|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К*** ***НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Анализ объявлений о продаже квартир***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_******\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ5-52Б\_\_\_\_\_\_\_ \_\_**\_\_\_\_\_\_\_\_ \_\_\_Пустовалов Г.В.\_\_\_\_**

(Группа) (Подпись, дата). (И.О.Фамилия)

Руководитель научно-исследовательской работы **\_\_\_\_\_\_\_\_\_\_ \_\_\_****К. Ю. Маслеников\_\_\_**

(Подпись, дата). (И.О.Фамилия)

*2023 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**ЗАДАНИЕ**

**на выполнение** **научно-исследовательской работы**

по дисциплине \_\_\_Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_ \_ \_ \_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_ИУ5-52Б\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Пустовалов Григорий Владимирович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Тема научно-исследовательской работы Проектирования операционного устройства и его

       составляющих – операционного автомата и управляющего устройства \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Направленность КР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_УЧЕБНАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_КАФЕДРА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения работы: 25% к \_\_ нед., 50% к \_\_ нед., 75% к \_\_ нед., 100% к \_\_ нед.

***Задание*** \_ Спроектировать операционное устройство и его

     составляющие – операционный автомат и управляющее устройство

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**Руководитель** **научно-исследовательской работы** \_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_К. Ю.Маслеников\_\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_** Г.В. Пустовалов **\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

# Введение

Необходимо исследовать базу данных и выявить закономерности в стоимости квартир.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

# Определение данных для анализа

# В качестве данных для анализа был выбран датасет «объявлений о продаже квартир»

Сейчас покупка, продажа и аренда квартиры стали тяжелым процессом для каждого человека. Особенно это касается крупного города, где особенно сильно варьируются варианты: как по ценам, так и по инфраструктуре.

Учитывая это, было сделано решение выбрать этот датасет. В датасете представлены данные о квартирах в Санкт-Петербурге и Ленинградской области: их стоимости, площади, близости парков и множестве других параметров.

1. **Описание данных**

Для анализа были собраны данные о продаваемых в 2018 году в Москве квартирах. В наборе данных содержатся:

1. **total\_images** – количество изображний в объявлении
2. **last\_price** – последняя цена, указанная в объявлении
3. **total\_area**– общая площадь квартиры, кв.м.
4. **first\_day\_exposition** – день, в который объявление было выложено.
5. **rooms**– количество комнат
6. **сeiling\_height**– высота потолков
7. **floors\_total** – количество этажей в доме
8. **living\_area**– жилая площадь квартиры, кв.м.
9. **floor –** этаж
10. **is\_apartament** **–** является ли апартаментами
11. **kitchen\_area –** кухня площадь, кв.м.
12. **locality\_name –** название населенного пункта
13. **cityCenter\_nearest –** расстояние до центра
14. **parks\_nearest –** ближайший парк
15. **days\_exposition –** количество дней объявлению

# Формулирование гипотез

В ходе первичного анализа были выдвинуты следующие гипотезы:

**Гипотеза 1:** чем больше комнат в квартире, тем она дороже.

**Гипотеза 2**: стоимость квадратного метра увеличивается каждый год

**Гипотеза 3:** чем ближе квартира к центру города, тем выше ее стоимость квадратного метра.

**Гипотеза 4:** если квартира находится в радиусе 9км от центра, то на стоимость квартиры не влияет расстояние до центра.

# Изучение общей информации

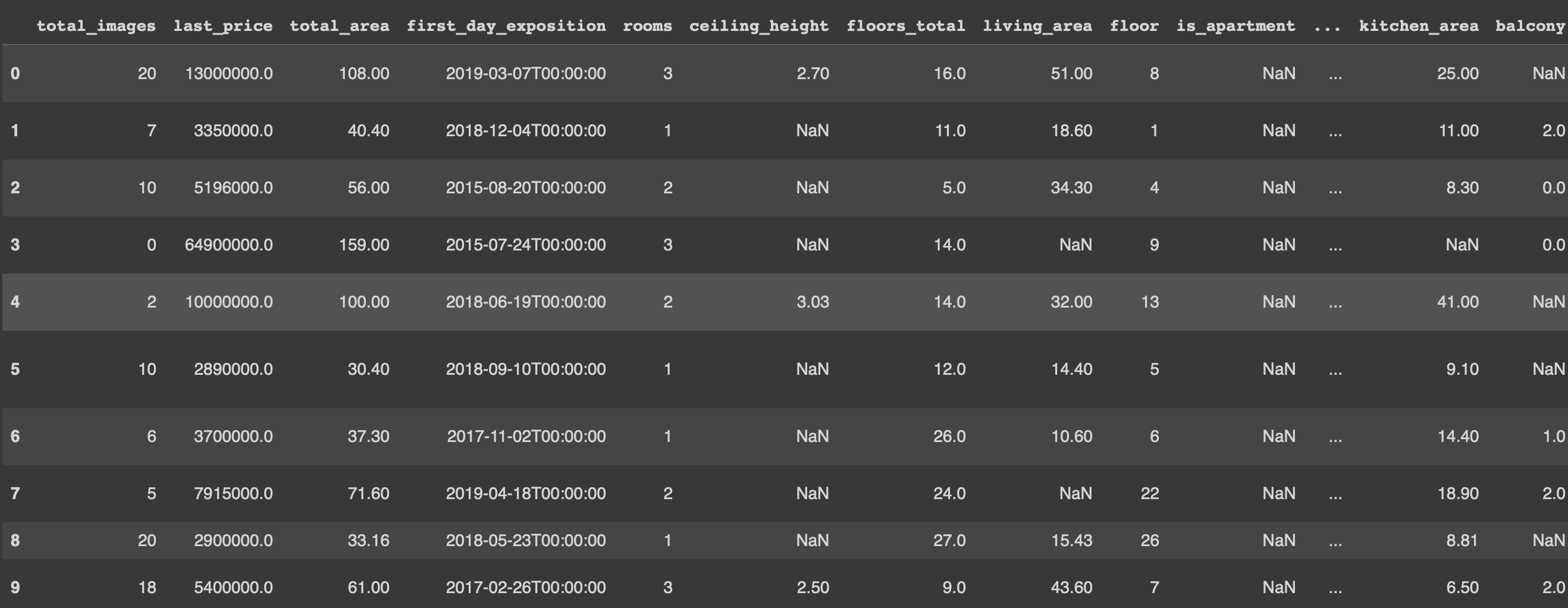
Загружаем датасет, подключаем необходимые библиотеки:

import pandas as pd

Откроем файл с данными.

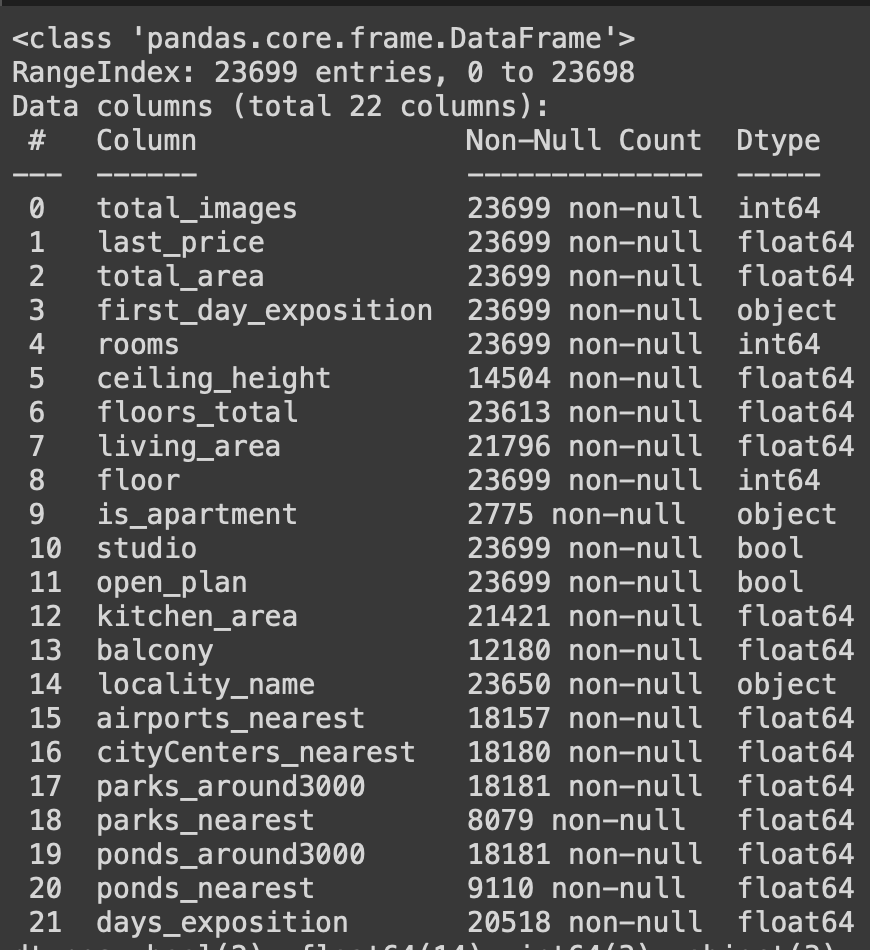
df = pd.read\_csv('/real\_estate\_data.csv', sep='\t')

df.head(10)



Получим информацию о датасете:

df.info()



Проверим на дубликаты.

df.duplicated().sum()



Займемся пропусками. Пропущенные значения количества балконов заменим на 0, так как, если человек не указал число балконов — скорее всего, их нет.

df['balcony'] = df['balcony'].fillna(0)

df['balcony'].isna().sum()

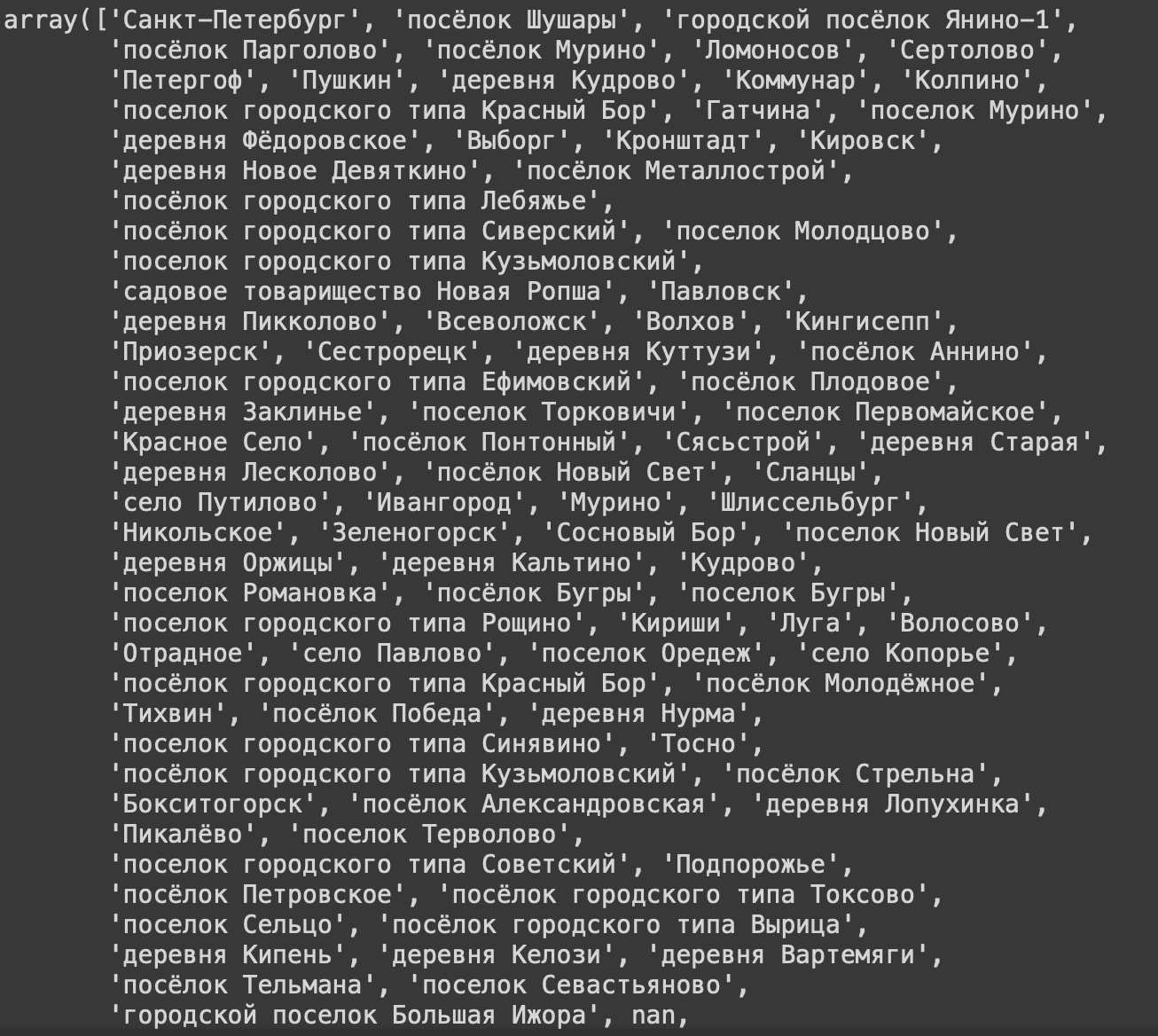
Аналогично поступим с количеством парков и прудов в радиусе 3 000 м. Если поле осталось пустым, скорее всего их просто нет.

df['parks\_around\_3000'] = df['parks\_around\_3000'].fillna(0)

df['ponds\_around\_3000'] = df['ponds\_around\_3000'].fillna(0)

Посмотрим, какие населенные пукты представленны.

df['locality\_name'].unique()



Слово "Поселок" записано по-разному, через буквы "е" и "ё". Заменим все буквы "ё" на "е".

len(df['locality\_name'].unique())

365

df['locality\_name'] = df['locality\_name'].str.replace('ё', 'е')

len(df['locality\_name'].unique())

331

Сделаем формат столбцов "balcony" и "last\_price" целочисленным, так удобнее воспринимать информацию из них. Есть ещё столбцы, которые можно было бы сделать целочисленным, но они имеют пропуски, которые нельзя заполнить логически.

df['balcony'] = df['balcony'].astype('int')

df['last\_price'] = df['last\_price'].astype('int')

df.head()

Добавим в таблицу стобцы со значением цены за квадратный метр "price\_per\_m2" и годом "year" размещения объявления.

df['price\_per\_m2'] = df['last\_price'] / df['total\_area']

df['year'] = df['first\_day\_exposition'].dt.year

df.head()

# Исследовательский анализ данных

## **Сравнение медианных цен с разными комнатами**

df.groupby('rooms')['last\_price'].median()\

.plot(kind = 'bar',\

figsize = (15,5),\

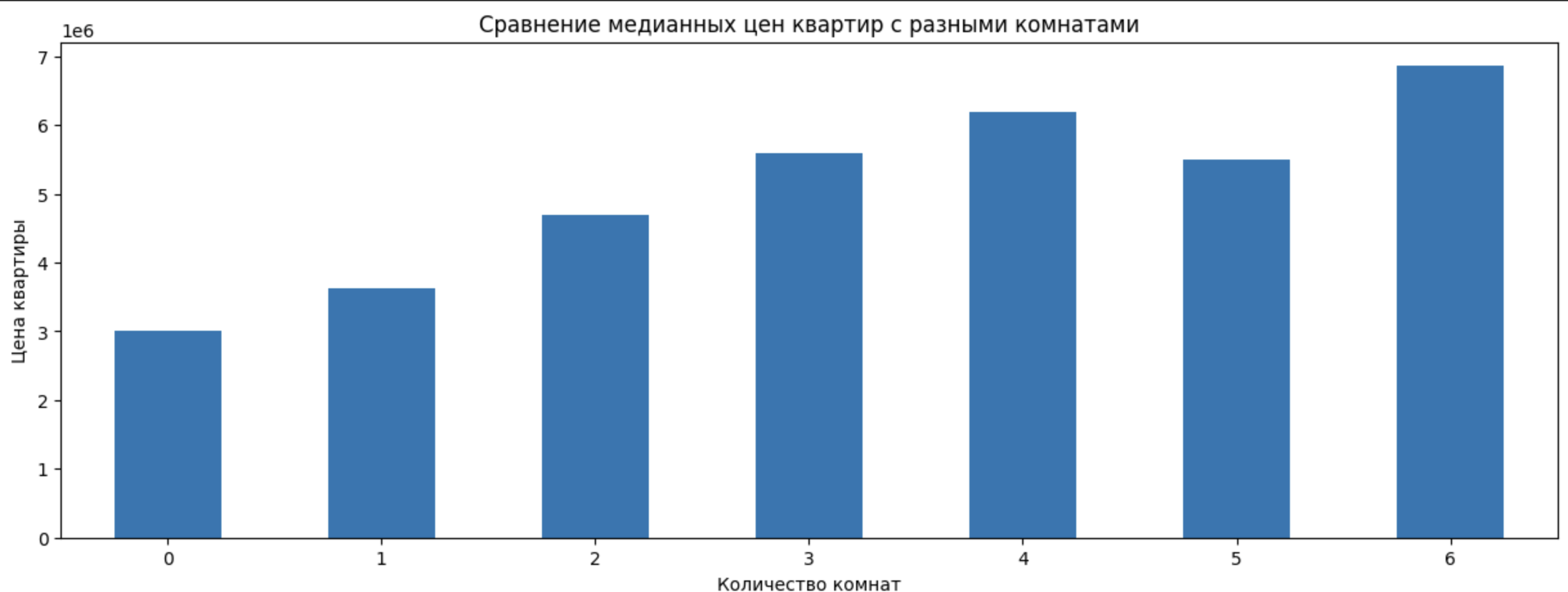
title = 'Сравнение медианных цен квартир с разными комнатами')

plt.xticks(rotation = 0)

plt.xlabel('Количество комнат')

plt.ylabel('Цена квартиры')

plt.show()



**Гипотеза 1:** соответствует действительности

## **Зависимость стоимости квадратного метра от года**



**Гипотеза 2**: возможно, в 2015 году был кризис недвижимости

## **Зависимость стоимости квадратного метра от расстояния до центра**



Гипотеза 3: на цену квартиры также влияет близость метро и прочих благ жизни.

Проанализируем сегмент квартир в центре.

df\_center = df.query('city\_centers\_nearest\_km < 9')

df\_center.head()

df\_center.plot(x='price\_per\_m2',

y='city\_centers\_nearest',

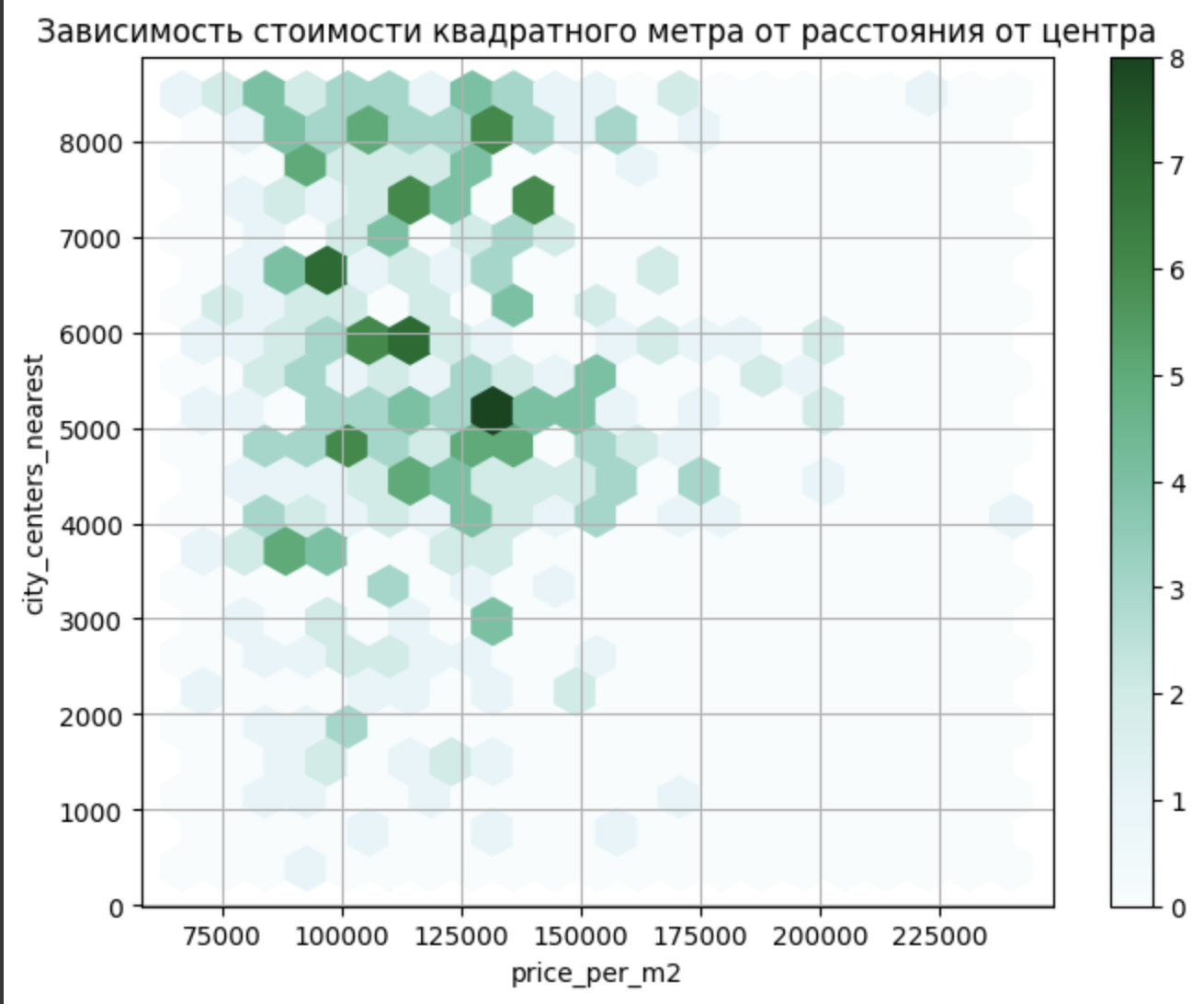
title='Зависимость стоимости квадратного метра от расстояния от центра',

kind='hexbin',

gridsize=20,

figsize=(8, 6),

sharex=False, grid=True)



**Гипотеза 4:** соответствует действительности