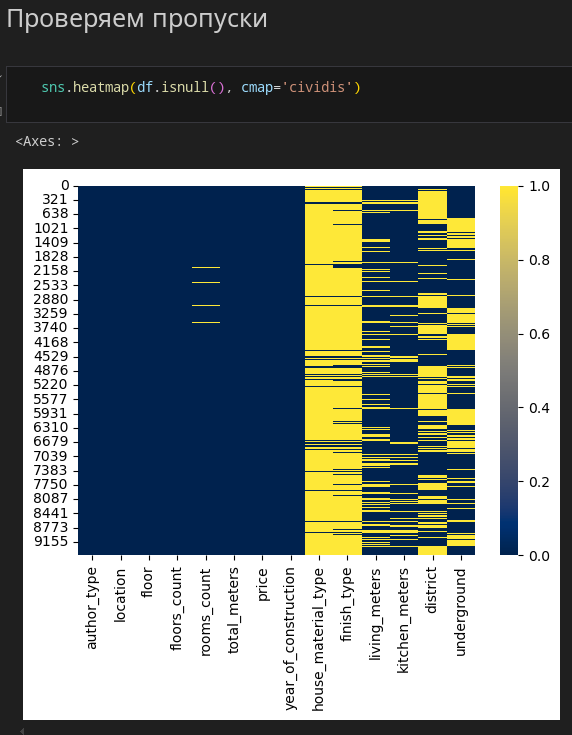
В нашем файле много неизвестных данных, которые не удалось запарсить и они помечены как «-1» в ячейках, они нам будут мешать, поэтому заменяем все возможные «-1» ячейки на Nan-значения и проверяем, всё ли правильно:



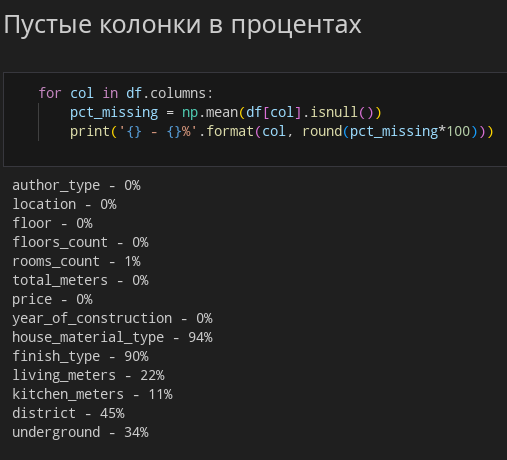
В файле есть возможность образования дубликатов, поэтому по хорошему их надо удалить, что мы и делаем:



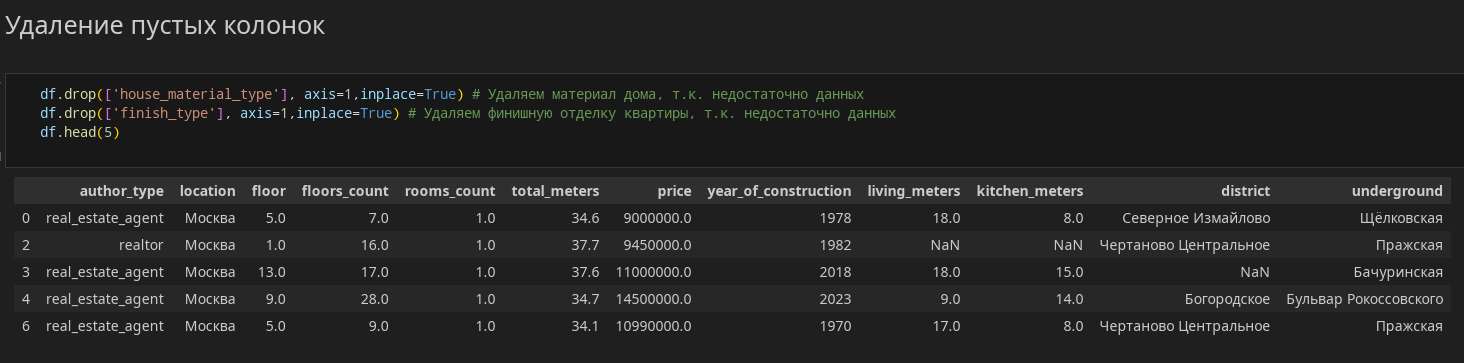
Далее выводим тепловую карту с пропусками в файле:

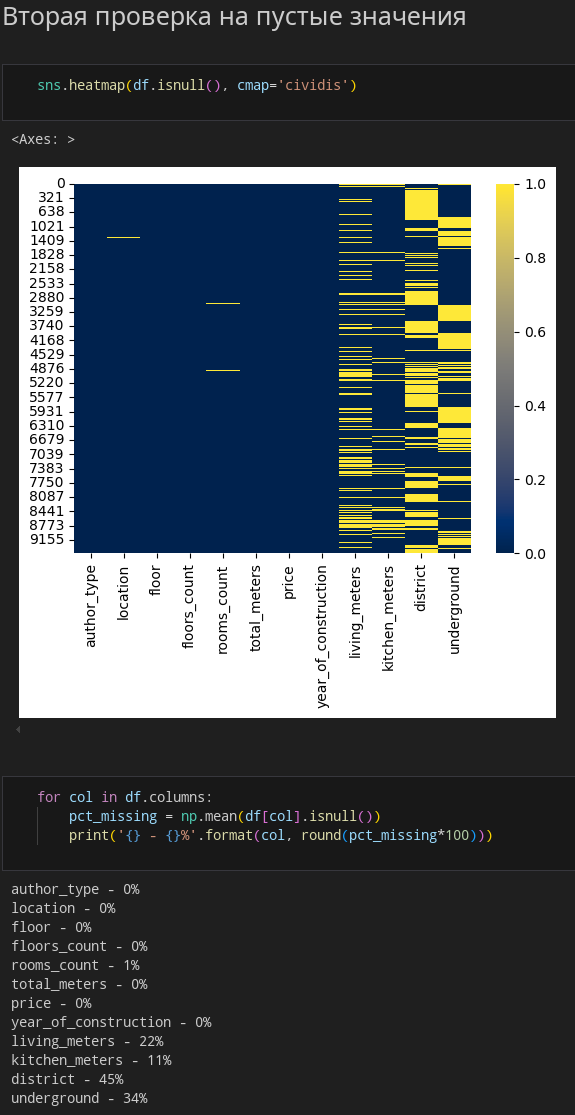


Изобразим эту карту в виде строк, чтобы пустые данные было легче подсчитать:



По результатам этого анализа пустых колонок мы узнали, что колонки с типом материала дома (house\_material\_type) и финишной отделкой дома (finish\_type) являются практически пустыми, т. к. их пустота оценивается в 94% и 90% соответственно. Поэтому их надо удалить:

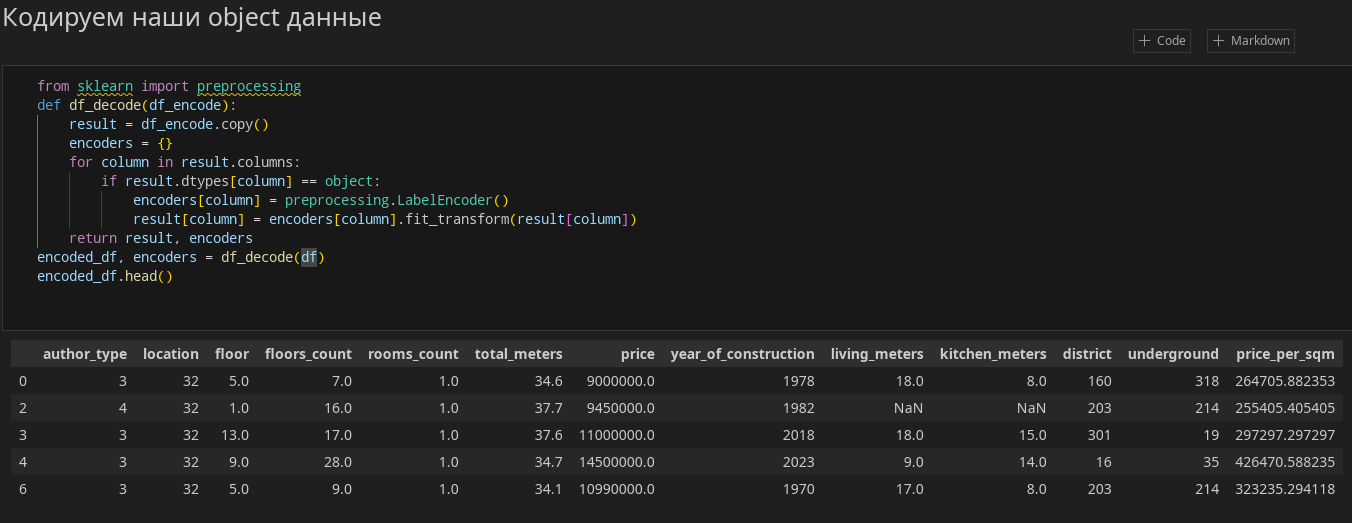


Проверяем тепловую карту пропусков и понимаем, что колонка «district» так же является практически пустой и ее надо удалить, но удалять эту колонку я не решил, т.к. эту колонку я использовал в построении графика

Для дальнейшей работы нам понадобится колонка, которая показывает цену квадратного метра каждой квартиры. Вычисляется она таким способом: цена квартиры/общий метраж квартиры.

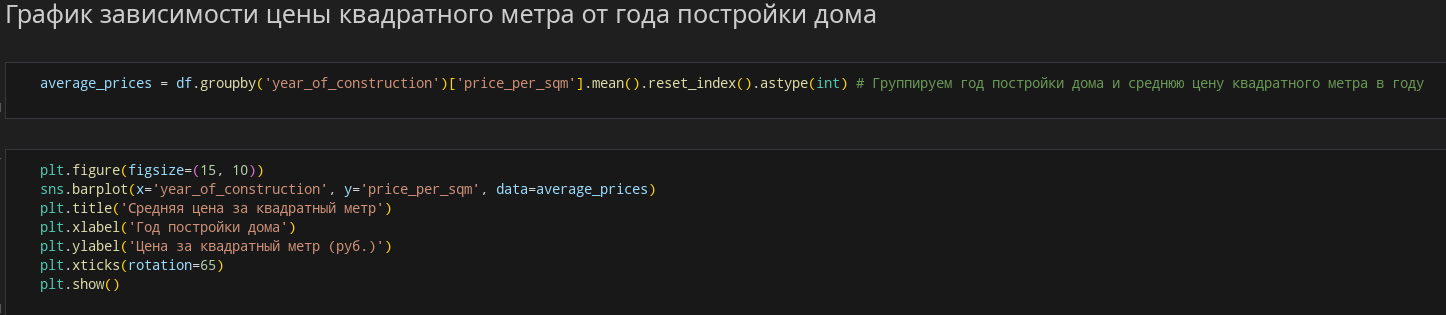


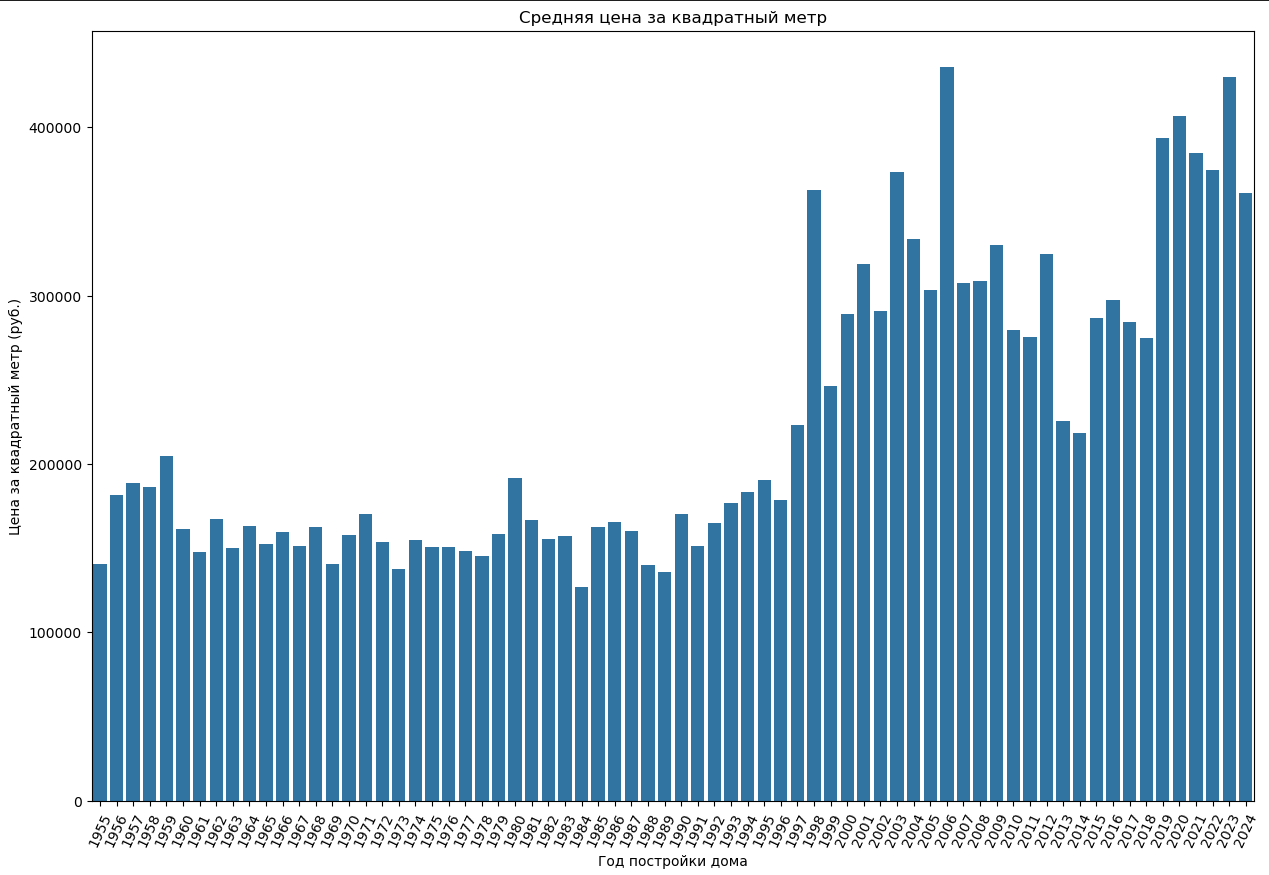
В корреляционной карте понадобятся данные не равные типу «object», поэтому с помощью библиотеки sklearn я закодировал весь файл



**Теперь перейдем к графикам.** Первый график, который я вывел — зависимость средней цены квадратного метра от года постройки дома, где находится эта квартира

Код к этому графику:

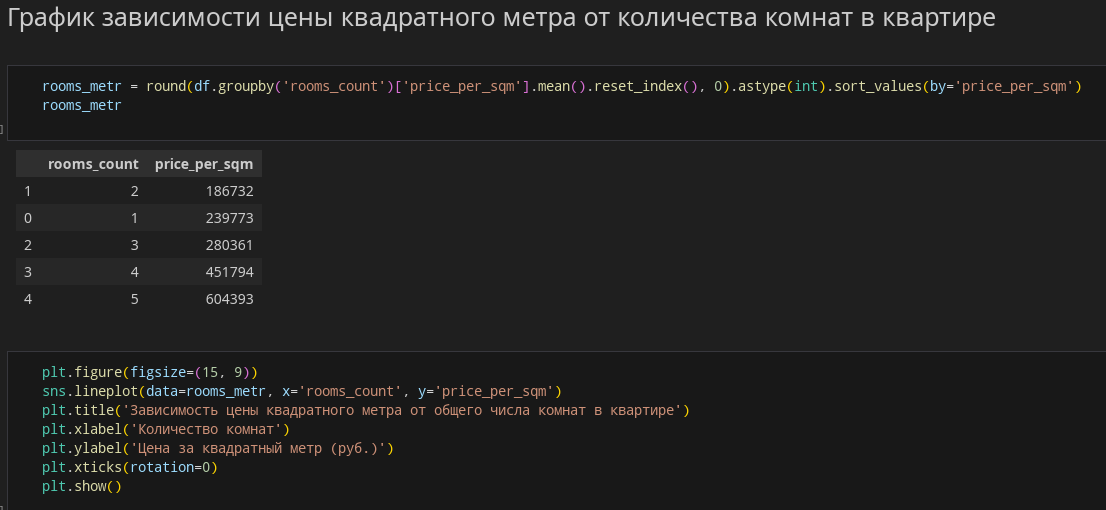
Результат этого кода и сам график:

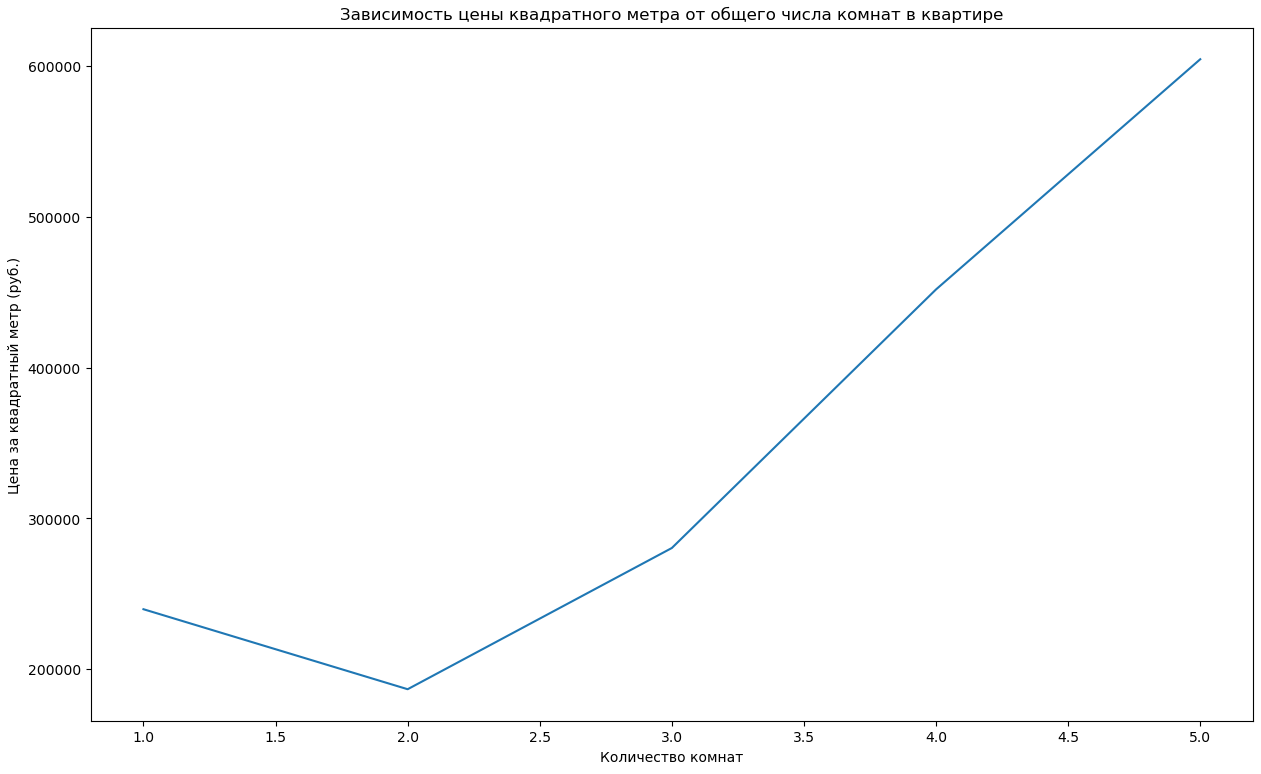


Глядя на график можно понять, что новостройки, построенные за последние 5 лет являются самыми дорогими при счете на квадратных метрах, а самые дешевые это старые здания до ~1995 года.

Далее я построил график зависимости цены квадратного метра от количества комнат в квартире

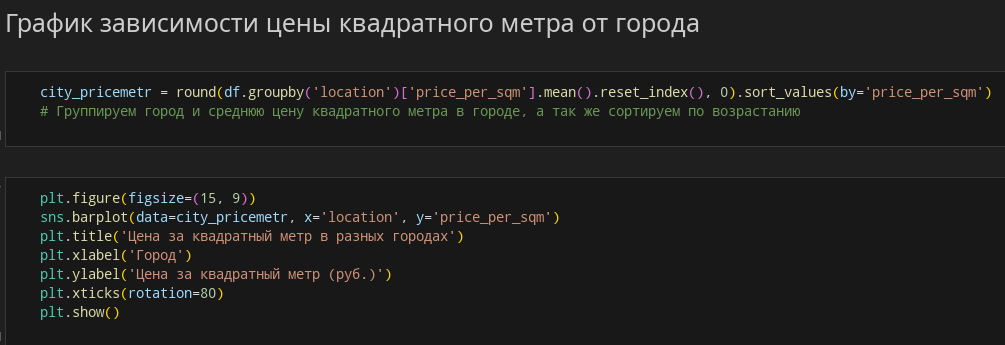
Код к этому графике:

Результат этого кода и сам график:

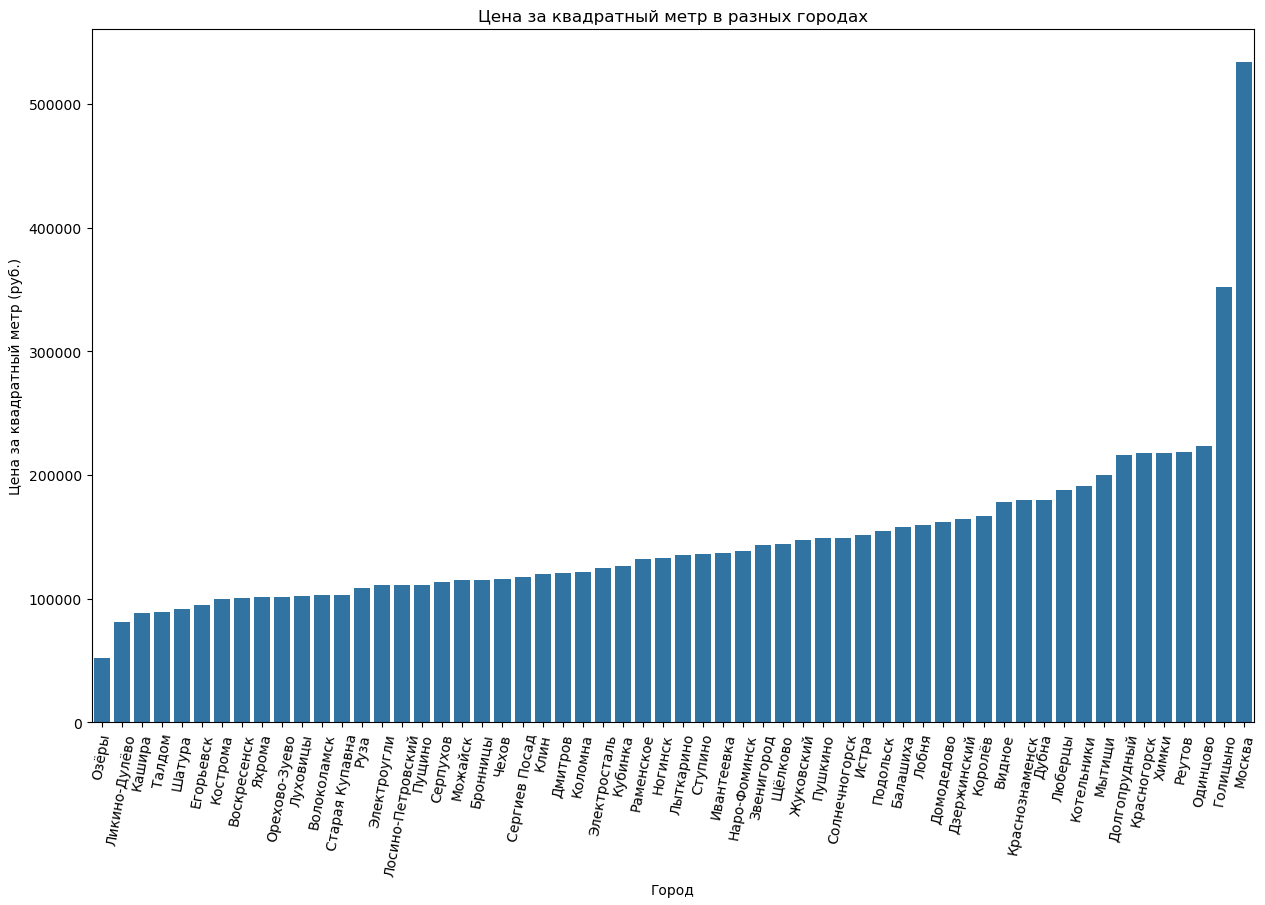
Результатом этого графика является линия, показывающая зависимость средней цены квадратного метра от количества комнат в квартирах. Квартиры с 1 и 2 комнатами являются достаточно дешевыми, а квартиры с 4 и 5 комнатами являются достаточно дорогими в сравнении с 1 и 2 комнатами.

Далее я построил график зависимости средней цены квадратного метра от города, где расположены эти квартиры.

Код к этому графику:

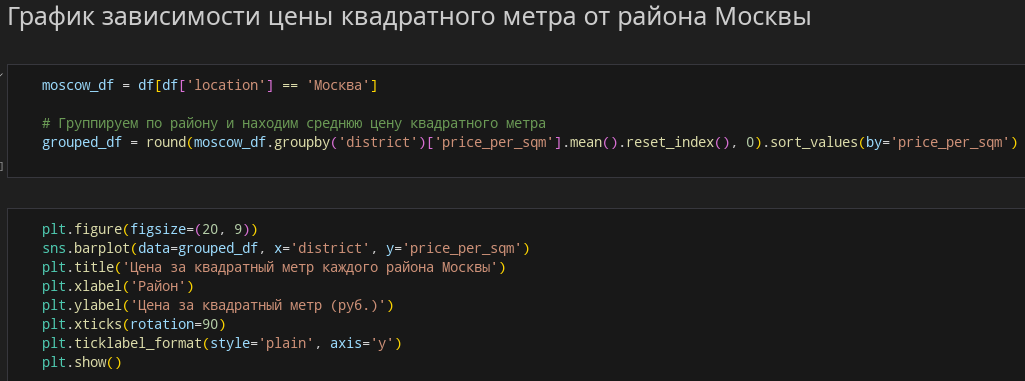


Результат этого кода и сам график:

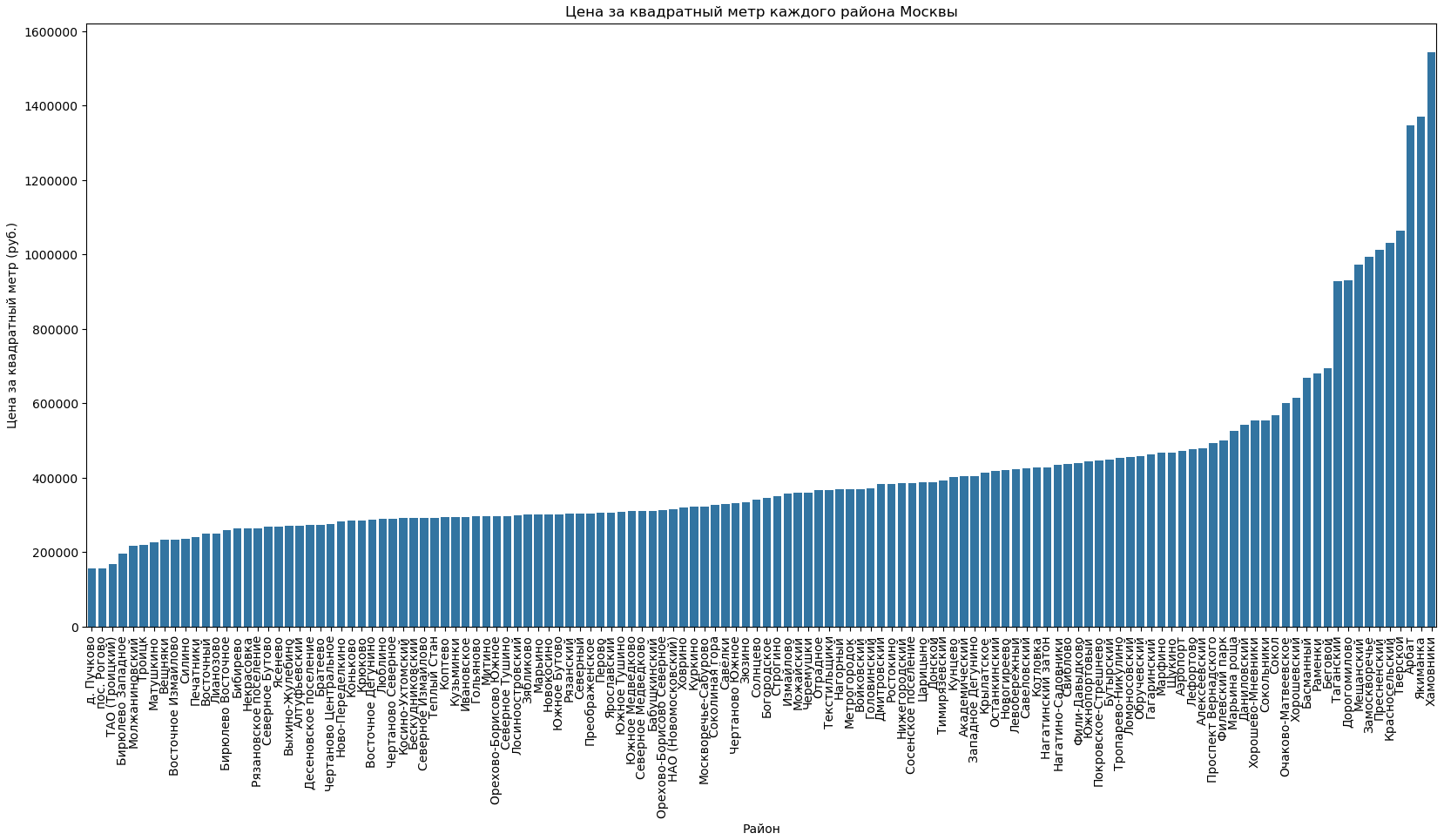
На графике отчетливо видно, что в таких городах как Москва и Голицыно средняя цена за квадратный метр является самой большой, а в городах Озёры, Ликино-Дулёво, Кашира и других цена квадратного метра маленькая и эти города отлично подходят к покупке квартиры, если у покупателя не так много денег.

Далее я построил график зависимости средней цены квадратного метра от района Москвы.

Код к этому графику:

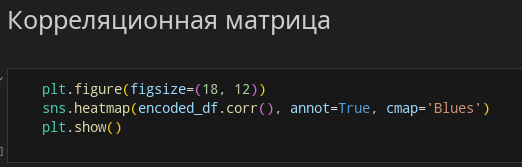


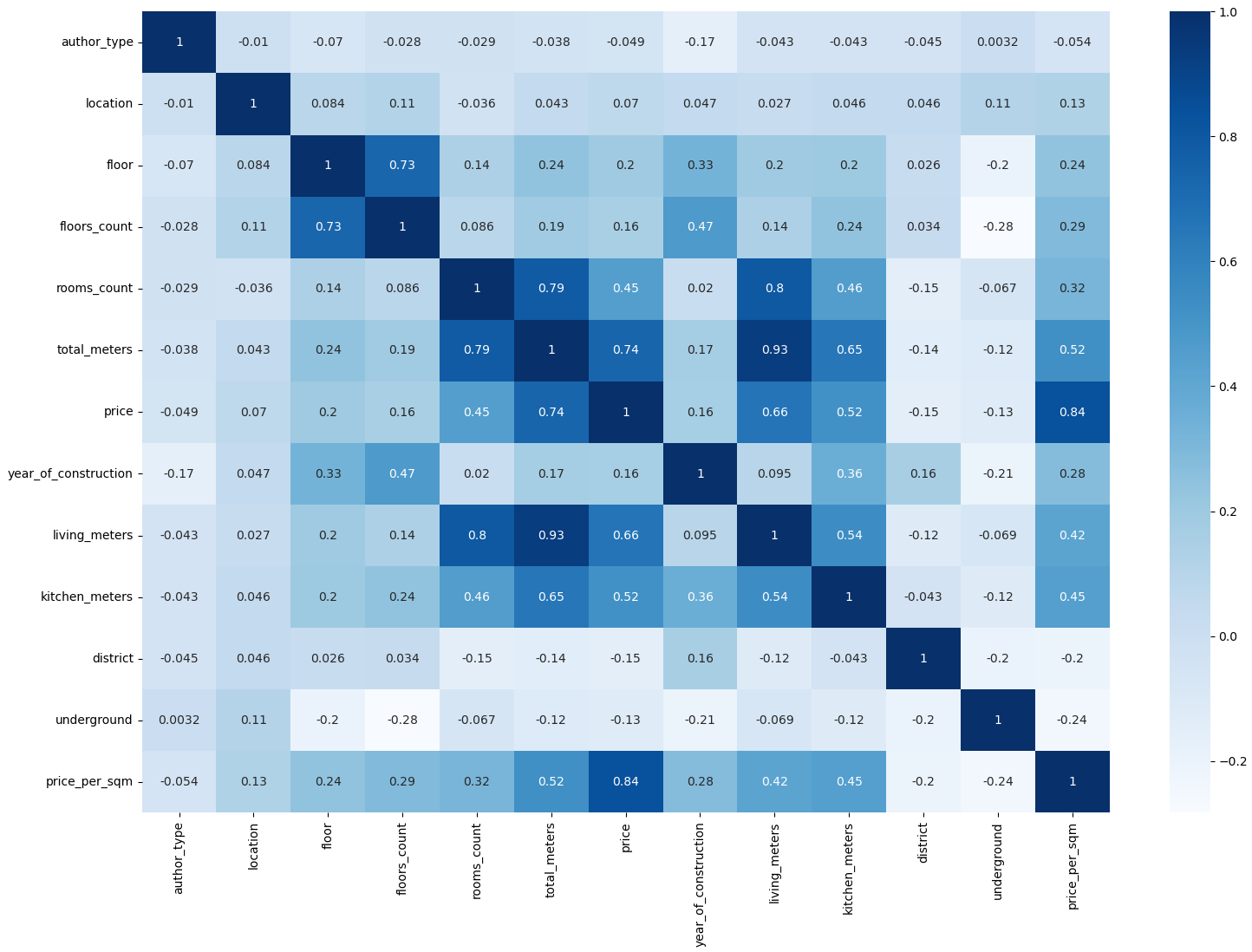
Результат этого кода и сам график:

График показывает в каких районах Москвы самые дорогие цены на квадратный метр: это Хамовники, Якиманка и Арбат. А самые дешевые квадратные метры в деревне Пучково и поселке Рогово и т. д. Некоторые значения не входят в районы Москвы, но являются частью Новой Москвы.

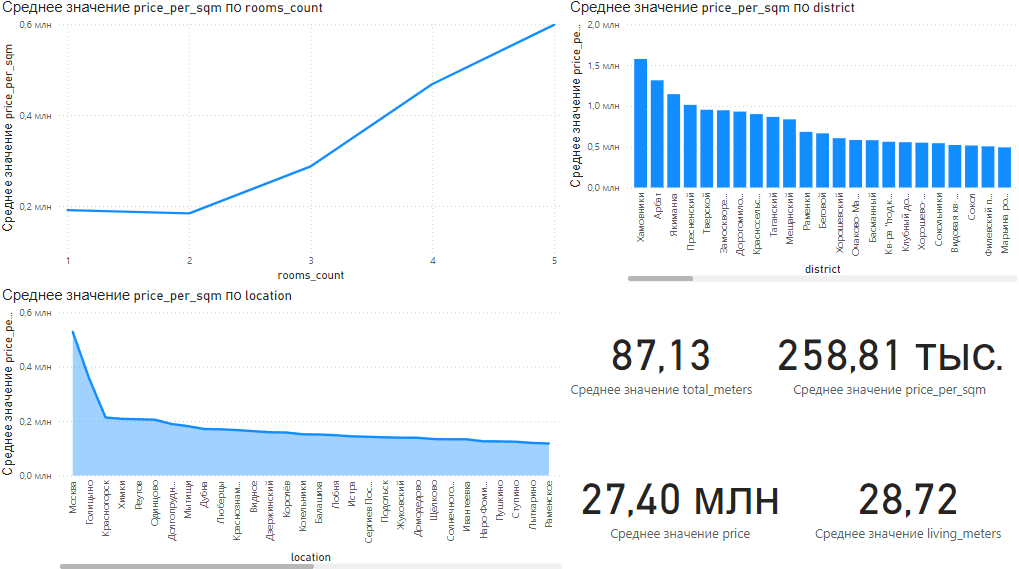
Последним я сделал корреляционную матрицу, которая показывает корреляции между всеми переменными.

Код этой корреляционной матрицы:



Результат этого кода и сама корреляционная матрица:

Цель нашего отчета - это понять от чего зависит цена на квартиру. Т. к. цена может быть разной от общей площади квартиры, то лучше всего смотреть от чего зависит какое-то универсальное значение, в нашем случае это — цена за квадратный метр. Смотря на корреляционную матрицу и на последнюю колонку, которая отвечает именно за корреляцию цены за квадратный метр, можно сказать что **цена квадратного метра зависит от общей площади квартиры, количества комнат в квартире, года постройки дома, местоположения, жилищного метража и метража кухни.**

Так же была выполнена работа в Power BI. Там я создал 3 графика и 4 значения: 

Слева-сверху график показывающий зависимость цены квадратного метра от кол-ва комнат в квартире; справа-сверху график показывающий зависимость цены квадратного метра от района, где расположена квартира; слева-снизу график показывающий цену квадратного метра в городах; справа-снизу 4 значений, показывающие среднюю общую площадь квартир, среднюю цену квадратного метра, среднюю цену квартиры и среднюю жилищную площадь.

**Заключение**

В этом аналитическом отчете были проделаны работы, такие как: парсинг данных, отбор и чистка данных, визуализация и построение модели. Главная цель отчета была выполнена и определяется как: **цена квадратного метра зависит от общей площади квартиры, количества комнат в квартире, года постройки дома, местоположения, жилищного метража и метража кухни.**

В будущем в этот аналитический отчет можно добавить новые параметры и данные, доработать очистку данных, сгенерировать новую визуализацию и исправить недочеты.