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 обработки и защиты данных»

Выполнил:

Студенты гр. М3О-236Б-22

Попов М.А

Кузьмин Д.Е

Преподаватель:

Александрова С.С.

Москва 2023

Оглавление

[Цель работы: 3](#_Toc150595144)

[Раздел 1. Изучите общую информацию. 4](#_Toc150595145)

[Раздел 2. Предобработка данных. 5](#_Toc150595146)

[Раздел 3. Добавление в таблицу данных. 8](#_Toc150595147)

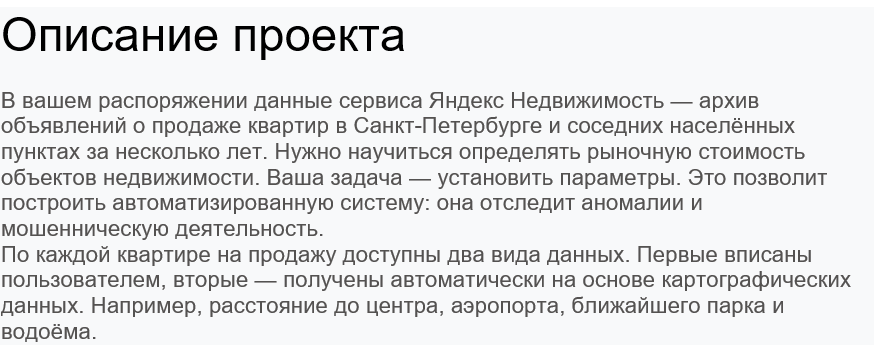
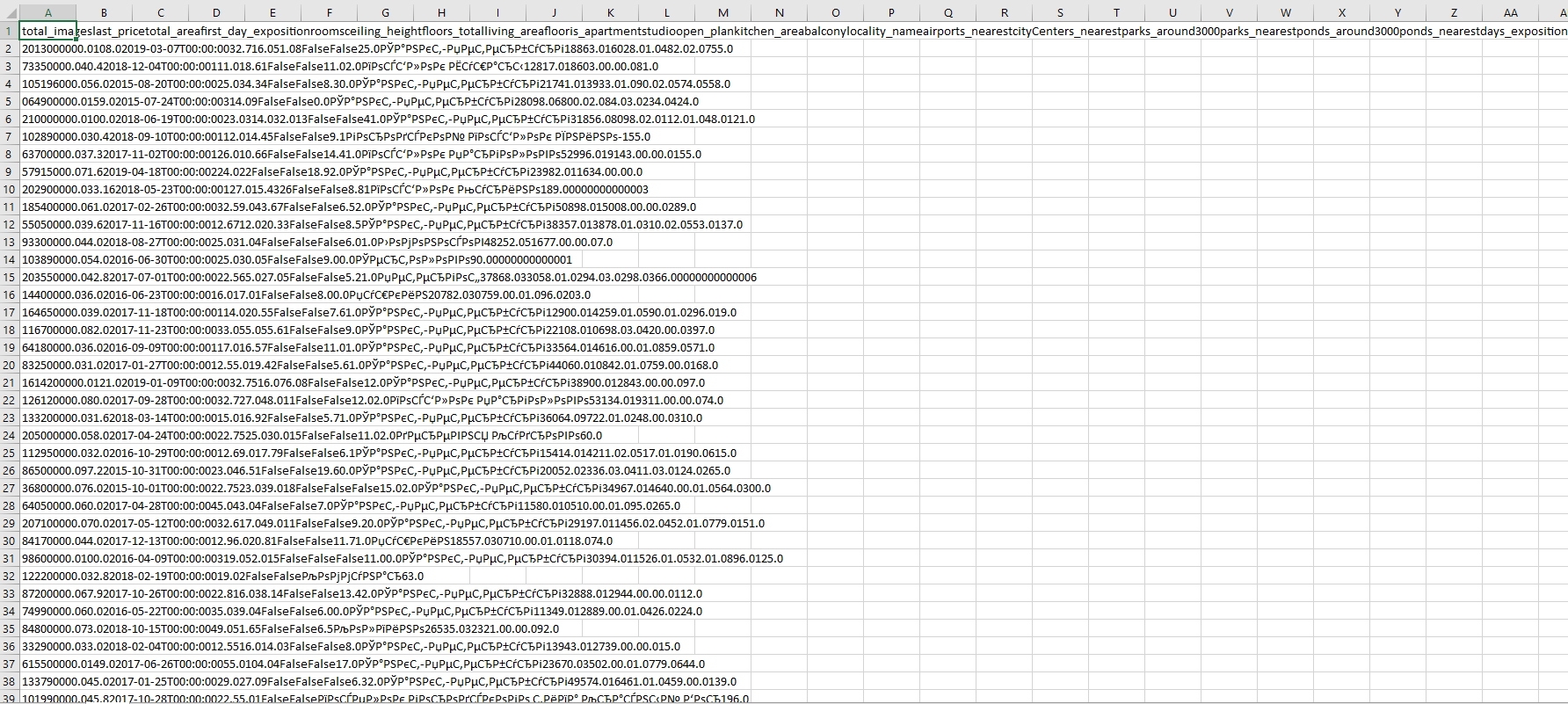
[Раздел 4. Проведение исследовательского анализа. 11](#_Toc150595148)

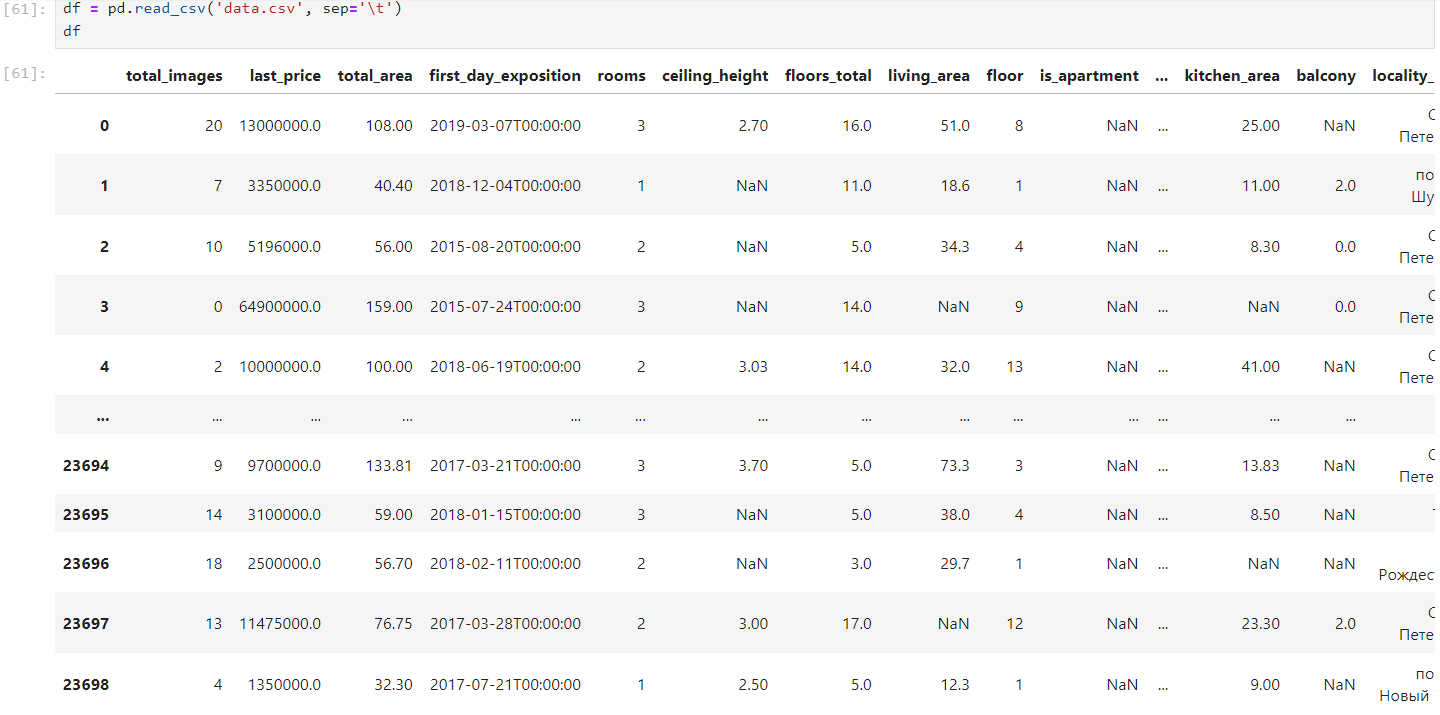
[Раздел 5. Вывод. 16](#_Toc150595149)

# Цель работы:

Научиться определять рыночную стоимость объектов недвижимости для дальнейшей работы с ними и построения автоматизированной системы расчётов.

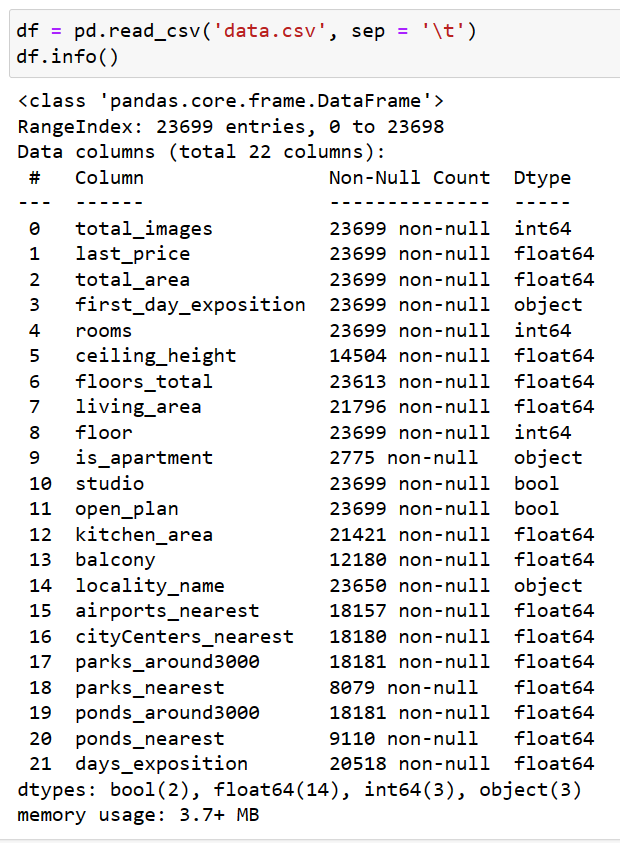
# Раздел 1. Изучите общую информацию.



Нами были изменены данные в изначальной таблице. Все названия столбцов заменены на англоязычные аналоги, а также название файла на data.csv.

# Раздел 2. Предобработка данных. Нами были изменены типы данных нескольких столбцов, а также заполнены пропуски в некоторых столбцах, где это требовалось, в остальных случаях заменили пропущенные значения медианными значениями столбцов до внесения изменений в них.



#ищем медианные значения для столбцов

median\_parks\_around3000 = df['parks\_around3000'].median()

median\_parks\_nearest = df['parks\_nearest'].median()

median\_ponds\_around3000 = df['ponds\_around3000'].median()

median\_ponds\_nearest = df['ponds\_nearest'].median()

#заполняем значения пустых ячеек (NA) медианными значениями или же (в balcony нулями )

df['ponds\_nearest'] = df['ponds\_nearest'].fillna(median\_ponds\_nearest)

df['ponds\_around3000'] = df['ponds\_nearest'].fillna(median\_ponds\_around3000)

df['parks\_nearest'] = df['ponds\_nearest'].fillna(median\_parks\_nearest)

df['parks\_around3000'] = df['ponds\_nearest'].fillna(median\_parks\_around3000)

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

df['is\_apartment'] = df['is\_apartment'].fillna(False)

#изменяем тип данных на дату для столбца first\_day\_exposition

df['first\_day\_exposition'] = pd.to\_datetime(df['first\_day\_exposition'])

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

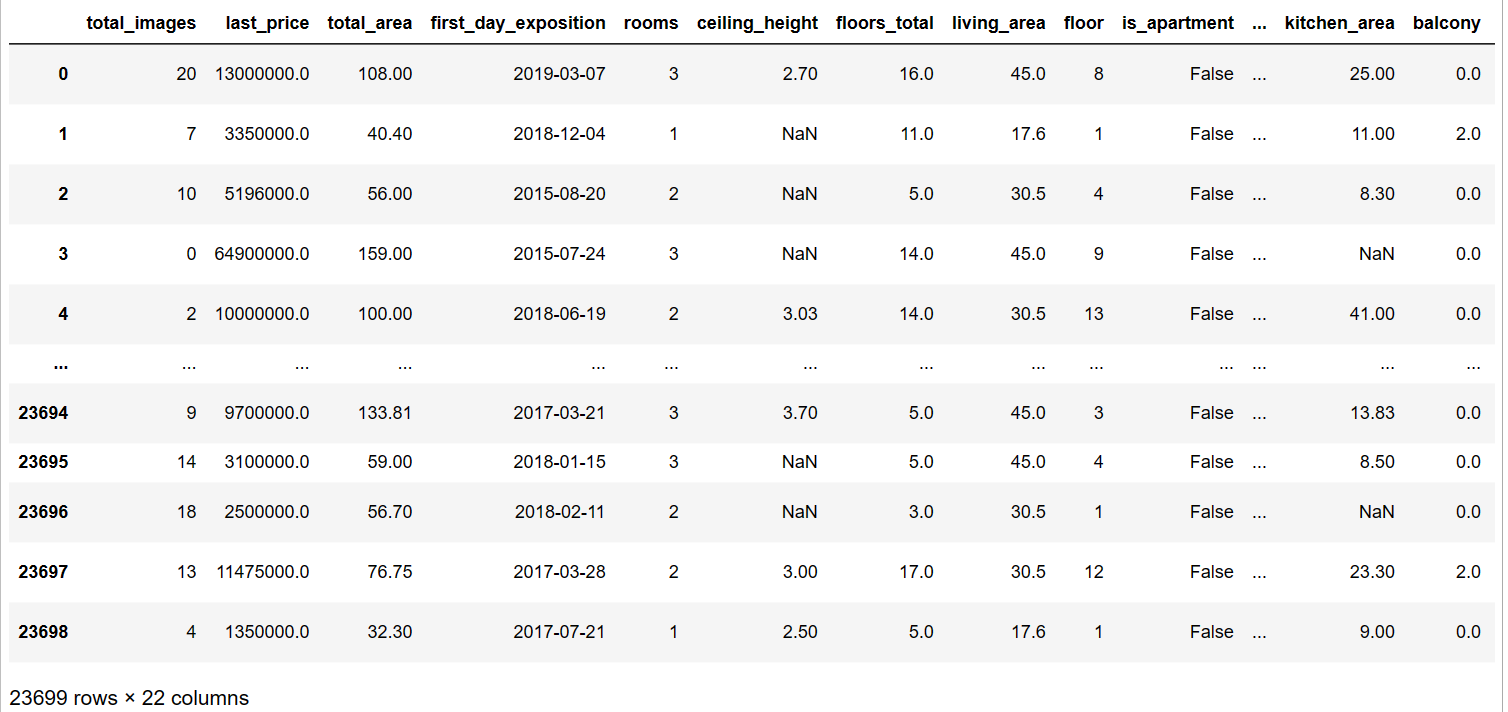
**for** i **in** range(**4**):

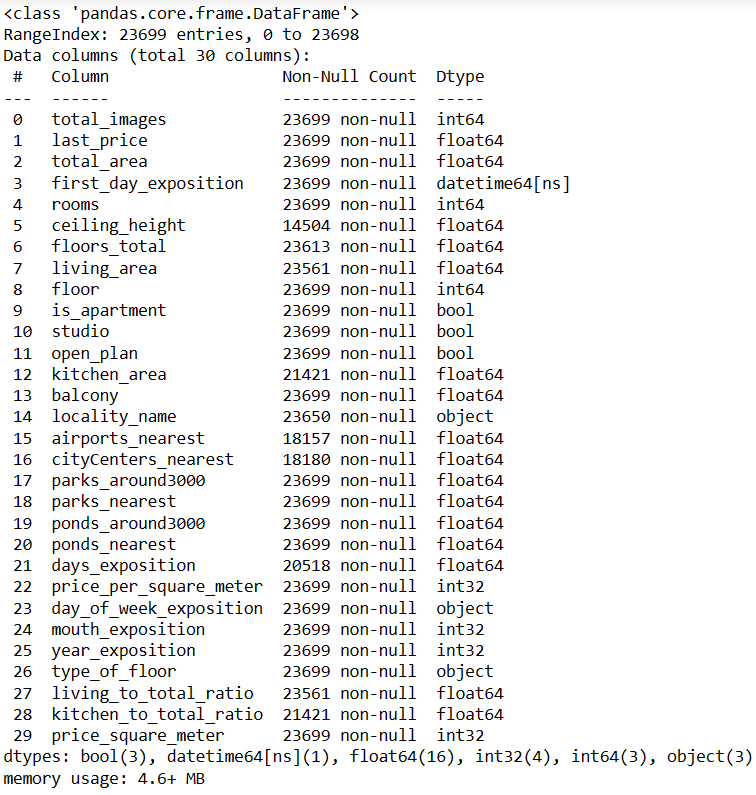
median\_value = df.loc[df['rooms'] == i, 'living\_area'].median()

df.loc[df['rooms'] == i, 'living\_area'] = median\_value

df.info()

df





# Раздел 3. Добавление в таблицу данных.

Были созданы и заполнены дополнительные столбцы с информацией, нужные для анализа рынка продаж.

**День недели, месяц и год публикации объявления**

price\_square\_meter – цена квадратного метра;

day\_of\_week\_exposition – день недели публикации объявления;

mouth\_exposition – месяц публикации объявления;

year\_exposition - год публикации объявления;

# цена квадратного метра

# считаем значение

df['price\_square\_meter'] = df['last\_price'] / df['total\_area']

# заполняем столбец и вписываем данные с типом данных int

df['price\_square\_meter']= df['price\_square\_meter'].astype('int')

df['price\_square\_meter']

#день недели, месяц и год публикации объявления

df['first\_day\_exposition'] = pd.to\_datetime(df['first\_day\_exposition'])

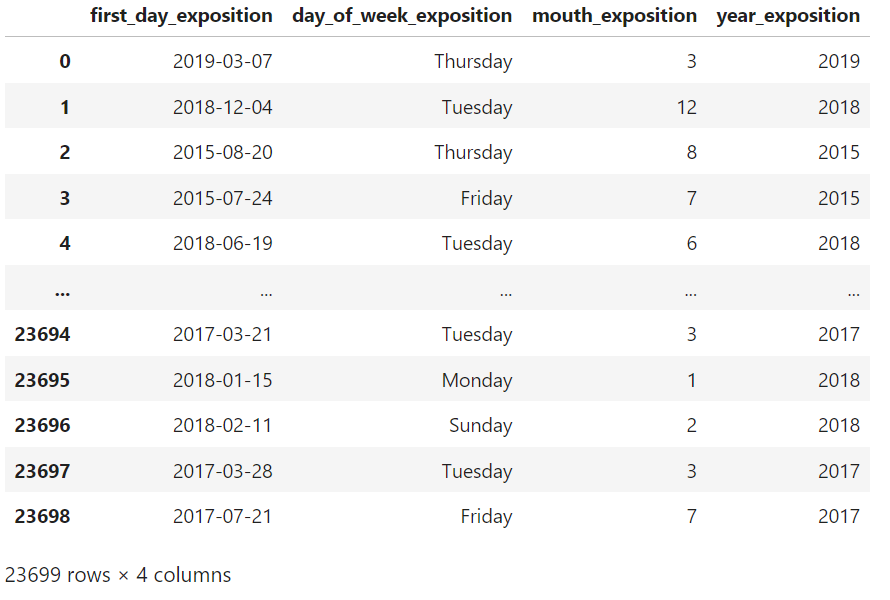
df['day\_of\_week\_exposition'] = df['first\_day\_exposition'].dt.day\_name()

df['mouth\_exposition']=df['first\_day\_exposition'].dt.month

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

# заполняем столбец

df[[ 'first\_day\_exposition', 'day\_of\_week\_exposition', 'mouth\_exposition', 'year\_exposition']]



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

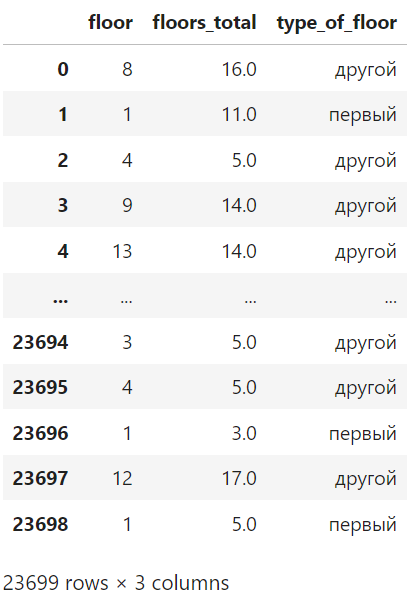
#этаж квартиры; варианты — первый, последний, другой

circs = [(df['floor']==**1**), (df['floor']==df['floors\_total']),]

values = ['первый', 'последний']

df['type\_of\_floor']=np.select(circs, values, default='другой')

df[['floor', 'floors\_total','type\_of\_floor']]

****

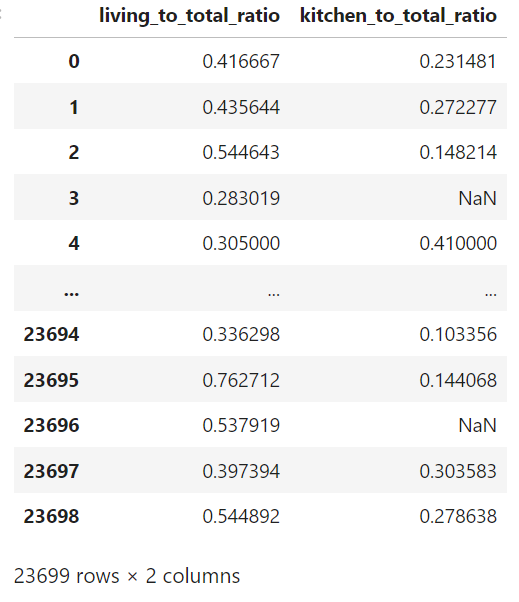
**Соотношение жилой и общей площади, а также отношение площади кухни к общей**#соотношение жилой и общей площади, а также отношение площади кухни к общей

df['living\_to\_total\_ratio'] = df['living\_area']/df['total\_area']

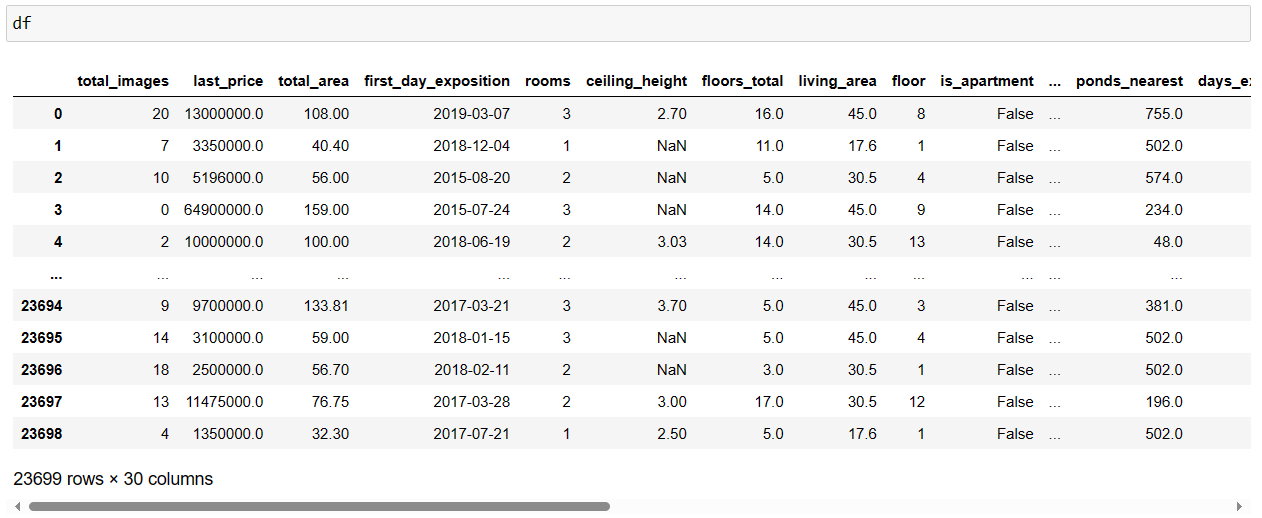
df['kitchen\_to\_total\_ratio'] = df['kitchen\_area']/df['total\_area']

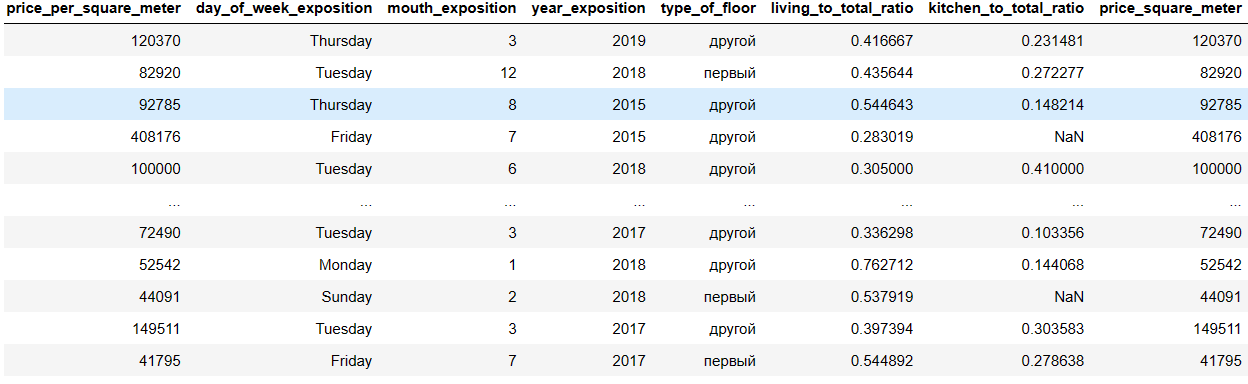
# заполняем столбец

df[['living\_to\_total\_ratio','kitchen\_to\_total\_ratio']]



Рассмотрим наш датасет





# Раздел 4. Проведение исследовательского анализа.

**Изучение следующих параметров: площадь, цена, число комнат, высота потолков. Построим гистограммы для каждого параметра**

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

plt.hist(df['total\_area'], bins=**100**, range=(**0**, **500**), color='orange')

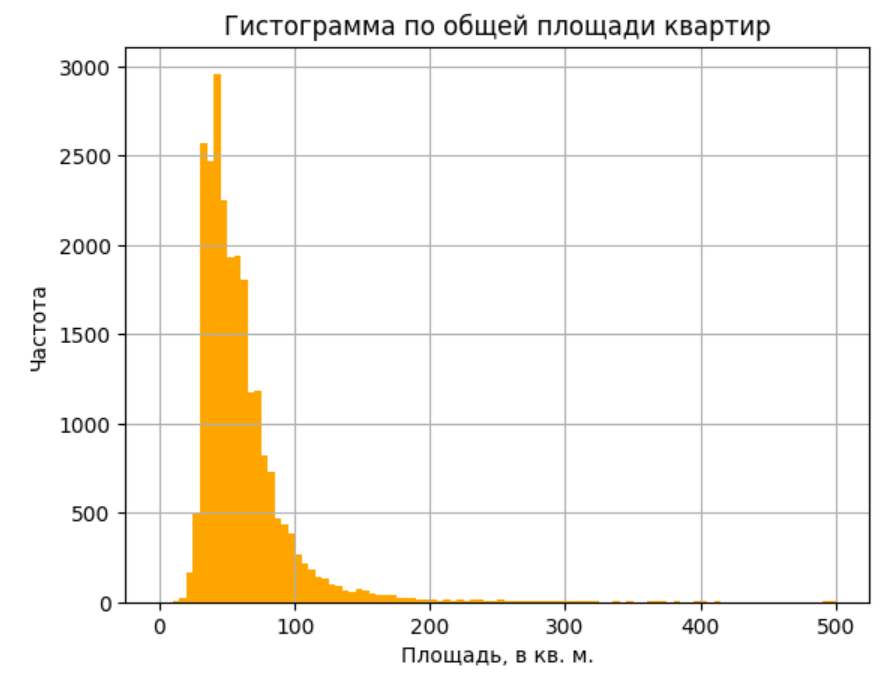
plt.title('Гистограмма по общей площади квартир')

plt.xlabel('Площадь, в кв. м.')

plt.ylabel('Частота')

plt.grid(True)

plt.show()

****

plt.hist(df['rooms'], bins=**30**, color='orange')

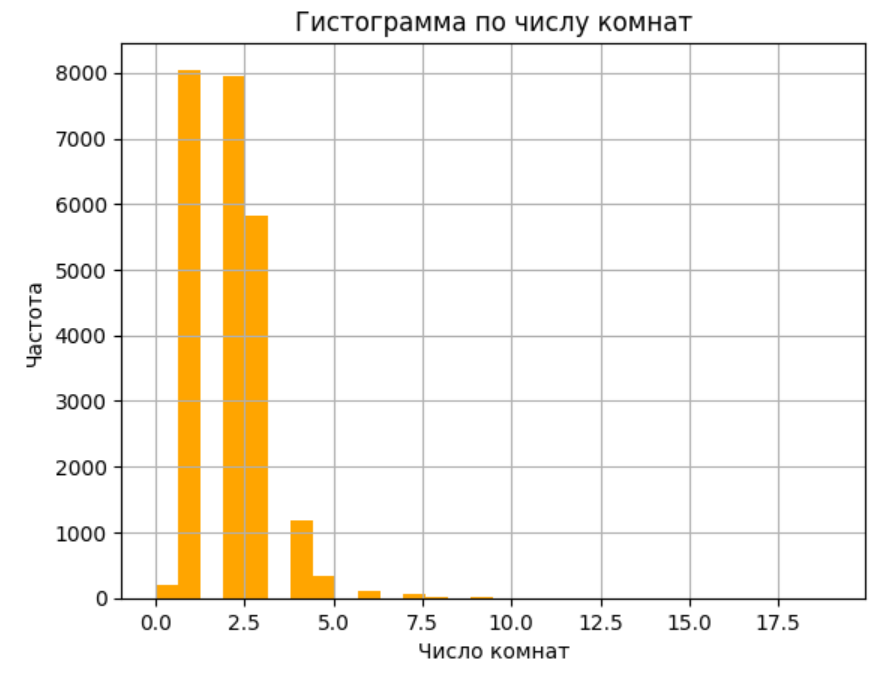
plt.title('Гистограмма по числу комнат')

plt.xlabel('Число комнат')

plt.ylabel('Частота')

plt.grid(True)

plt.show()



plt.hist(df['ceiling\_height'], bins=**30**, range= (**2**,**5**), color='orange')

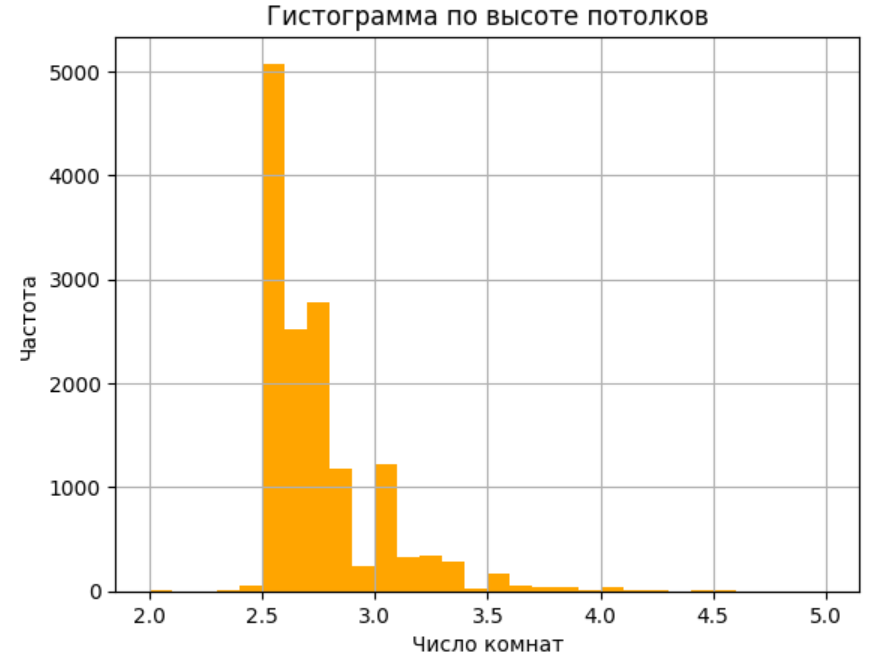
plt.title('Гистограмма по высоте потолков')

plt.xlabel('Число комнат')

plt.ylabel('Частота')

plt.grid(True)

plt.show()

****

#Изучите время продажи квартиры. Постройте гистограмму.

#Посчитайте среднее и медиану. Опишите, сколько обычно занимает продажа

plt.hist(df['days\_exposition'], bins=**50**, color='orange')

plt.title('Гистограмма времени продажи квартир')

plt.xlabel('Дни продажи')

plt.ylabel('Частота')

plt.show()

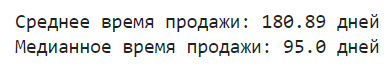
# Находим среднее времени продажи квартир и медиану

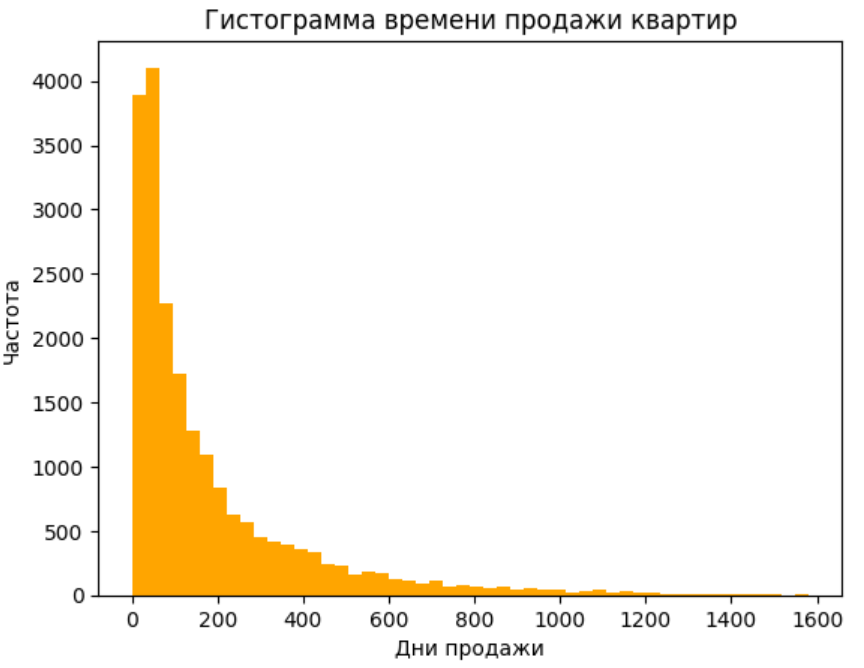
mean\_days = df['days\_exposition'].mean()

median\_days = df['days\_exposition'].median()

**print**(f'Среднее время продажи: {mean\_days:.2f} дней')

**print**(f'Медианное время продажи: {median\_days} дней')

****

****

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

plt.figure(figsize=(**10**, **6**))

plt.scatter(df['total\_area'], df['last\_price'], alpha=**0.5**, color='orange')

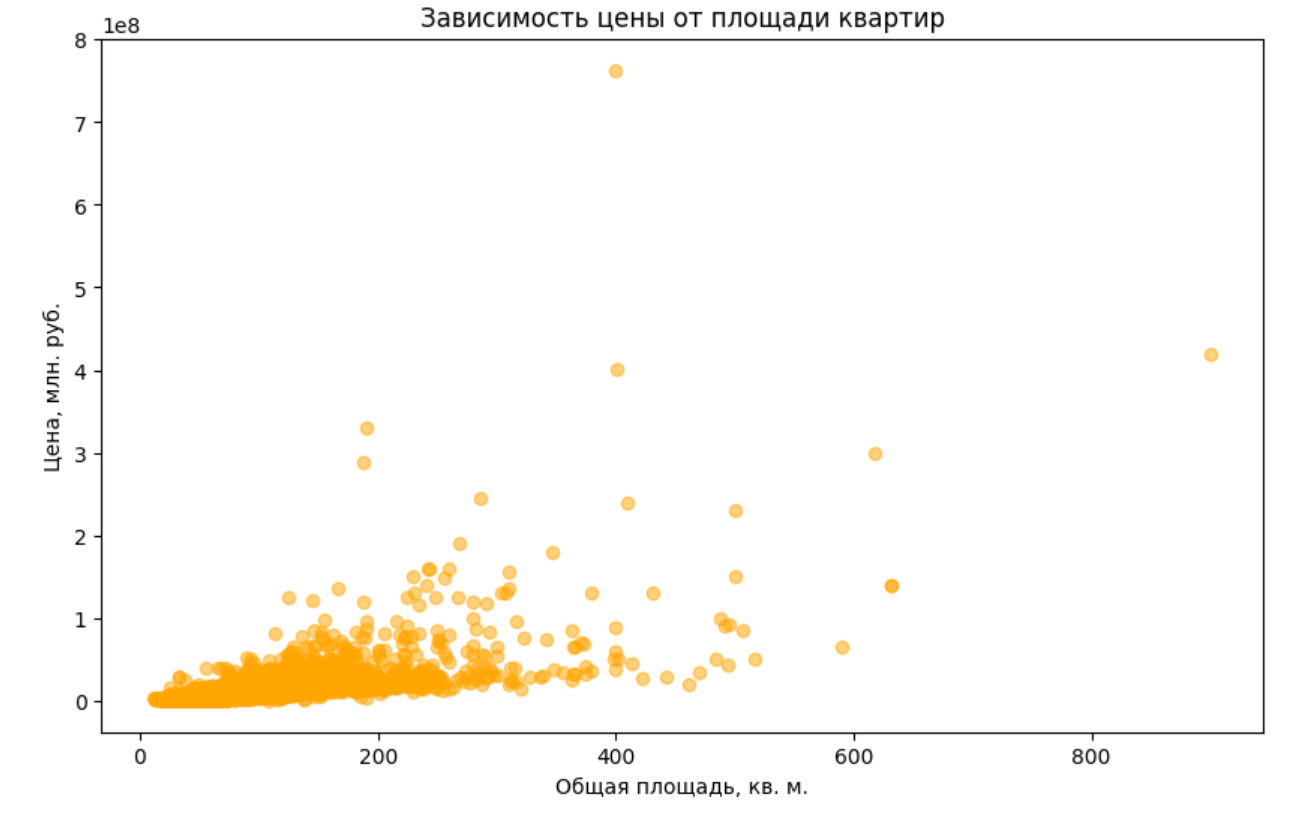
plt.title('Зависимость цены от площади квартир')

plt.xlabel('Общая площадь, кв. м.')

plt.ylabel('Цена, млн. руб.')

plt.show()

Из графика следует, что зависимость цены от общей площади слабая, но есть – с увеличением площади растет и стоимость квартир.



# Раздел 5. Вывод.

Нами была выполнена задача по обработке и анализу данных по анализу продаж недвижимости в сервисе Яндекс Недвижимость. В ходе работы были получены новые данные и отредактированы старые, а также проведён анализ всех полученных нами данных. Для решения этой задачи нами были использованы следующие библиотеки python: pandas, numpy, matplotlib. Задача была решена в среде разработки Jupyter Notebook.