ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(национальный исследовательский университет)» (МАИ)

Направление подготовки: 27.03.05 «Инноватика»

**Лабораторная работа №6**

по дисциплине «Прикладные методы и алгоритмы обработки и защиты данных»

Выполнил:

Студент гр. М3О-236Б-22

Попов И.А., Стaндровский П.Д.

Преподаватель:

Александрова С.С.

Москва 2023

Оглавление

[**Цель работы**: 3](#_Toc150722510)

[Шаг 1. Откройте файл с данными и изучите общую информацию 3](#_Toc150722511)

[Шаг 2. Предобработка данных 3](#_Toc150722512)

[Шаг 3. Посчитайте и добавьте в таблицу 5](#_Toc150722513)

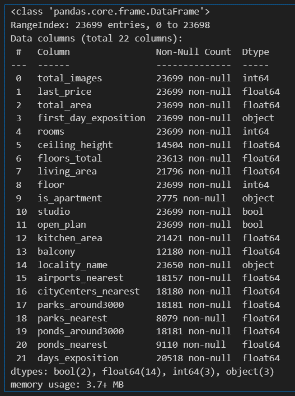
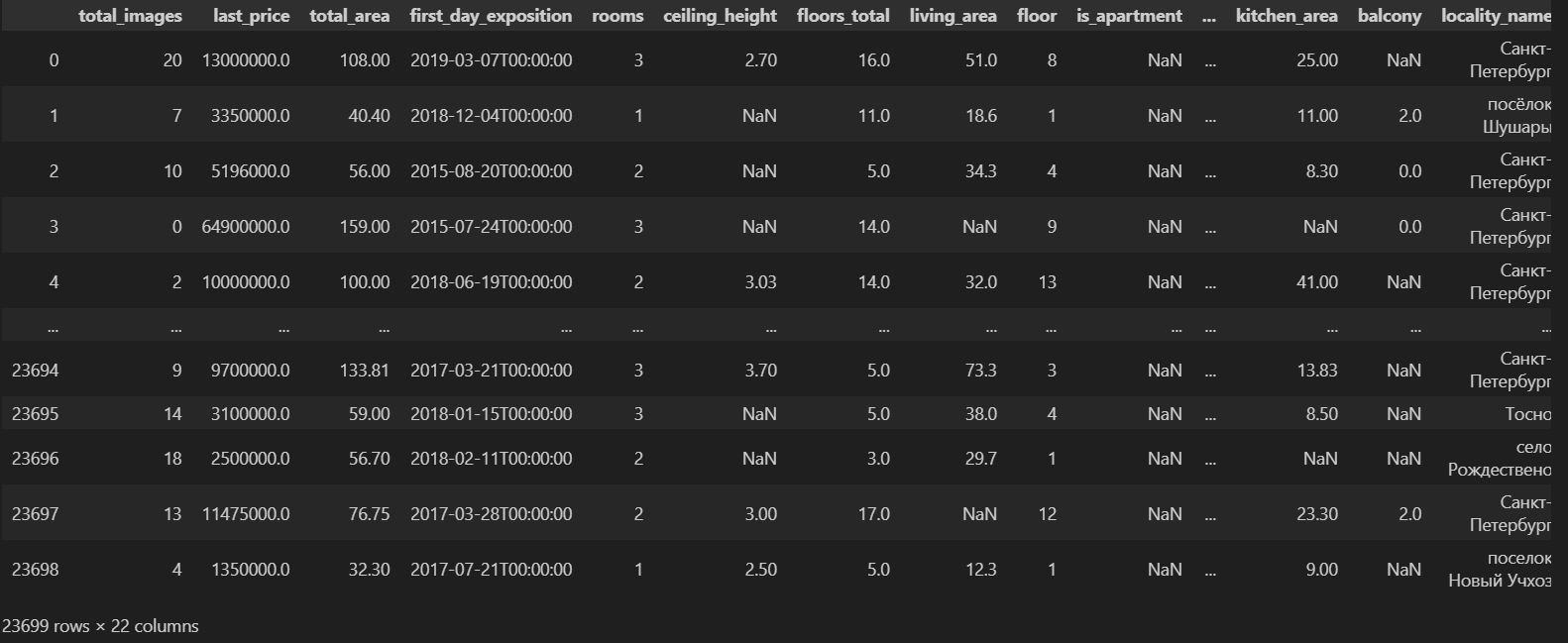
[Шаг 4. Проведите исследовательский анализ данных и выполните инструкции 6](#_Toc150722514)

[Ссылка на GitHub: 12](#_Toc150722515)

[**Вывод:** 12](#_Toc150722516)

**Цель работы**: анализ данных о продаже квартир в Санкт-Петербурге и соседних населенных пунктах с целью определения ключевых параметров, влияющих на рыночную стоимость недвижимости.

# Шаг 1. Откройте файл с данными и изучите общую информацию

Из предоставленной информации видно, что в датафрейме содержится 23699 записей о продаже квартир с 22 колонками. Некоторые столбцы имеют пропущенные значения, например, "ceiling\_height", "floors\_total", "living\_area", "is\_apartment" и другие. Некоторые столбцы, такие как "is\_apartament" и "first\_day\_exposition" имеют неправильный тип данных.

# Шаг 2. Предобработка данных

1. Изучение пропущенных значений

Столбец ceiling\_height:

- Пропущенных значений: 9195

- Логичная замена: можно предположить, что высота потолков может быть стандартной. Заполним пропуски медианным значением по всему столбцу.

Столбец floors\_total:

- Пропущенных значений: 86

- Логичная замена: можно предположить, что это частный дом или здание с одним этажом. Заполним пропуски значением 1.

Столбец living\_area:

- Пропущенных значений: 1903

- Логичная замена: можно предположить, что жилая площадь может быть связана с общей площадью и количеством комнат. Заполним пропуски медианным значением внутри каждой комнаты.

Столбец is\_apartment:

- Пропущенных значений: 20924

- Логичная замена: здесь логично предположить, что если не указано, что это апартаменты, то это не апартаменты. Заполним пропуски значением False.

Столбец kitchen\_area:

- Пропущенных значений: 1278

- Логичная замена: можно предположить, что площадь кухни может быть связана с общей площадью. Заполним пропуски медианным значением по всему столбцу.

Столбец balcony:

- Пропущенных значений: 11519

- Логичная замена: если не указано количество балконов, то, вероятно, их нет. Заполним пропуски значением 0.

Столбец locality\_name:

- Пропущенных значений: 249

- Логичная замена: На данном этапе лучше оставить пропуски, так как название населенного пункта может быть важным параметром и не подлежит легкой замене.

Столбец airports\_nearest, cityCenters\_nearest, parks\_around3000, parks\_nearest, ponds\_around3000, ponds\_nearest:

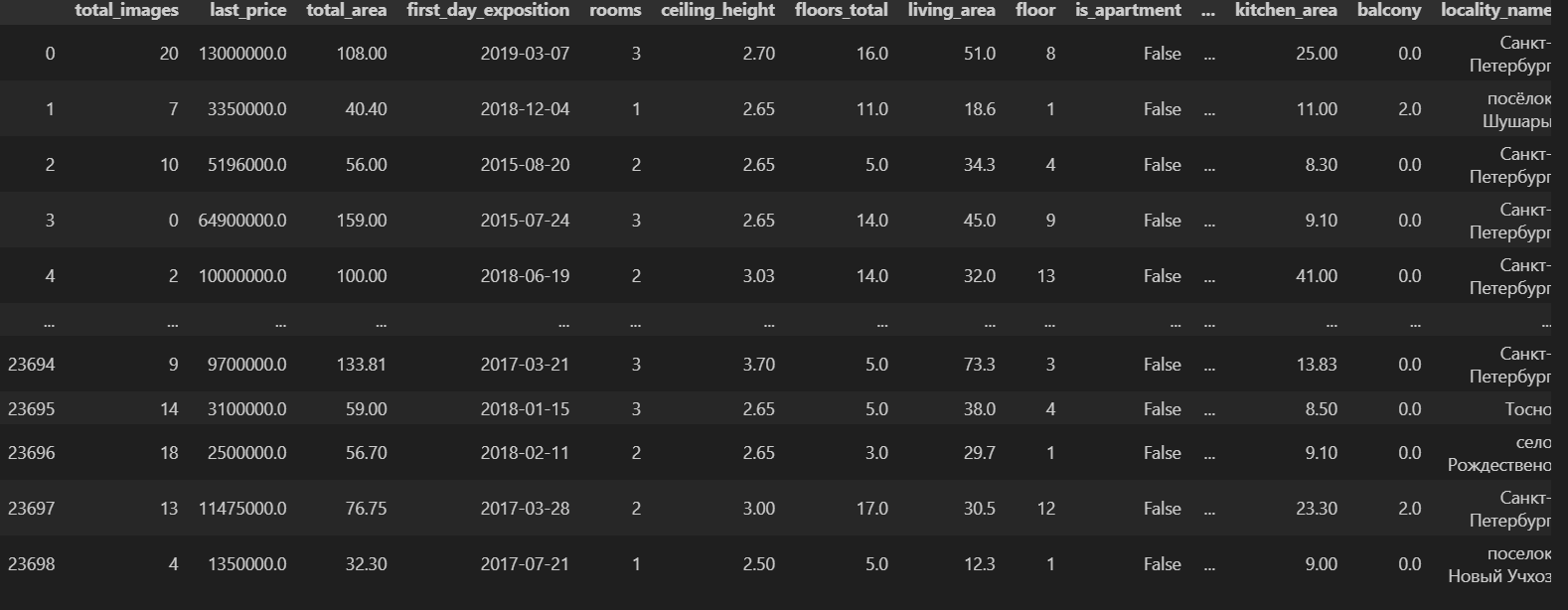
- Возможные логичные замены: Пропуски в этих столбцах могут быть связаны с удаленностью объектов от квартиры. Возможно, эти пропуски можно заполнить медианными значениями.

2. Приведение данных к нужным типам:

Столбец first\_day\_exposition: преобразим в тип datetime, чтобы легче работать с датами.

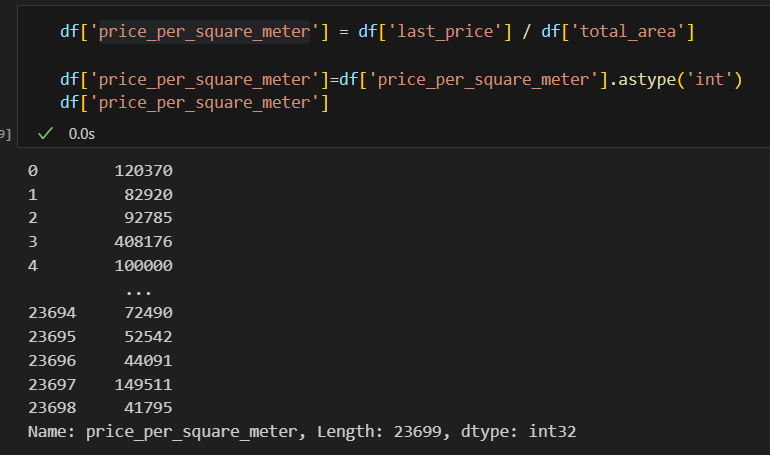
Столбец is\_apartment: преобразим в булев тип, так как этот столбец должен содержать только значения True или False.



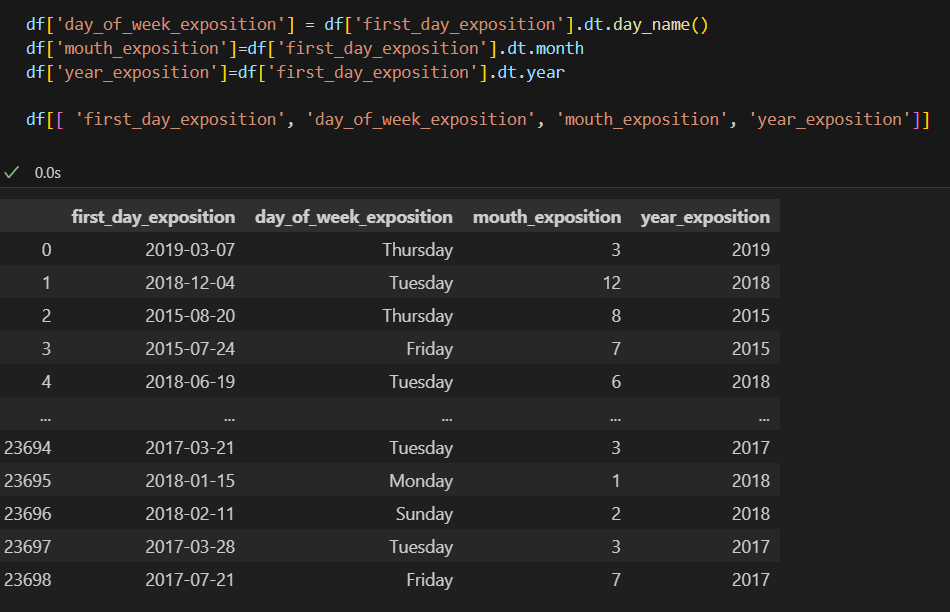


# Шаг 3. Посчитайте и добавьте в таблицу

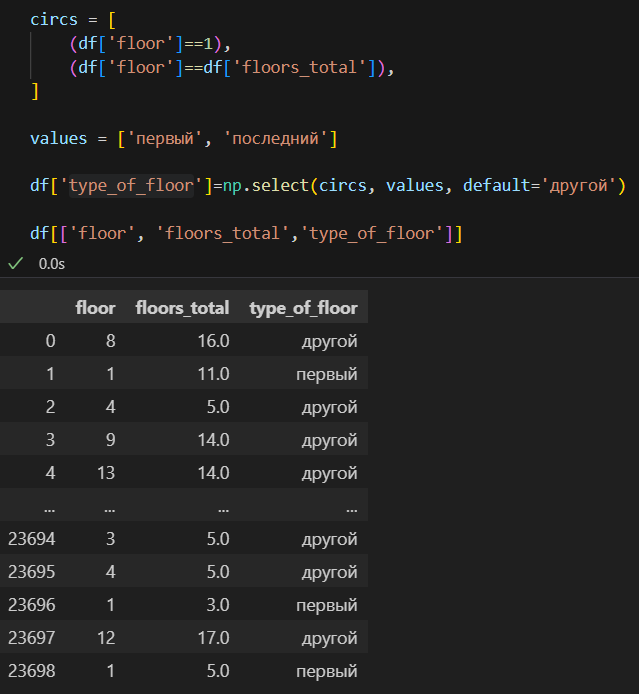
* цену квадратного метра;



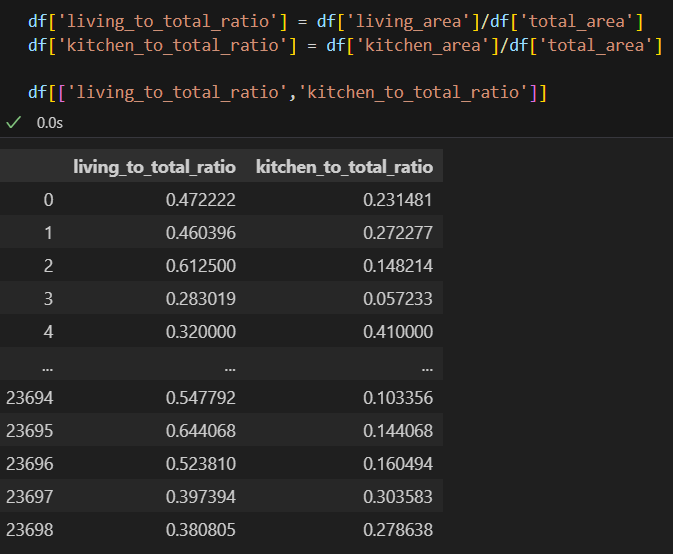
* день недели, месяц и год публикации объявления;



* этаж квартиры; варианты — первый, последний, другой;

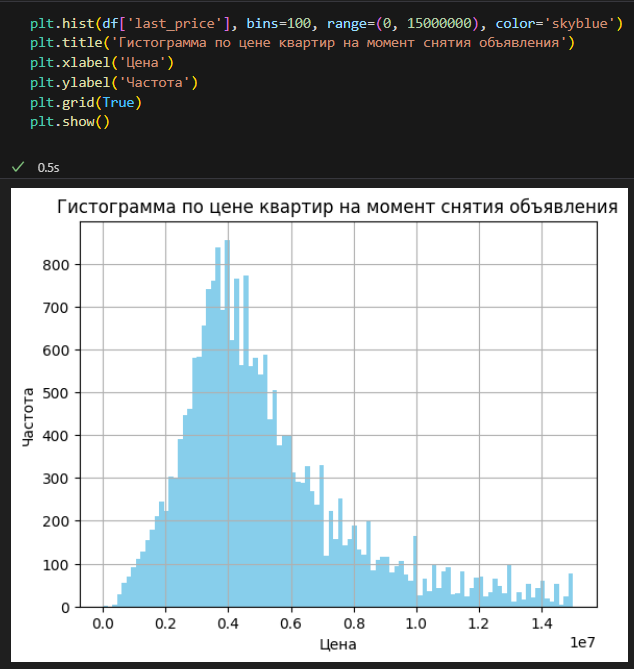
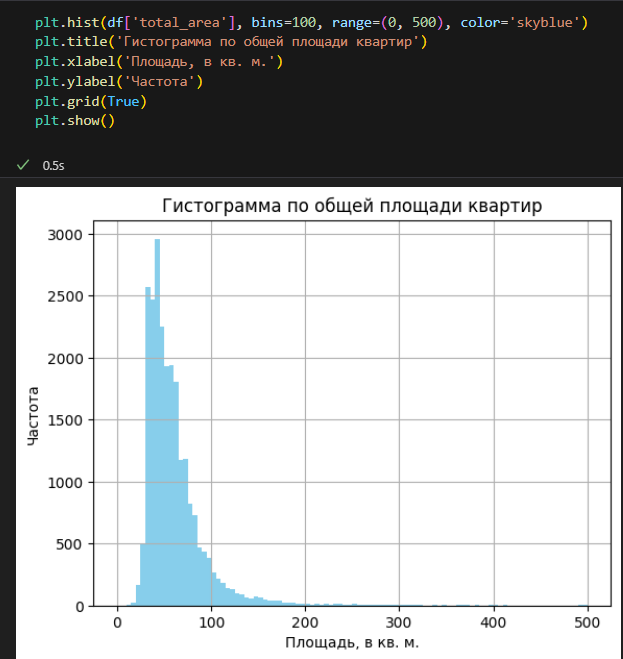


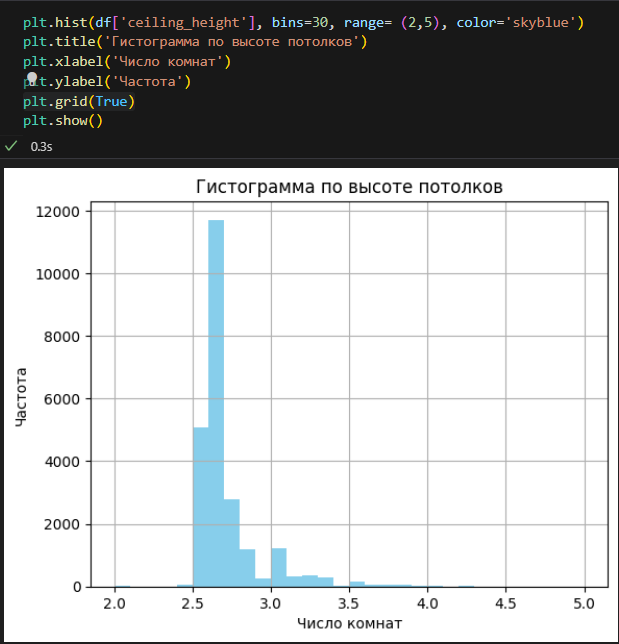
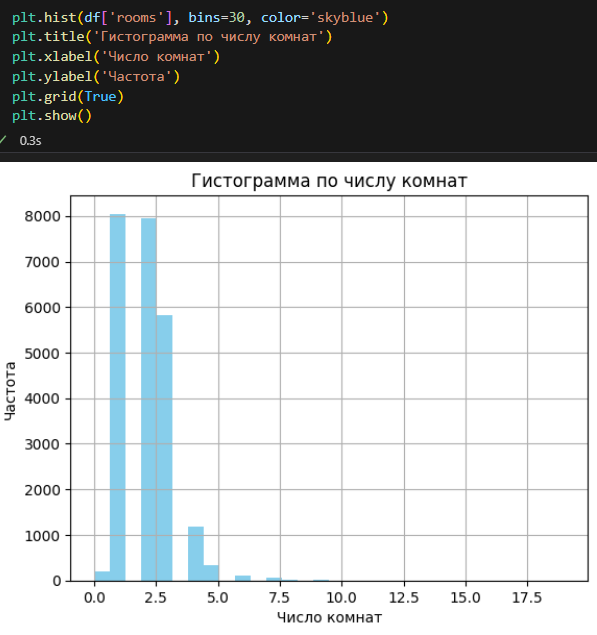
* соотношение жилой и общей площади, а также отношение площади кухни к общей.



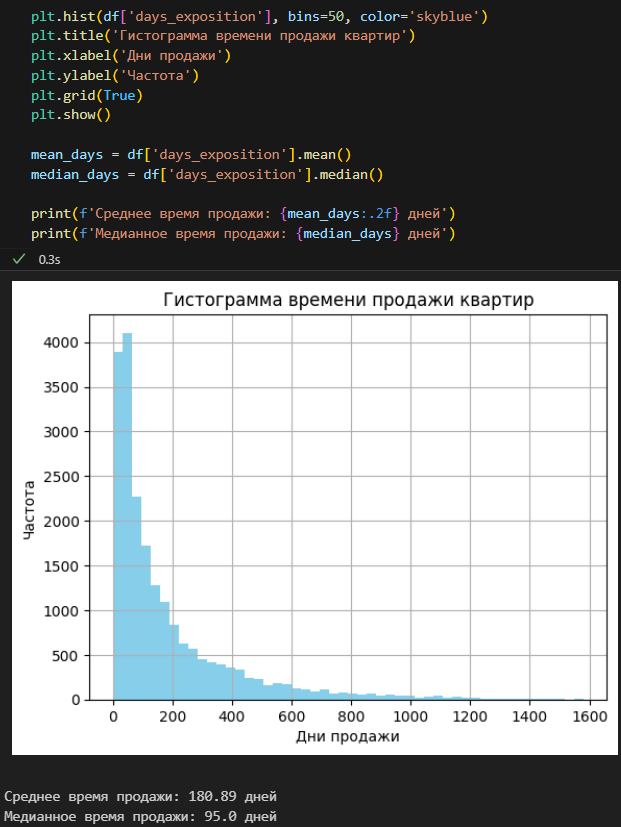
# Шаг 4. Проведите исследовательский анализ данных и выполните инструкции

* Изучите следующие параметры: площадь, цена, число комнат, высота потолков. Постройте гистограммы для каждого параметра.





* Изучите время продажи квартиры. Постройте гистограмму. Посчитайте среднее и медиану. Опишите, сколько обычно занимает продажа.



Судя по гистограмме, среднее время продажи квартир составляет 180 дней, в то время как медианное значение равно 95 дням, что указывает на то, что большинство квартир продается относительно быстро, но существуют выбросы с длительным временем продажи.

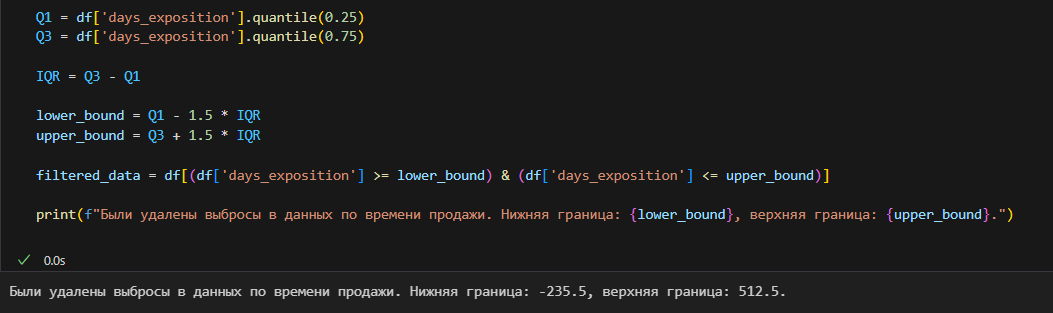
* Уберите редкие и выбивающиеся значения. Опишите, какие особенности обнаружили.

Для удаления выбросов в данных по времени продажи будем использовать алгоритм на основе интерквартильного размаха (IQR).

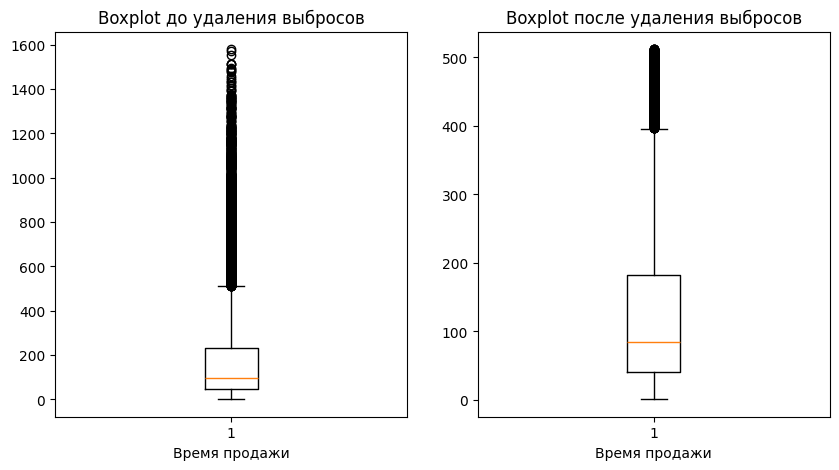
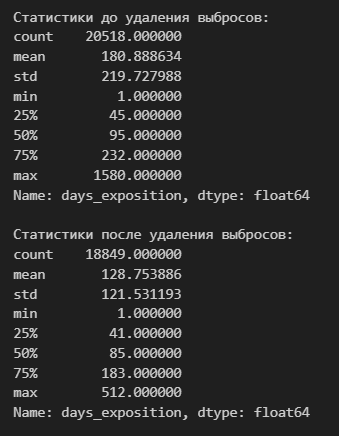
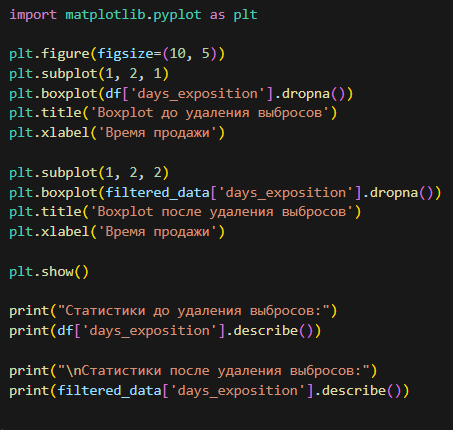
- Верхний квартиль – это такая оценка, выше которой только 25% оценок.

- Нижний квартиль – это такое значение, ниже которого только 25% оценок.

- Межквартильный размах (IQR) – это разница между 75% и 25% квартилем. Внутри этого диапазона лежит 50% наблюдений.



Теперь сравним время продажи до и после удаления выбросов.



После удаления выбросов в данных о времени продажи квартир, наблюдаются следующие изменения в статистиках:

До удаления выбросов:

- Среднее время продажи составляло 180.89 дней.

- Медианное время продажи было равно 95 дням.

- 75% квартир продавались в течение 232 дней, что является верхним квартилем.

После удаления выбросов:

- Среднее время продажи уменьшилось и составило 128.75 дней.

- Медианное время продажи также уменьшилось и стало равным 85 дням.

- 75% квартир продается в течение 183 дней.

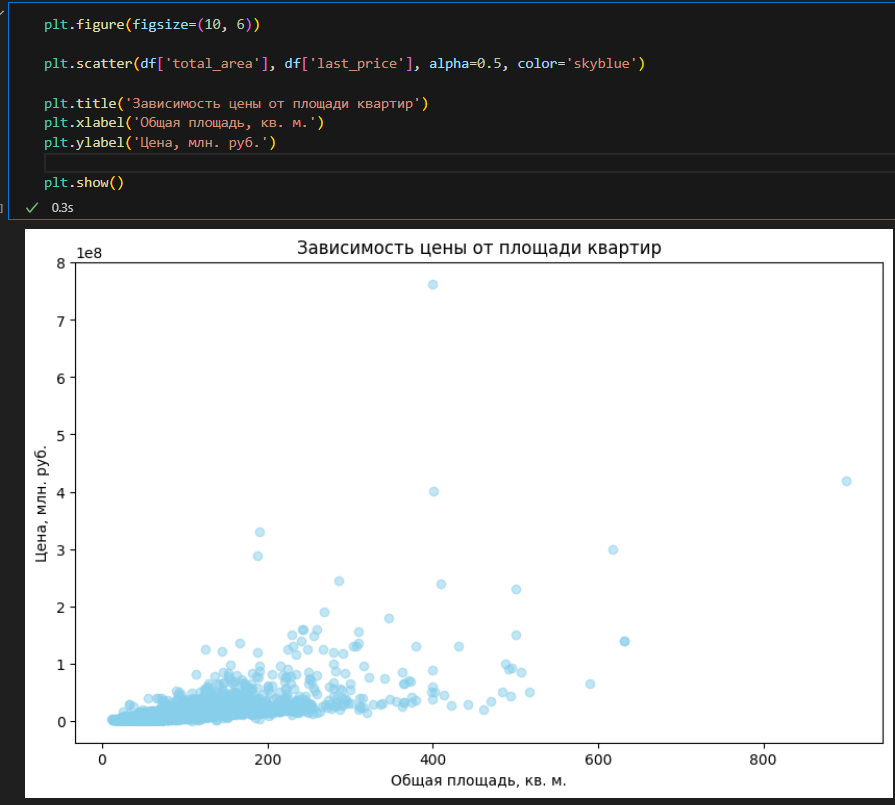
Выводы:

- Удаление выбросов сократило среднее и медианное время продажи, что указывает на то, что выбросы в виде очень долгих продаж влияли на среднее значение.

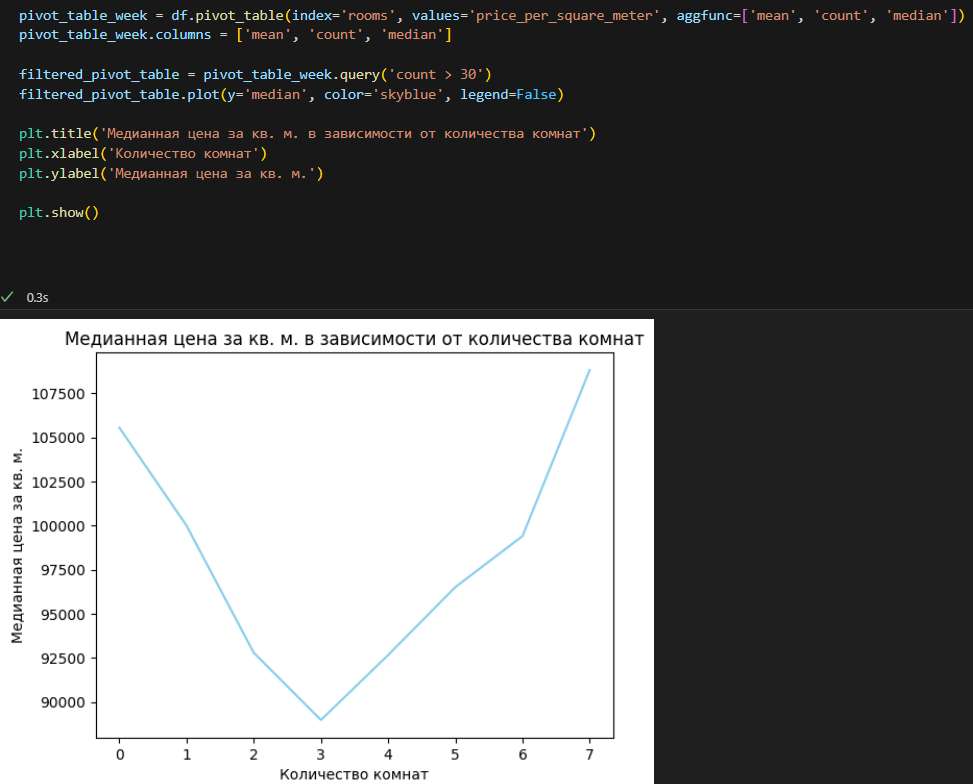
- Медианное значение осталось близким к среднему, что может свидетельствовать о том, что распределение времени продажи стало более симметричным.

- Уменьшение значения 75% квартиля после удаления выбросов указывает на сокращение времени продажи для большинства квартир.

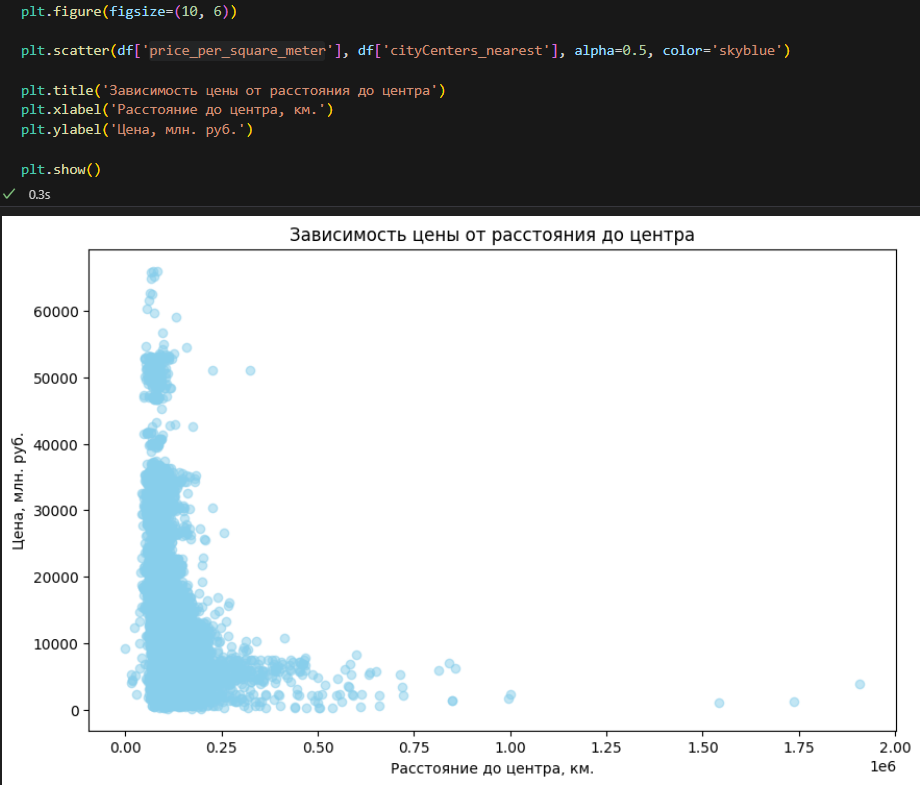
* Какие факторы больше всего влияют на стоимость квартиры? Изучите, зависит ли цена от площади, числа комнат, удалённости от центра. Изучите зависимость цены от того, на каком этаже расположена квартира: первом, последнем или другом. Также изучите зависимость от даты размещения: дня недели, месяца и года. Постройте графики.



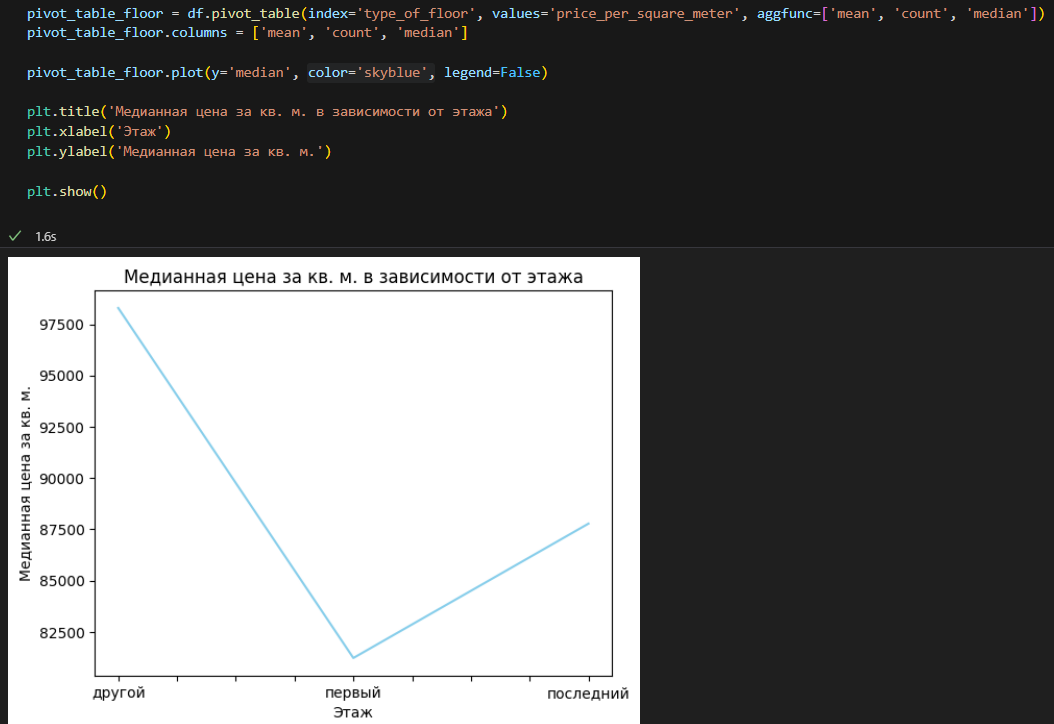
Из графика можно сделать вывод, что зависимость цены от общей площади слабая, но есть - с увеличением площади растет и стоимость.



Из графика видно, что чем ближе количество комнат к трем, тем ниже цена кв. м.



Из графика видна явная зависимость цены кв. м. от расстояния до центра - чем ближе к центру, тем выше цена.



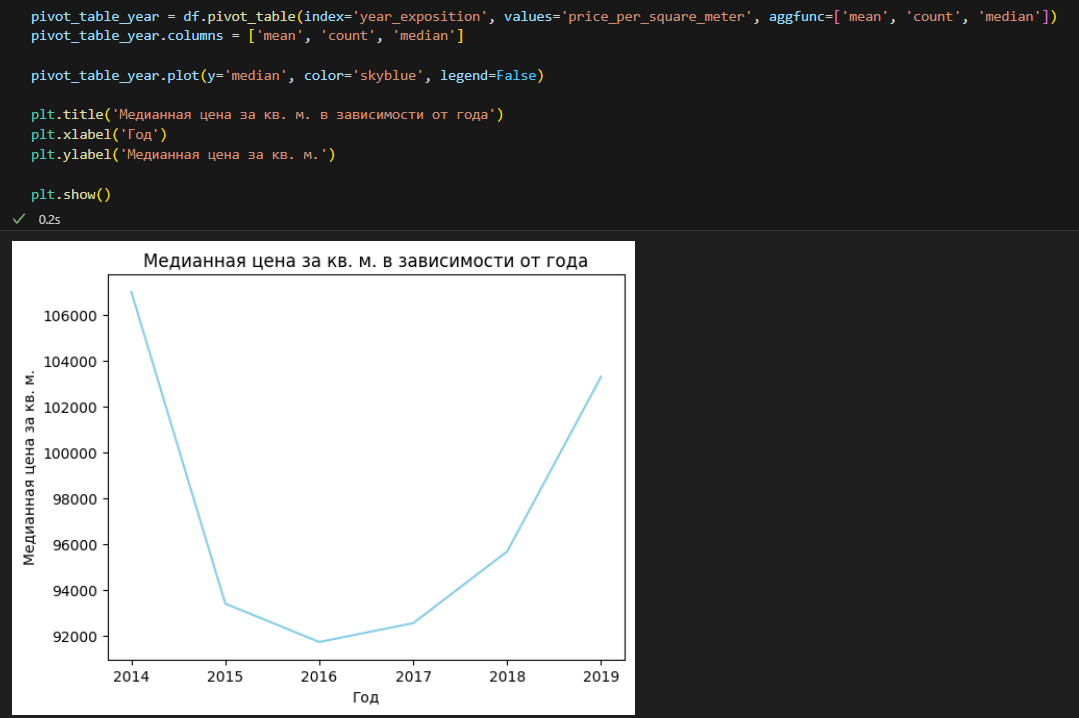
Из графика видно, что цена квартиры на первом этаже в среднем ниже, чем на последнем и другом этаже. Большую же стоимость имеют квартиры, расположенные на этажах, отличных от первого и последнего.



Из графика можно сделать вывод, что цены на квартиры, объявления которых были выложены во вторник гораздо ниже, чем в другие дни недели.

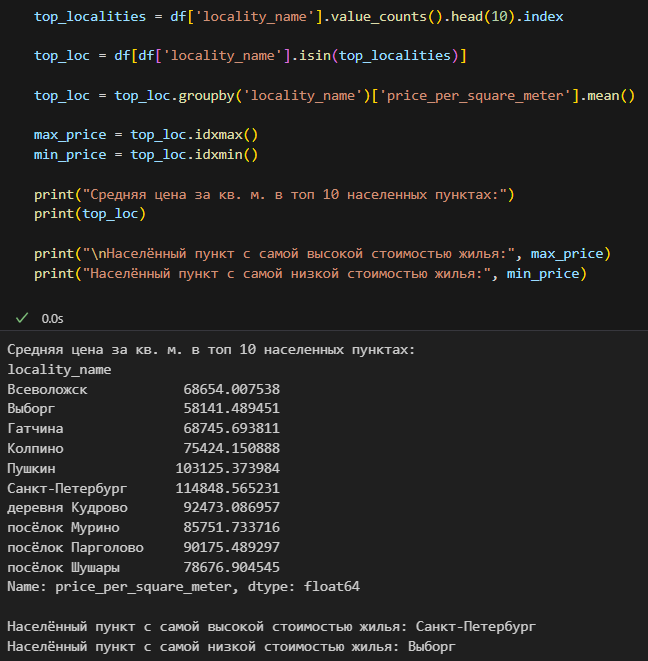


Из графика можно сделать вывод, что самая низкая стоимость у квартир, объявления которых были выложены в мае и июне, а сама высокая - в апреле и августе.



По графику видно, что стоимость квартир падала в 2014-2016 годах, но после стала резко возрастать.

* Выберите 10 населённых пунктов с наибольшим числом объявлений. Посчитайте среднюю цену квадратного метра в этих населённых пунктах. Выделите населённые пункты с самой высокой и низкой стоимостью жилья. Эти данные можно найти по имени в столбце locality\_name.
* Населённый пункт с самой высокой стоимостью жилья: Санкт-Петербург

Таким образом, мы нашли топ 10 населенных пунктов с самым большим количеством объявлений и населённые пункты с самой низкой стоимостью жилья и самой большой стоимостью.

# Ссылка на GitHub:

https://github.com/setusq/PMOZDlabN6/tree/main/lr6

**Вывод:** в ходе выполнения лабораторной работы с использованием библиотек Pandas и Matplotlib был проведен анализ данных о продаже квартир в Санкт-Петербурге и соседних населенных пунктах. Процесс предобработки данных включал заполнение пропущенных значений и приведение данных к необходимым типам, что обеспечило точность и надежность последующего анализа.