Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Аналитический отчет**

***«Обработка данных и моделирование цен на квартиры сайта циан»***

Работу выполнил:

Студент группы ИСП-21

Бойчук Карина

Проверила:

преподаватель

Базяк Г.В.

Долгопрудный, 2024

# Введение

**Целью** данного отчета является анализ рынка недвижимости в Москве и Московской области с целью выявления различий в ценовой динамике и спросе на жилье в этих двух регионах. В свете текущих экономических и общественно-политических событий, такой анализ является актуальным и важным для граждан, планирующих приобрести недвижимость в данном регионе.

Отчет будет включать в себя данные о ценовой динамике на рынке недвижимости, спросе на жилье, особенностях предложения и другие важные аспекты, необходимые для принятия обоснованных решений о покупке жилья в Москве и Московской области. В результате исследования мы надеемся выявить различия и сходства между этими двумя регионами, а также предоставить полезную информацию для потенциальных покупателей недвижимости.

Исходя из поставленной цели, можно сформировать следующие **задачи**:

* Провести парсинг данных по квартирам на продажу;
* Подготовить данные для анализа;
* Визуализировать взаимосвязь между параметрами и определить признаки, оказывающие наибольшее влияние.

# МЕТОДОЛОГИЯ

В данном исследовании использовались методы и инструменты для сбора и анализа данных, полученных с сайта недвижимости Циан. Основные этапы методологии включают сбор данных, их предварительную обработку, анализ и визуализацию результатов. Ниже представлены подробности каждого из этапов.

1.1 Сбор данных

Данные были собраны с сайта Циан с использованием парсинг. Для этого был применен Python-библиотека cianpraser, которая позволяет извлекать данные из HTML-кода веб-страниц. Сбор данных включал в себя следующие шаги:

* Определение целевых страниц для сканирования (например, страницы с объявлениями о продаже и аренде недвижимости).
* Извлечение необходимых атрибутов, таких как цена, площадь, количество комнат, местоположение и другие характеристики объектов недвижимости.
* Сохранение собранных данных в формате CSV для дальнейшего анализа.

1.2 Выбор респондентов

В данном случае респондентами являются объекты недвижимости, представленные на сайте Циан. Выборка была сделана исходя из места положения объектов, что позволяет обеспечить выборку.

1.3 Предварительная обработка данных

Данные были загружены в Google Colab, где была выполнена предварительная обработка:

* Удаление дубликатов.
* Преобразование типов данных (например, преобразование цен из строкового формата в числовой).
* Обработка пропущенных значений.

Для анализа пропущенных данных были использованы следующие методы:

1. Процентный список пропущенных данных

С помощью Python и библиотеки pandas был вычислен процент пропущенных значений для каждого признака. Это позволяет быстро оценить качество собранных данных.

1. Гистограмма пропущенных данных

Гистограмма была построена для визуализации распределения пропусков по признакам. Это дает представление о том, какие признаки имеют наибольшее количество пропусков.

1. Тепловая карта

С помощью библиотеки seaborn была построена тепловая карта для визуализации корреляции между признаками и наличием пропусков. Это помогает выявить взаимосвязи между различными переменными.

1.4 Визуализация данных с помощью Power

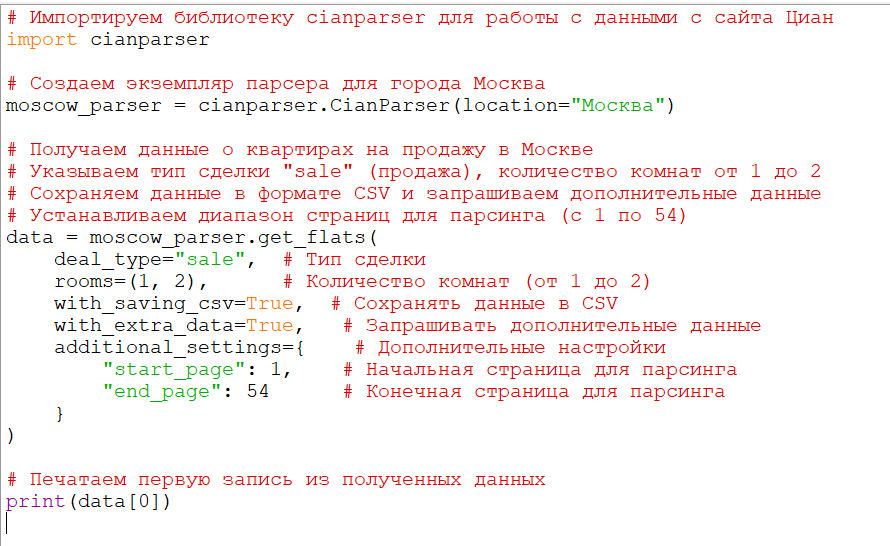
BI Для более глубокого анализа и визуализации результатов использовался Power BI:

* Построены интерактивные дашборды, позволяющие исследовать данные по различным фильтрам (например, по регионам или типам недвижимости).
* Созданы диаграммы, показывающие распределение цен, площади и других характеристик объектов недвижимости.

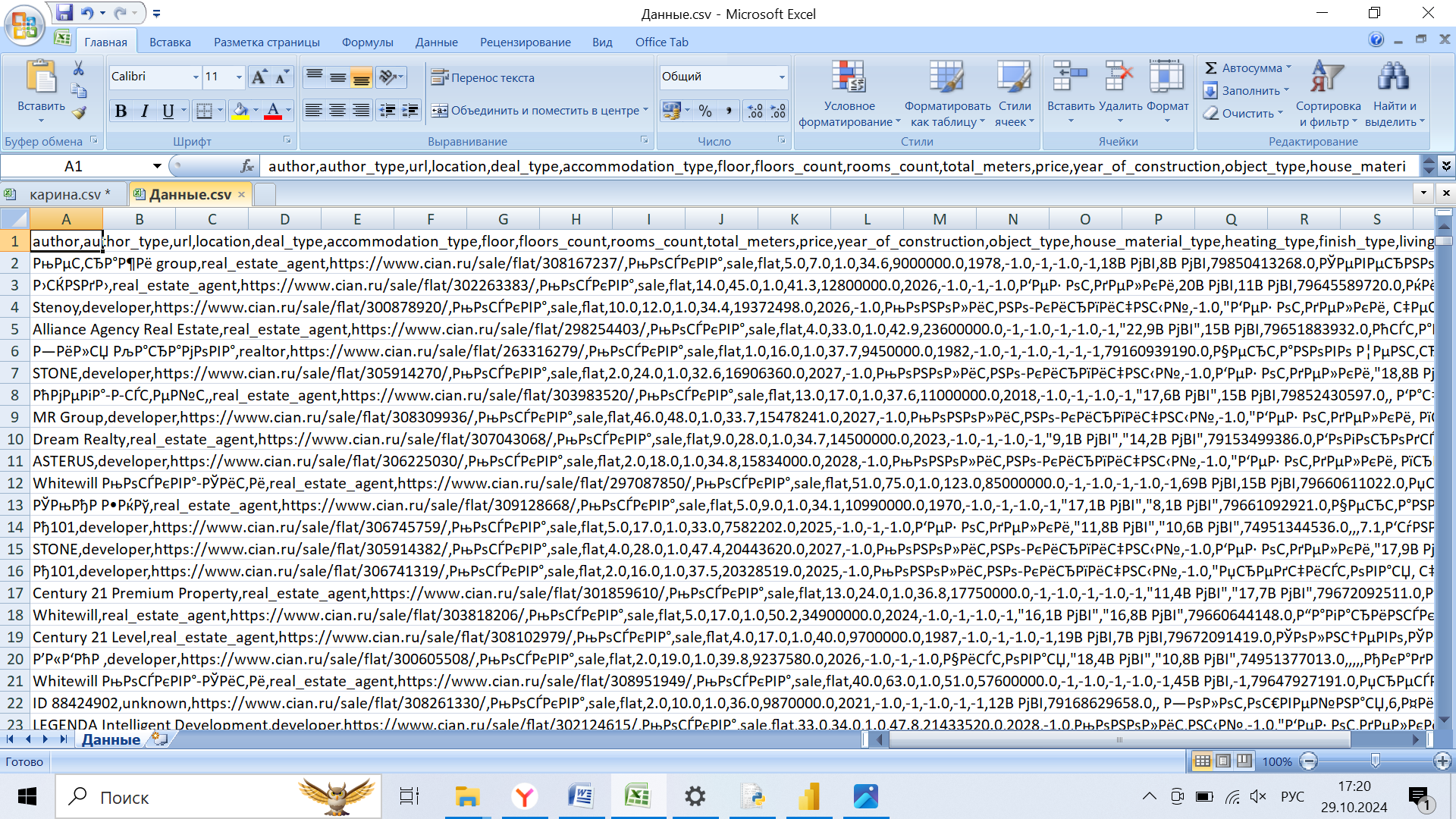
# ОСНОВНАЯ ЧАСТЬ

Для сбора данных нам была предоставлена библиотека cianpraser, предназначенная для извлечения информации с сайта объявлений об аренде и продаже недвижимости Циан.

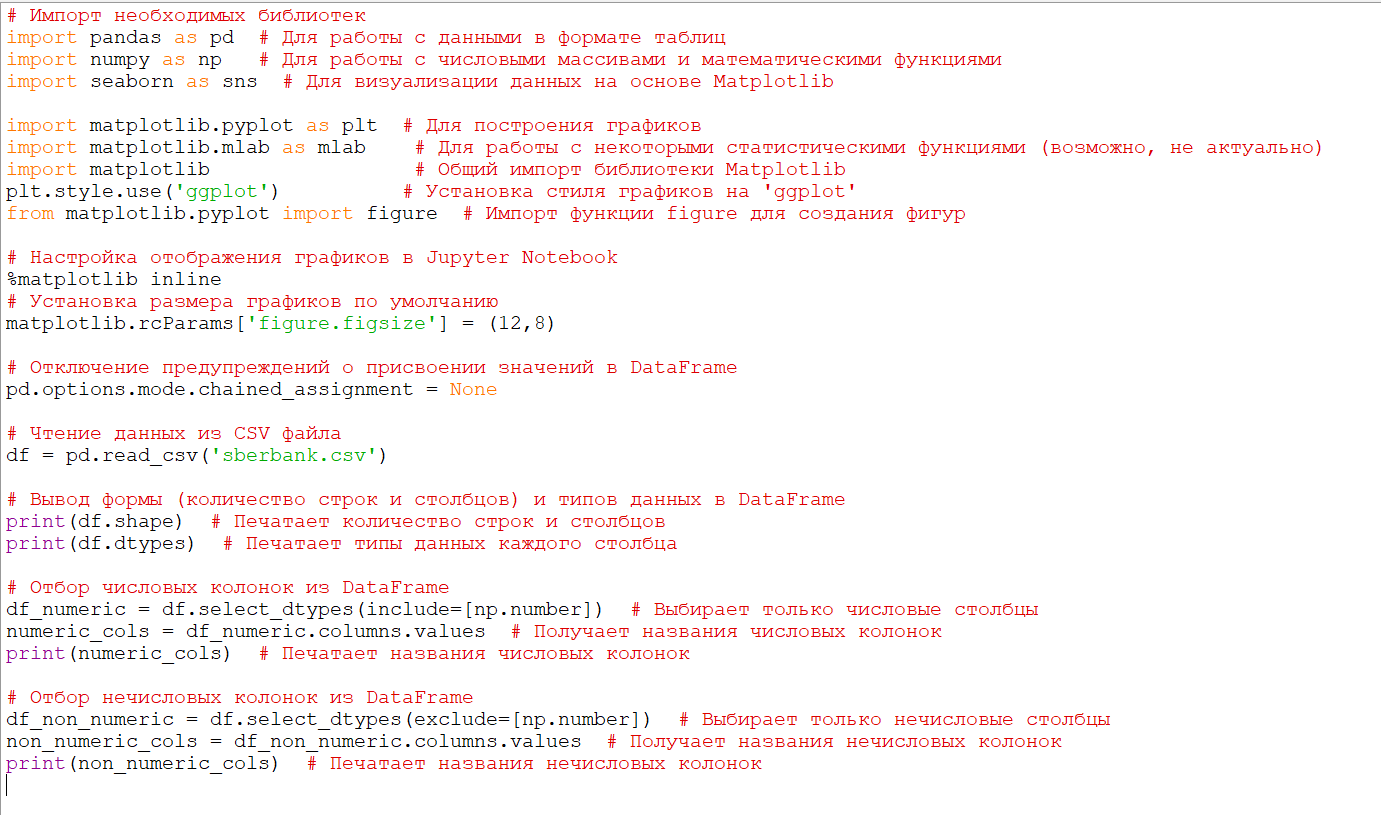
Парсер



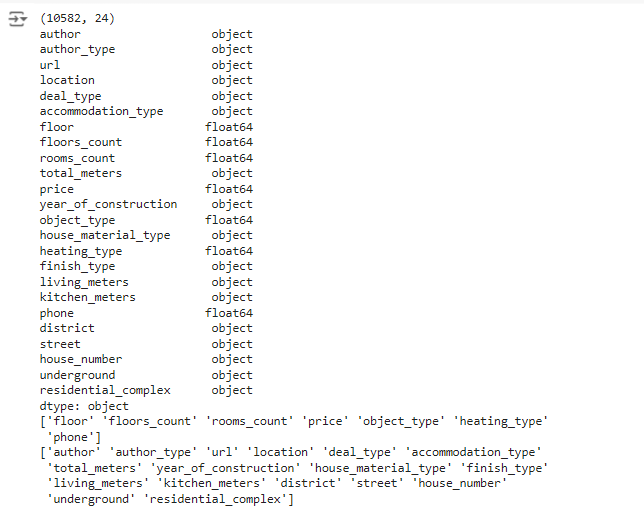
По итогам парсинга мы (с одногруппницей) получили неочищенные данные, содержащие дубликаты, пропуски и другие некорректные значения. Необходимо провести очистку данных, чтобы улучшить их качество и подготовить для дальнейшего анализа.



Для дальнейшей работы с данными я решила воспользоваться Google Colab, так как эта платформа предоставляет удобные инструменты для обработки и анализа данных. С помощью Google Colab я смогла загрузить необходимые наборы данных, а затем применить различные методы преобразования и предобработки. Это включает в себя очистку данных, обработку пропусков и нормализацию значений. Использование Google Colab также позволяет мне легко делиться результатами и сотрудничать с коллегами, что значительно упрощает процесс анализа.

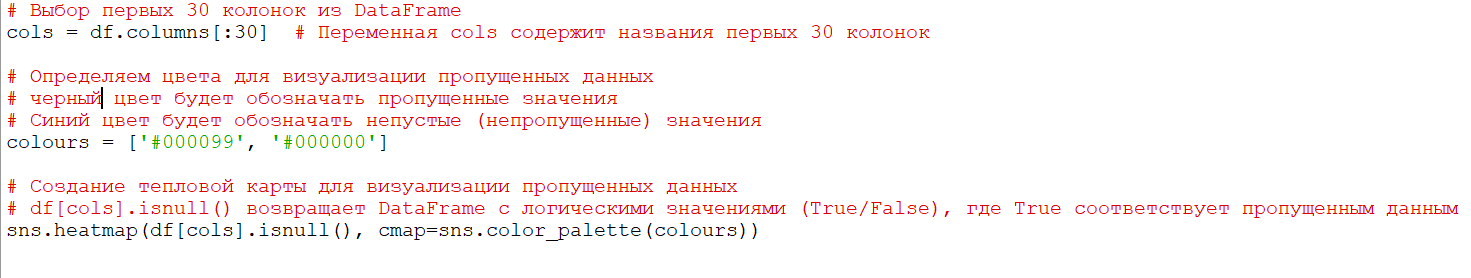


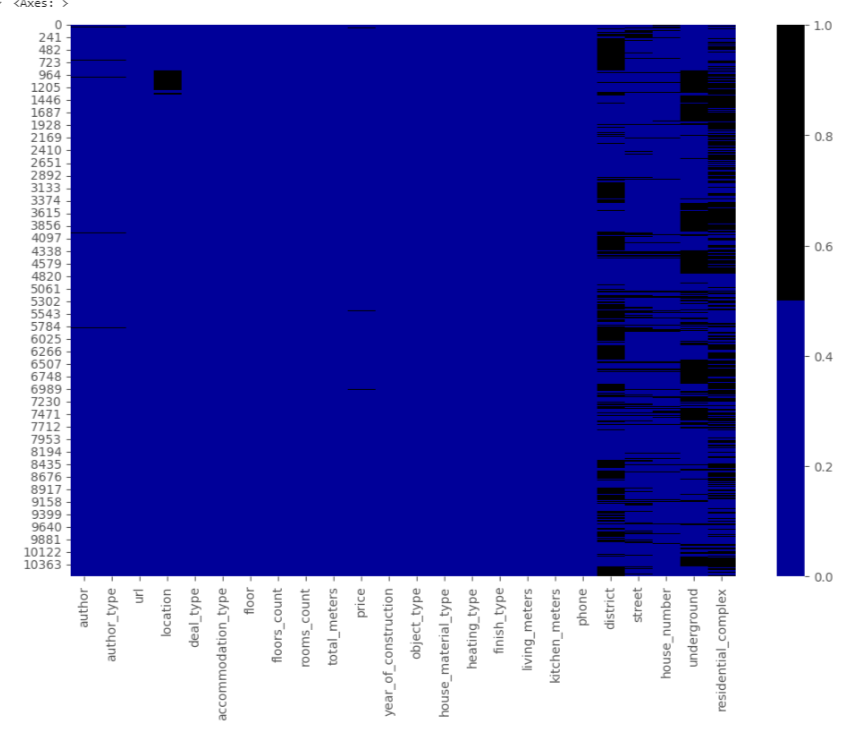
Выводим какие данные у нас есть и их тип:



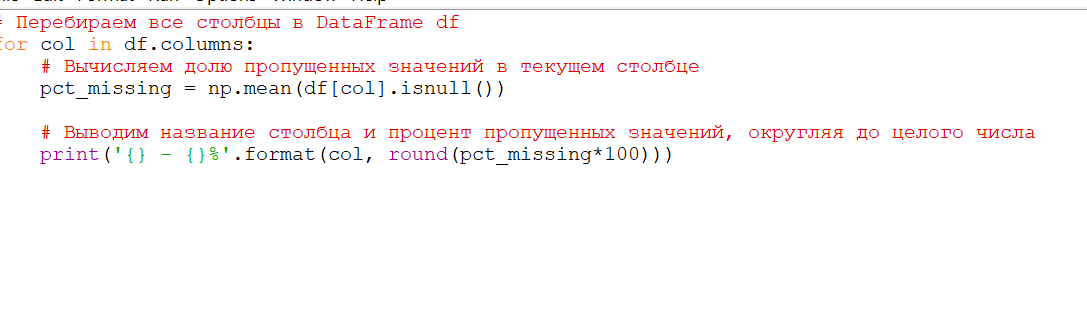
Этот код покажет нам, что набор данных состоит из 10582 строки и 24 столбцов. Теперь мы можем пробежаться по чек-листу.

**Визуализируем пропущенные значения с помощью тепловой карты**.

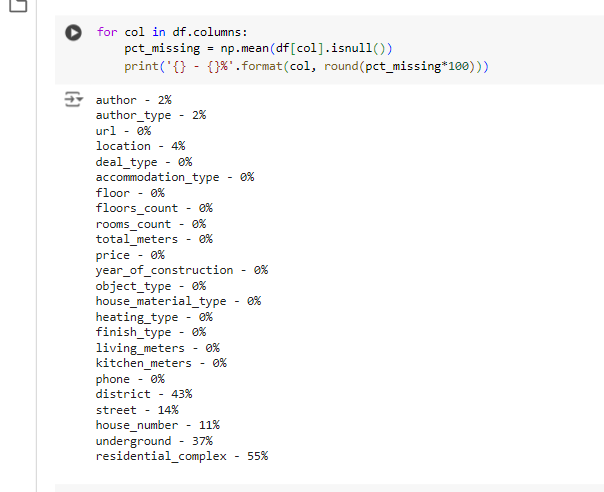




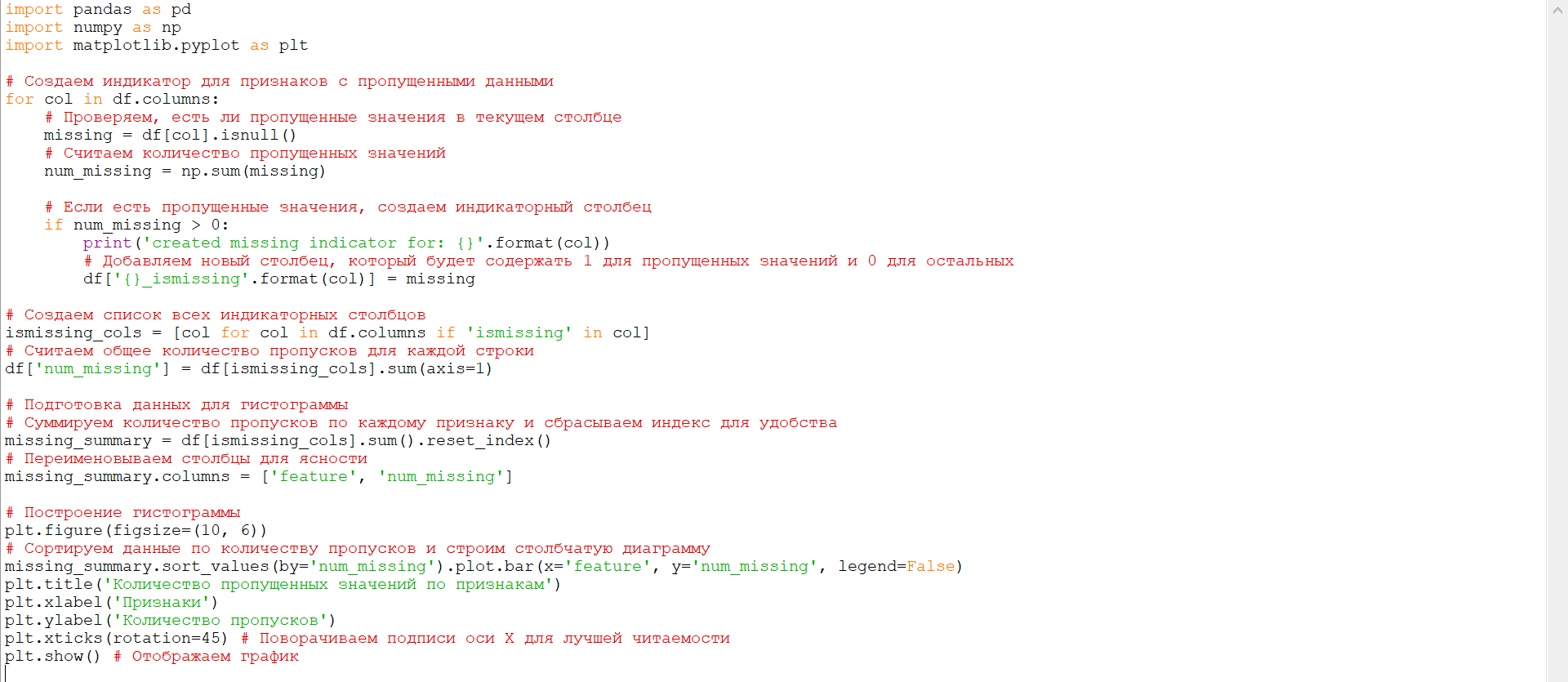
**Процентный список пропущенных данных**

****

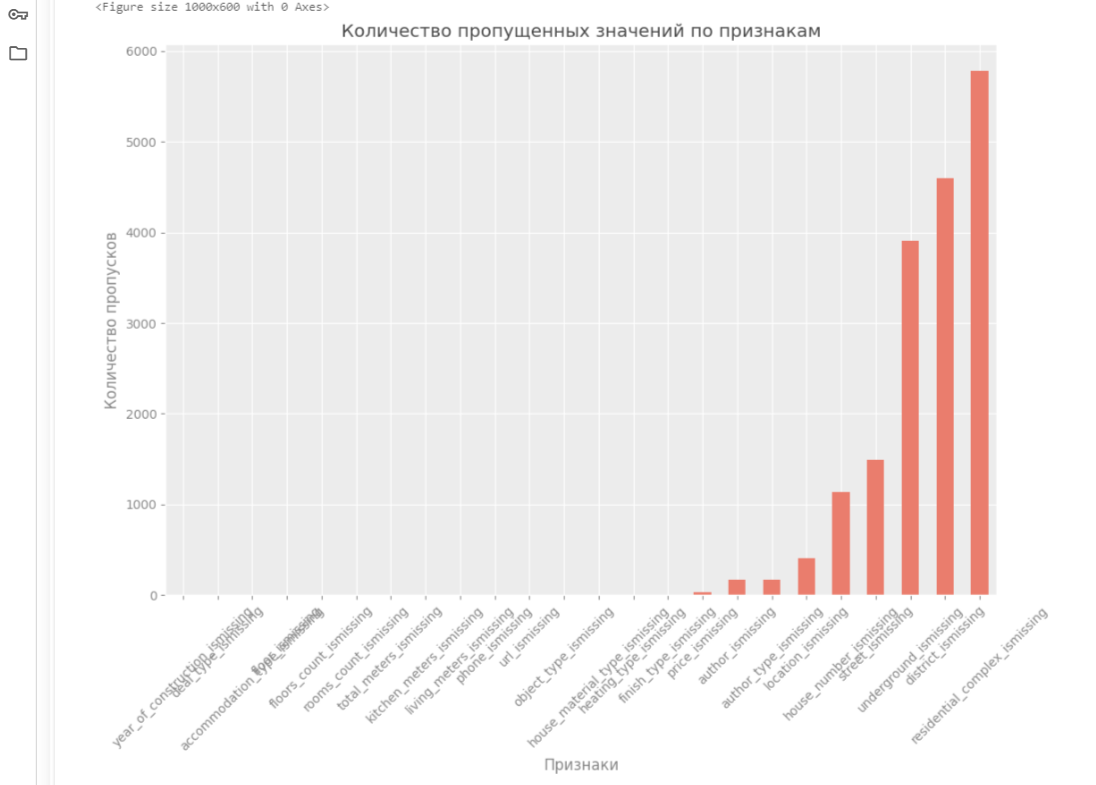
Вывод



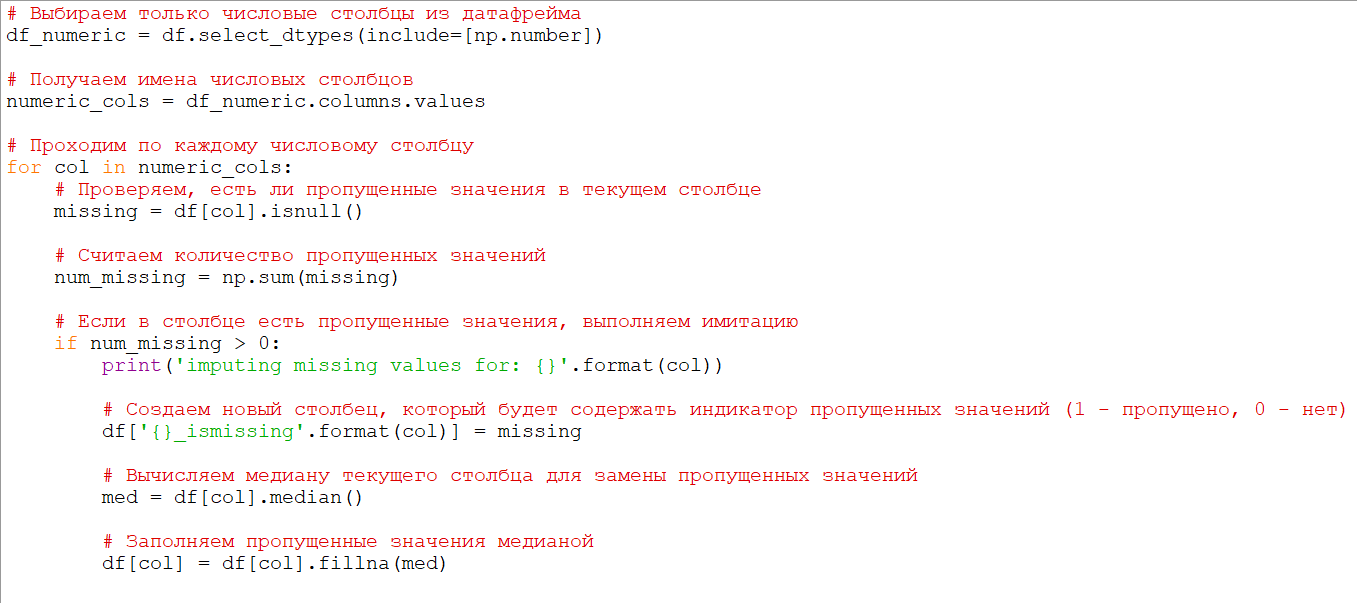
**Гистограмма пропущенных данных**



Вывод

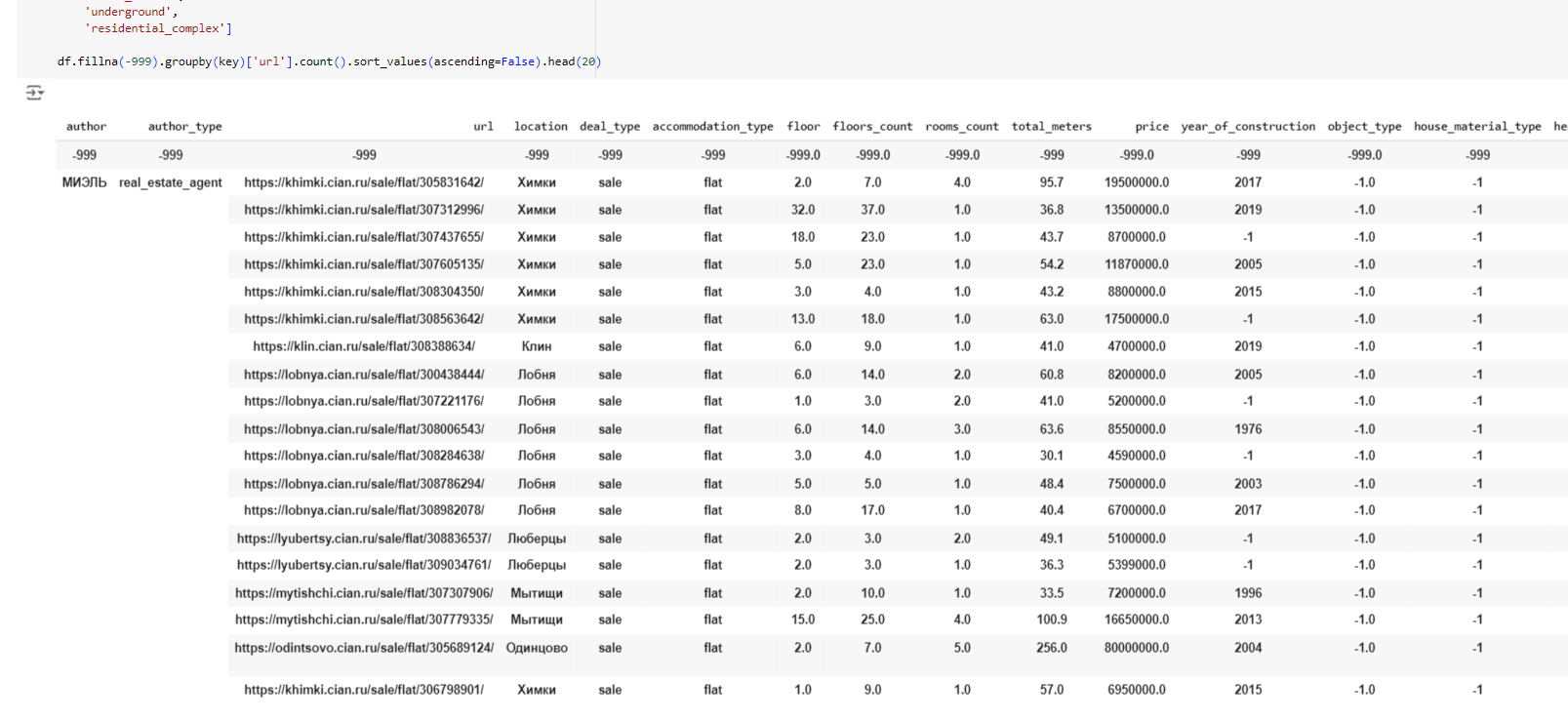


В процессе анализа данных я пришла к выводу, что удаление каких-либо значений недопустимо, так как все данные имеют важное значение для точного прогнозирования цен. Каждое наблюдение может вносить свой вклад в общее понимание динамики рынка, и потеря информации может негативно сказаться на качестве анализа. Поэтому я приняла решение заполнить пропущенные значения принудительно. Этот подход позволит сохранить целостность данных и обеспечит более надежные результаты при построении модели.

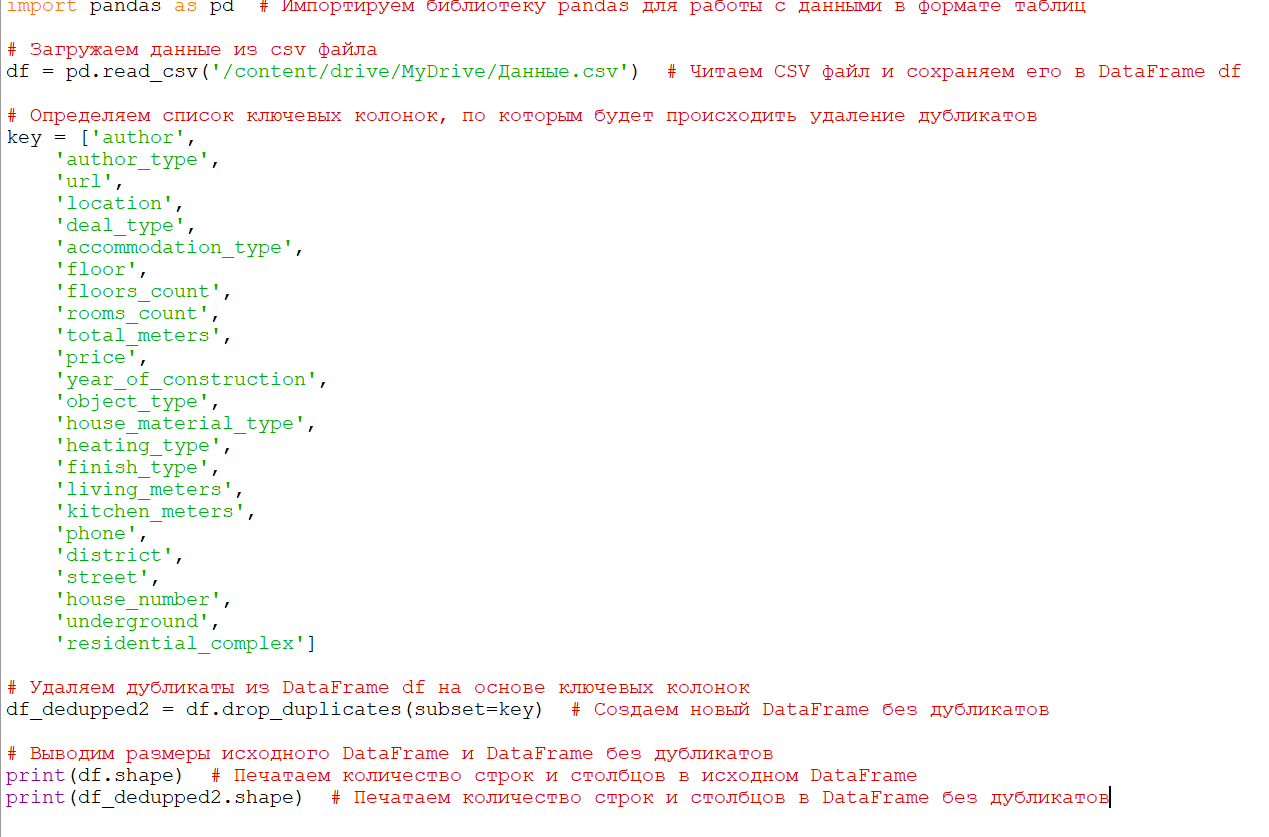


Очевидно, что повторяющиеся записи нам не нужны, значит, их нужно исключить их из набора.

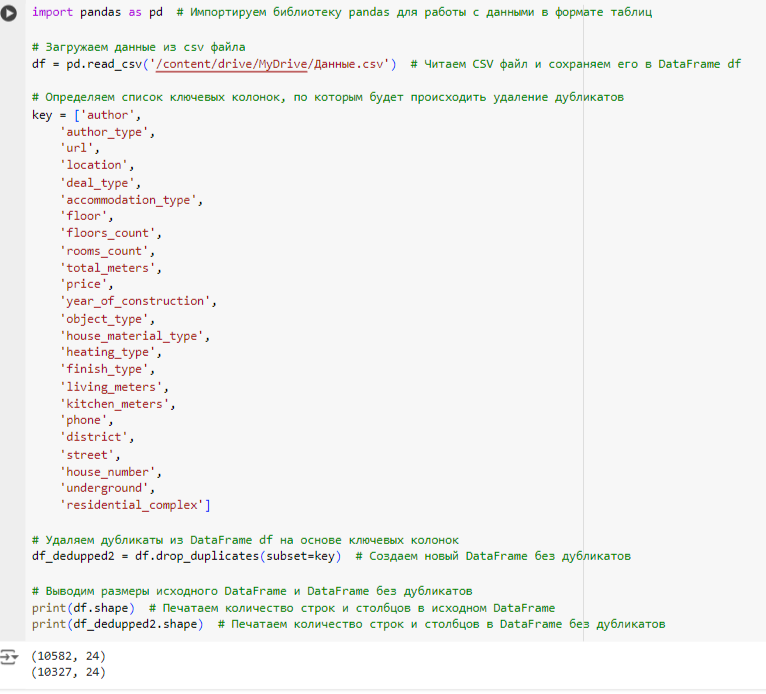




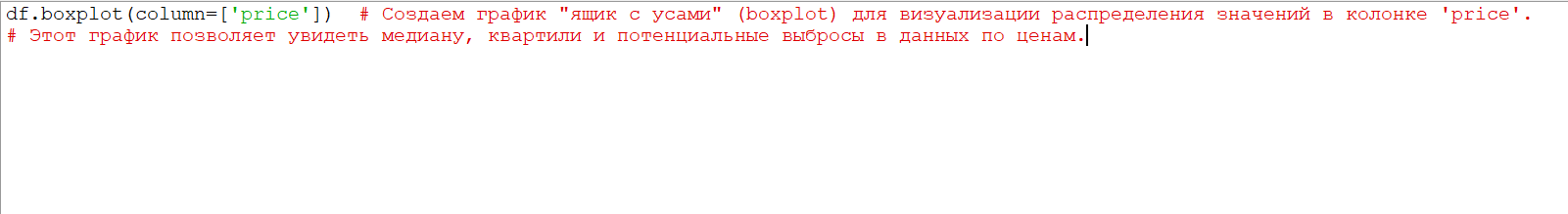
Мы получили дубликаты, теперь нужно их удалить

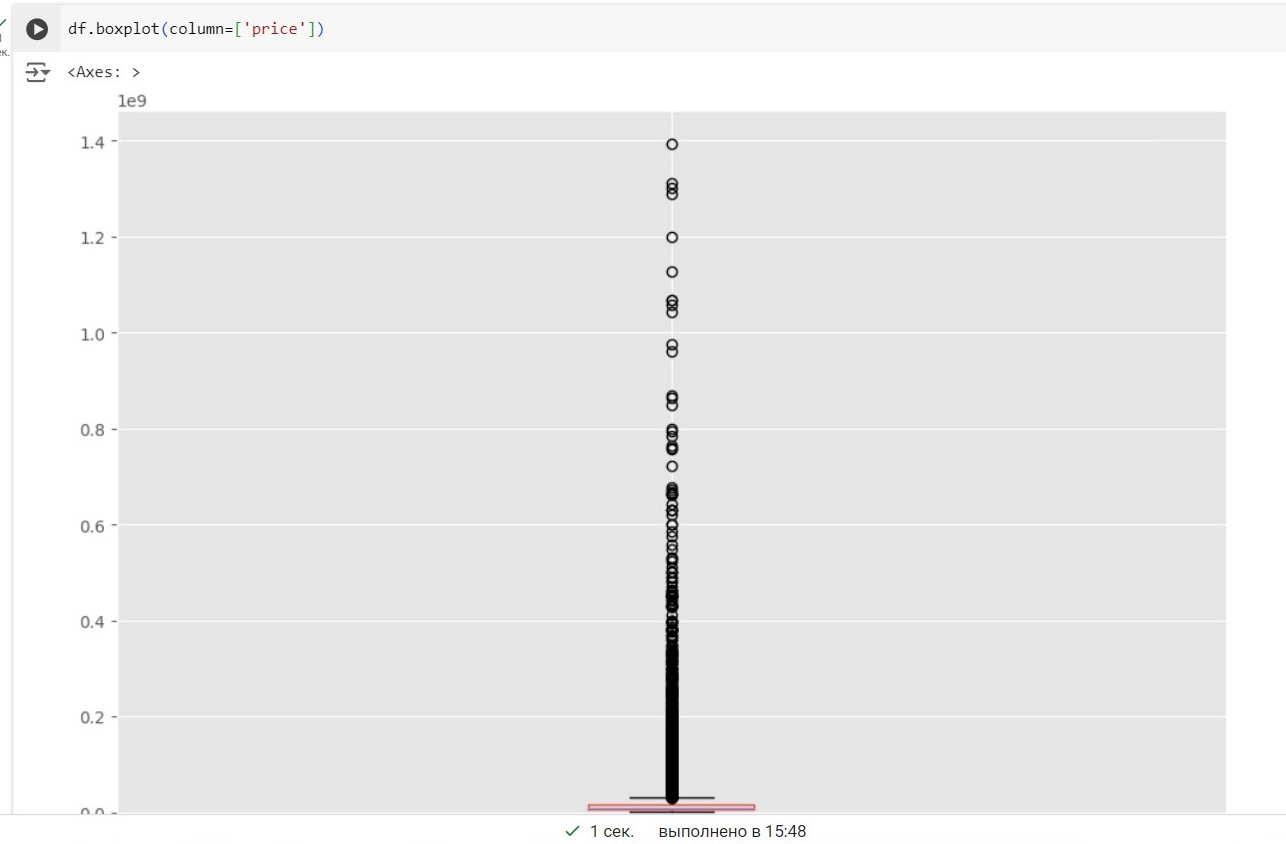


Мы удалили дублированные записи и на выходе видим, что данных стало меньше



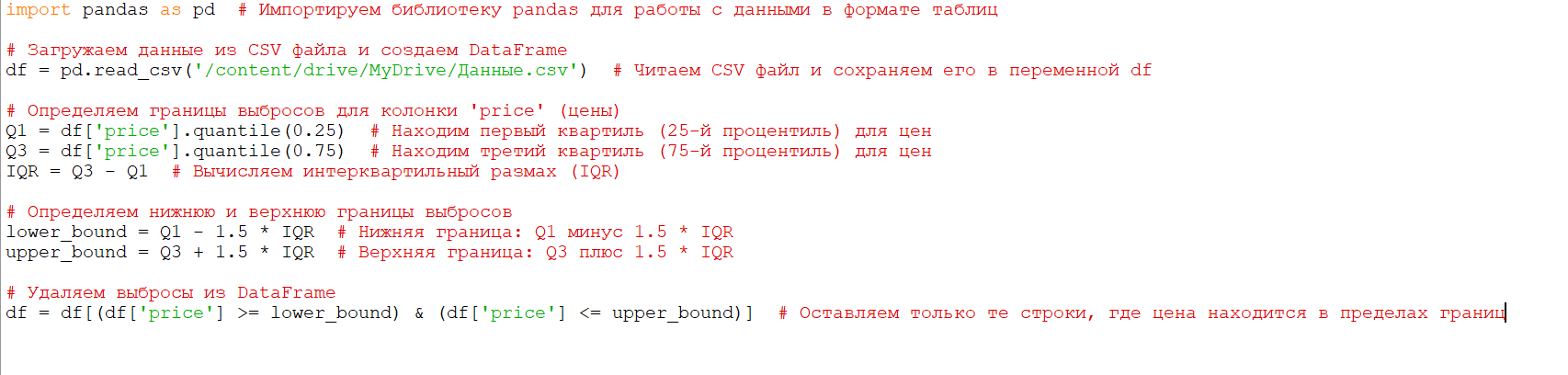
Т.к. цена ключевое в нашем исследовании я решила проверить выброс именно на нее

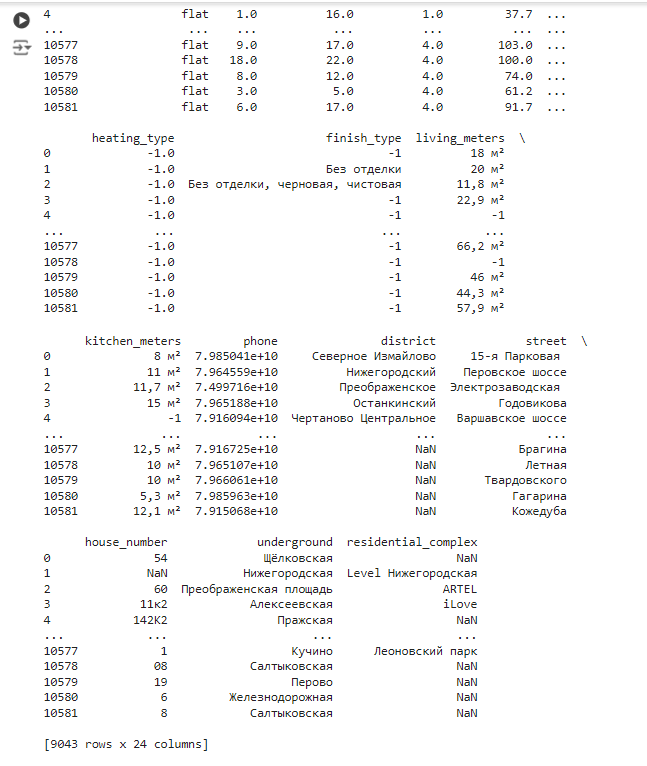




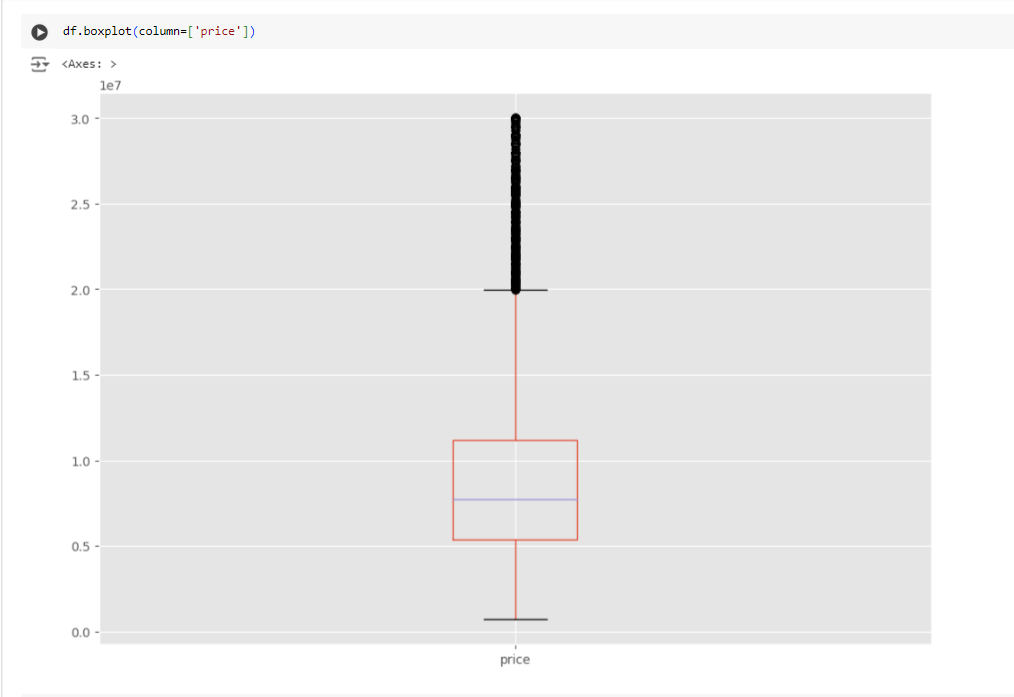
По выбросу я сделала вывод цена начиная от 0.8 слишком велики и могут помешать в объективной оценке данных

Поэтому я взяла диапазон и все не входящие в него данные удалила

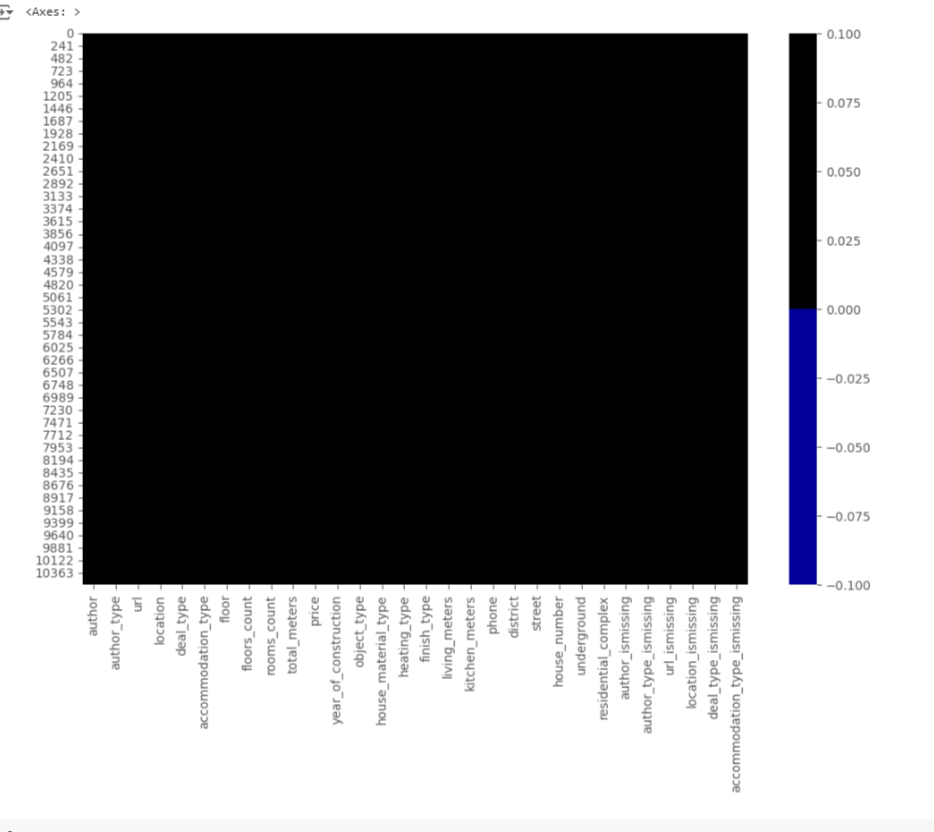




Данные сократились до 9043 и теперь на выбросе мы можем увидеть такую картину

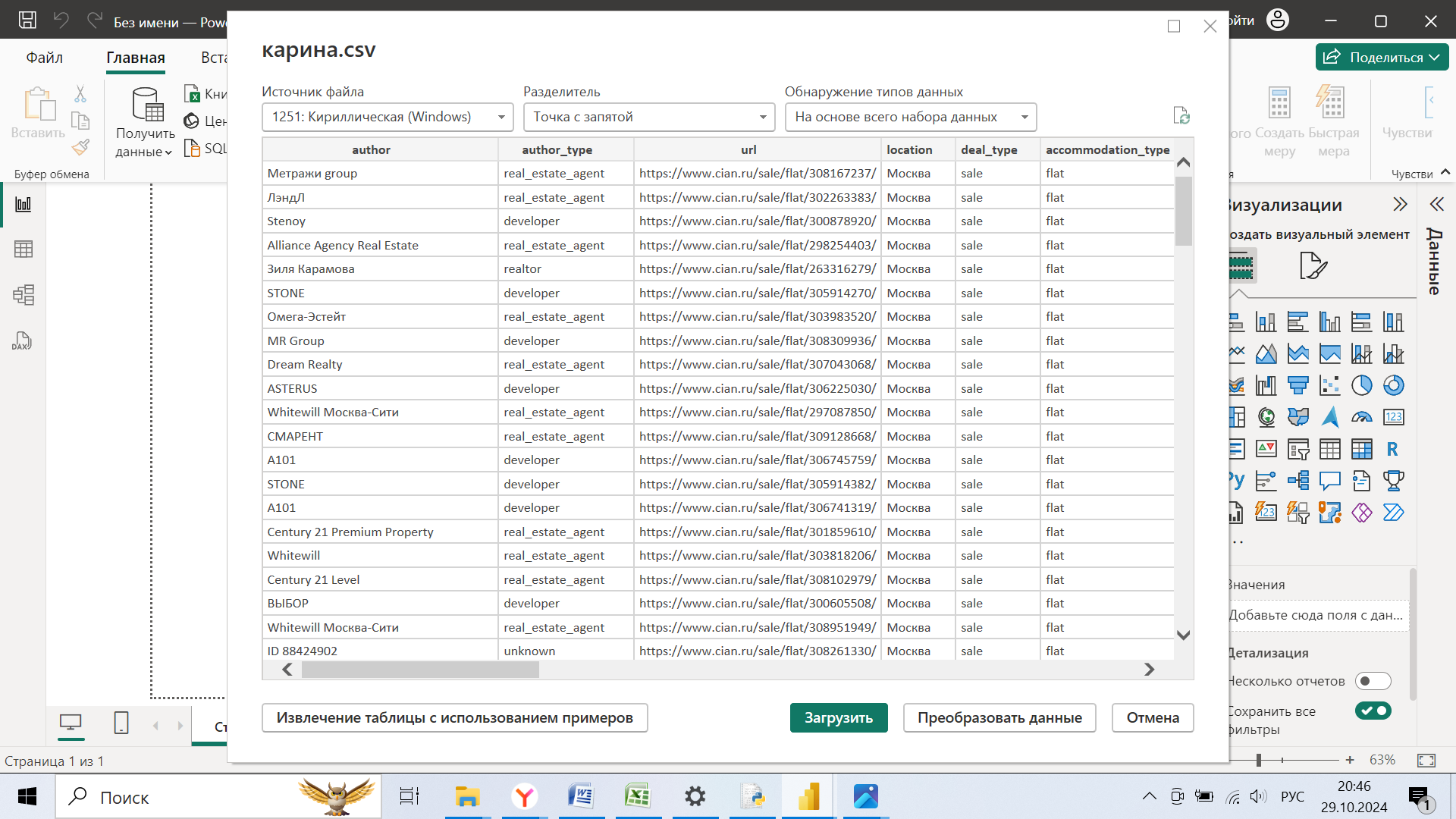


Мы получили корректные данные



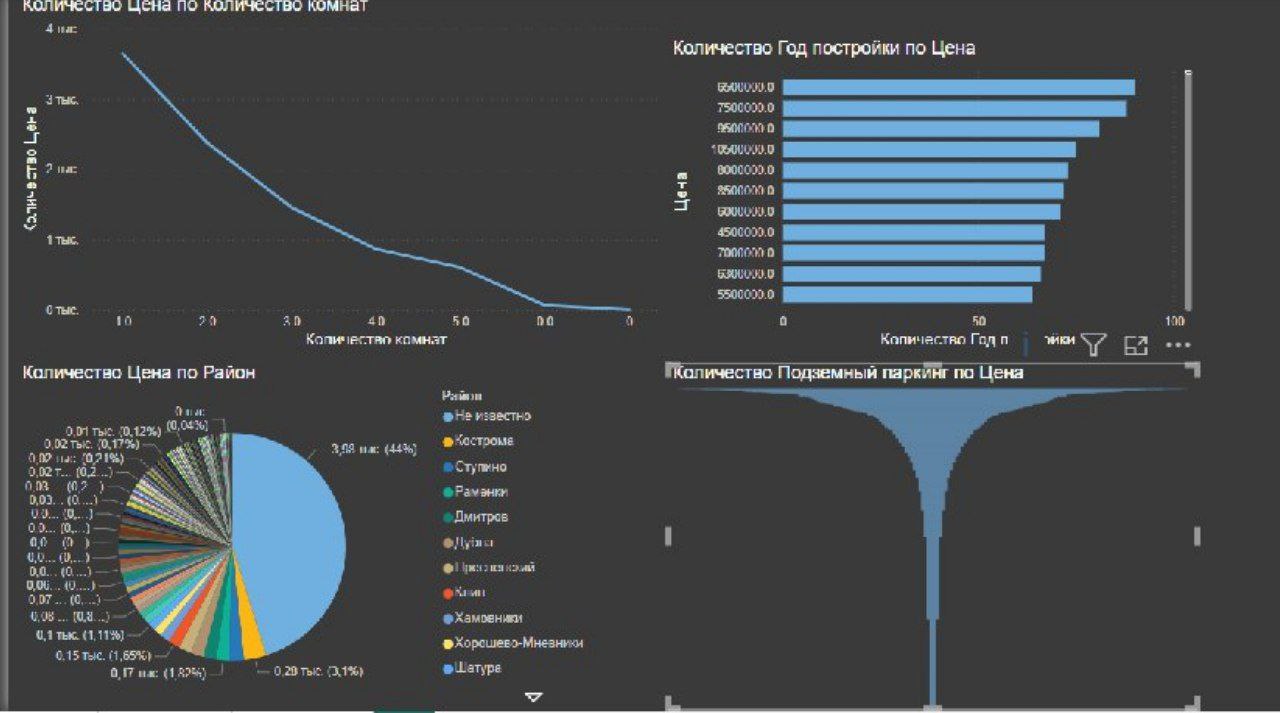


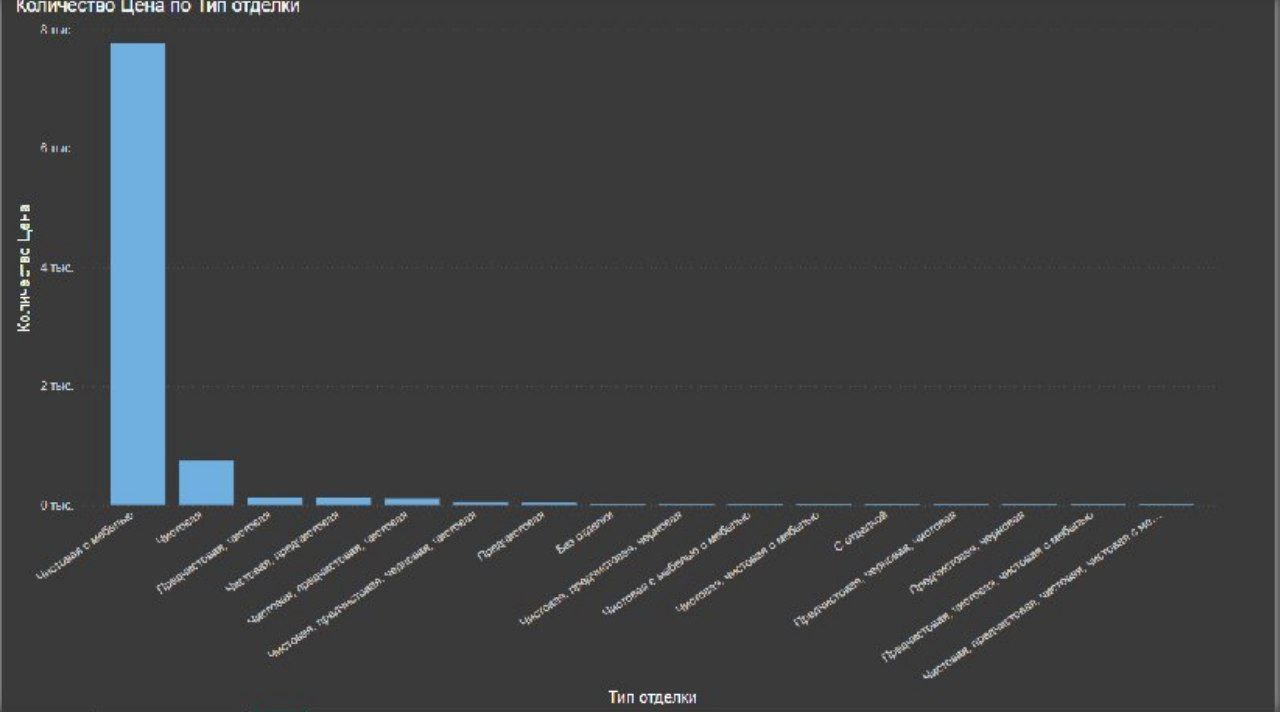
Дальнейшая обработка прошла в Power BI

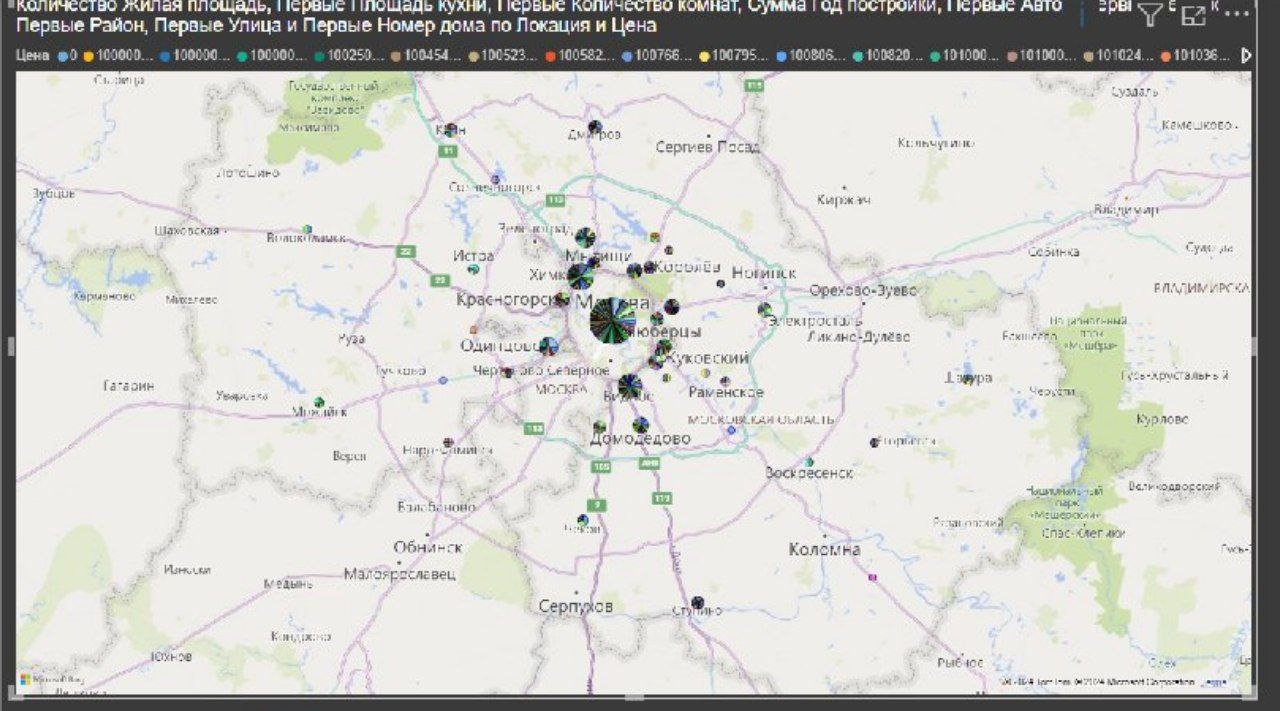


На стадии загрузки данных приняла решение удалить строки object\_type и heating\_type, т.к. они имели лишь одно значение (-1) которое приравнивается к значению Non.

Дашборд







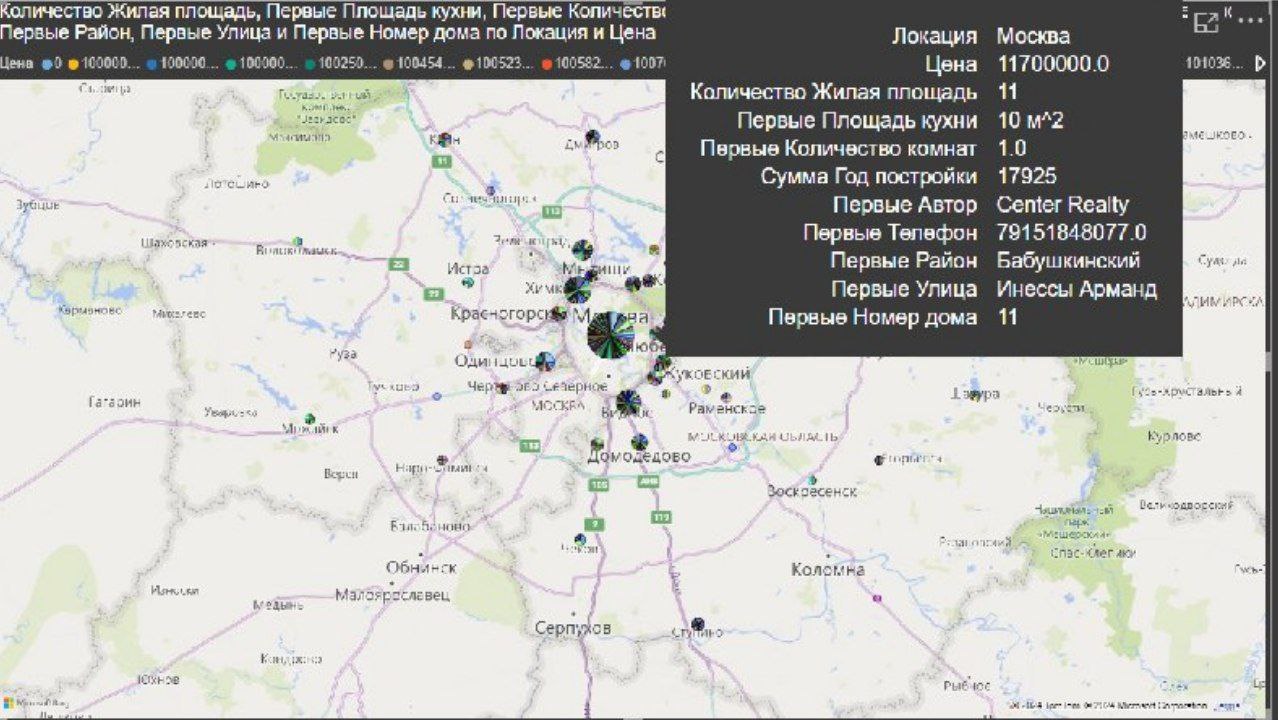
# АНАЛИЗ ДАННЫХ

Зависимость цены квартиры от отделки является одним из ключевых факторов, влияющих на ее стоимость. Отделка квартиры включает в себя такие элементы, как качество строительных материалов, сантехническое оборудование, отделка стен и полов, двери, окна и т.д.

Квартиры с дорогостоящей и качественной отделкой обычно имеют более высокую цену, поскольку покупатели ценят уровень комфорта, удобства и престижности жилья. Такие квартиры часто требуют минимальных дополнительных вложений после покупки, что делает их более привлекательными для покупателей.

С другой стороны, квартиры с неотделанными или низкокачественными отделочными материалами обычно имеют более низкую цену. Такие квартиры могут потребовать дополнительных финансовых затрат на ремонт и модернизацию, что может отразиться на итоговой стоимости жилья.

Поэтому при осмотре квартиры и принятии решения о ее покупке важно учитывать не только текущее состояние отделки, но и ее качество, поскольку это существенно влияет на ее стоимость и комфортность проживания. График, отображающий зависимость цены квартиры от отделки, поможет потенциальным покупателям принять взвешенное решение и оценить потенциальные затраты на дополнительные работы по улучшению жилья.



Данная карта представляет собой график зависимости цены квартиры от ее расположения по Москве и Московской области, а также от общей площади квартиры. Изучив данный график, можно сделать ряд выводов.

Во-первых, можно отметить, что цена квартиры сильно зависит от ее расположения. В данном случае видно, что квартиры, расположенные в центре Москвы имеют более высокую цену, чем квартиры на окраине или в пригородах Московской области. Это может быть связано с близостью к центру города, наличием развитой инфраструктуры и транспортной доступностью.

Во-вторых, можно заметить, что цена квартиры также зависит от ее общей площади. Обычно чем больше площадь квартиры, тем выше ее цена. Это связано с тем, что крупные квартиры обычно имеют более высокий спрос, так как могут удовлетворить потребности семей с детьми или людей, желающих больше пространства.

Таким образом, анализируя данную карту, можно сделать вывод о том, что цена квартиры зависит от ее расположения и общей площади.

# Заключение

В ходе проведенного анализа были исследованы различные факторы, оказывающие влияние на цену квадратного метра недвижимости в Московском регионе. Также оказалось, что жилье в крупных городах более дорогое, чем за пределами городов.

Было обнаружено, что наличие парковки во дворе или подземном гараже также оказывает влияние на цену квадратного метра недвижимости. Квартиры с собственным местом для автомобиля обычно более ценятся покупателями и могут иметь более высокую стоимость.

Кроме того, количество комнат и общая площадь квартиры также влияют на цену за квадратный метр. Например, квартиры с большим количеством комнат или общей площадью могут иметь более высокую цену за квадратный метр, чем квартиры с меньшим количеством комнат или площадью.

Таким образом, при определении цены квадратного метра недвижимости в Московском регионе необходимо учитывать не только расположение квартиры и наличие ремонта, но и другие факторы, такие как наличие парковки, количество комнат и общая площадь квартиры.