Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Отчёт по кейсу «Самолёт»:**

Работу выполнил:

Студент группы № ИСП-21

Новасельский Артём

Долгопрудный, 2024

# Введение

В данном отчёте рассматриваются выводы, полученные после анализа данных в области «Квартиры в Москве и Московской области».

# Цель

Собрать данные и проанализировать их для будущего использования, например, обучение машины на основе выводов.

# Задачи

* Собрать данные с помощью данных инструментов.
* Совершить работу над ними, а точнее удаление ненужных данных, дополнение необходимых и т.д.
* Визуализация данных. Нахождение взаимосвязи между данными или её полное отсутствие для отчёта.

# Основная часть

Существует небольшой выбор источников, для сбора данных, нам был предложен интернет-ресурс «Циан». С помощью языка Python и библиотеки “CianParser” было собрано свыше десяти тысяч объявлений в нужных регионах.

После сбора всей информации воедино и уборки дубликатов, можно посмотреть, какого типа наши данные (рис.1).

Теперь мы смотрим, какие данные у нас не смогли собраться(рис.2) при помощи библиотеки seaborn и функции heatmap. Как мы видим, например, колонка residential\_complex сильно пустеет, что означает мы вынуждены их удалить, так как строить анализ будет невозможно.

Далее фильтруем ненужные данные и форматируем столбцы, чтобы их было можно анализировать(рис.3). После объёмной чистки данных, нужно проверить их состояние – смотрим внутрь файла и бегло проверяем на аномалии, в случае их отсутствия приступаем к кодовой проверке данных.

После полной очистки данных, вручную и программно, можно сохранить очищенную базу данных, а после, приступать к постройке графиков и аналитической работе при помощи библиотеки matplotlib для вывода графических изображений. Например, будет 5 графиков:

1. Цена за м^2 в зависимости от этажа, на котором квартира.
2. Цена за м^2 по району.
3. Цена за м^2 по городу.
4. Цена за м^2 в зависимости от года постройки самой квартиры.
5. Количество объявлений по количеству комнат.

(рис.4-6)

Ещё можно вывести корреляционную матрицу, или матрицу корреляций, которая напрямую показывает зависимость значений друг от друга(рис 7).

В итоге получаются графики(рис.10), на основе которых уже можно проводить анализ.

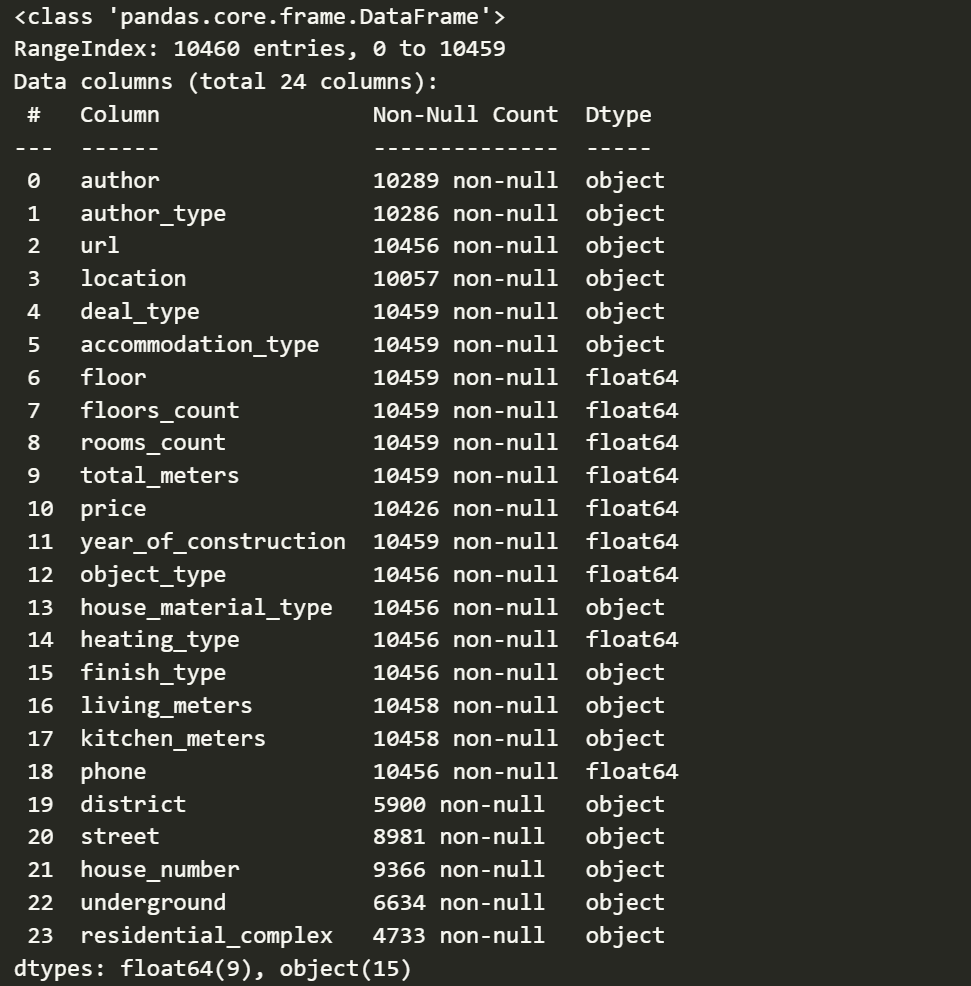
# Аналитика данных

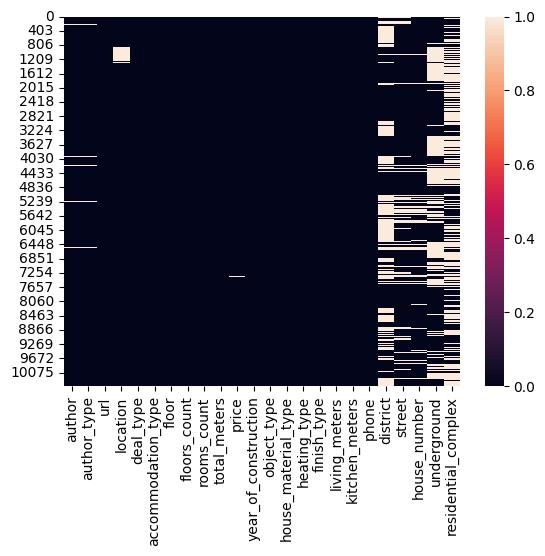
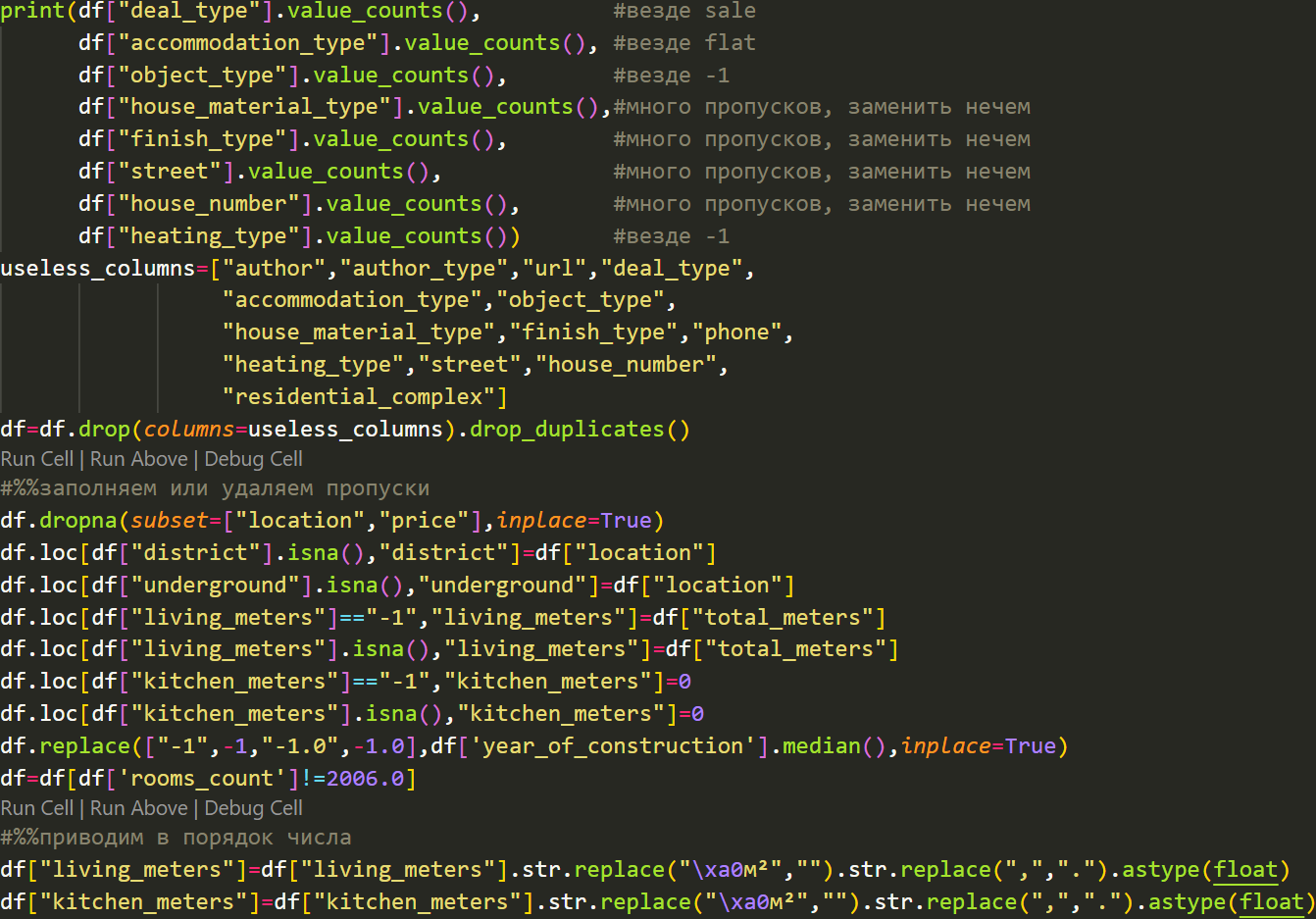
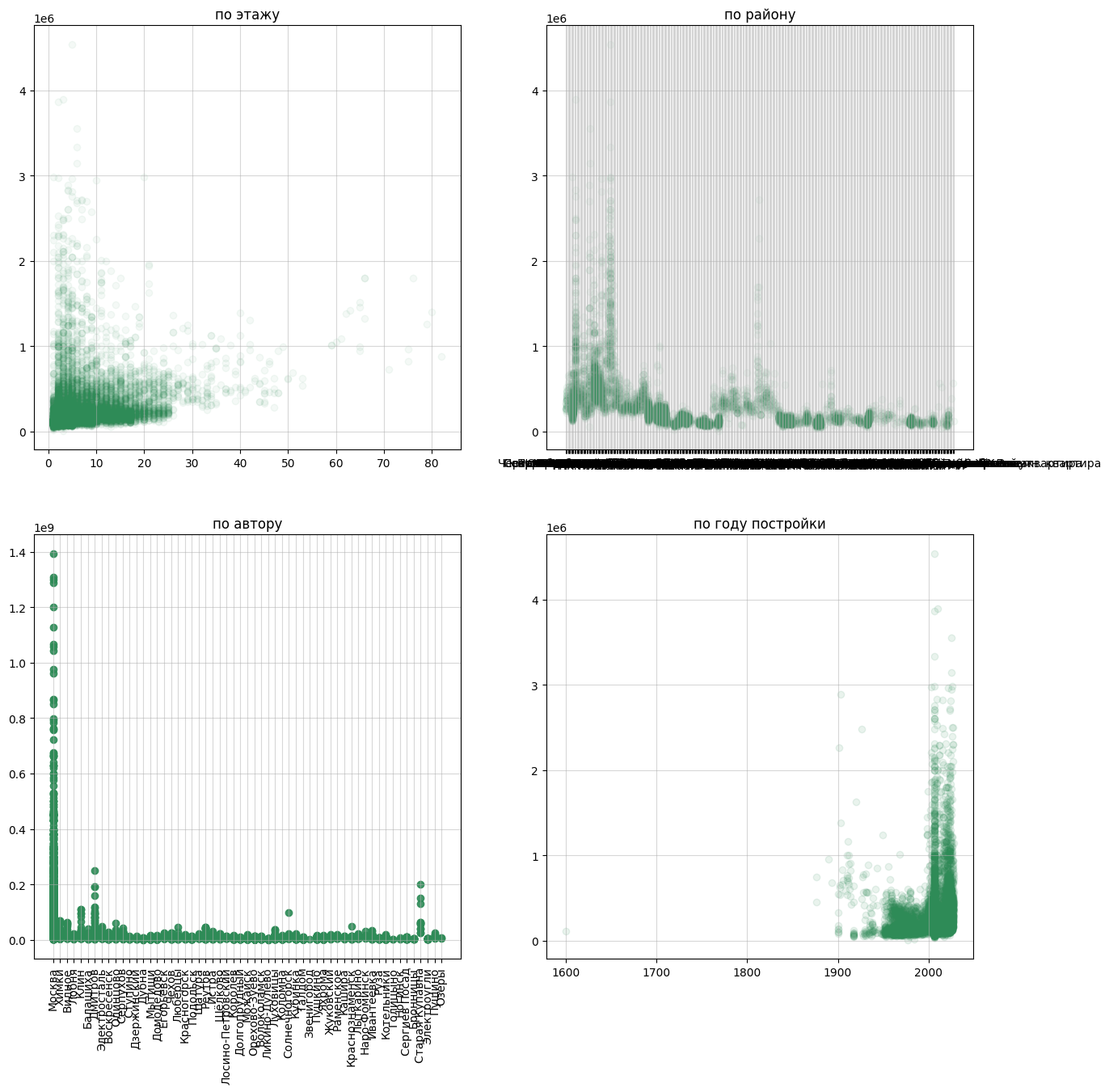
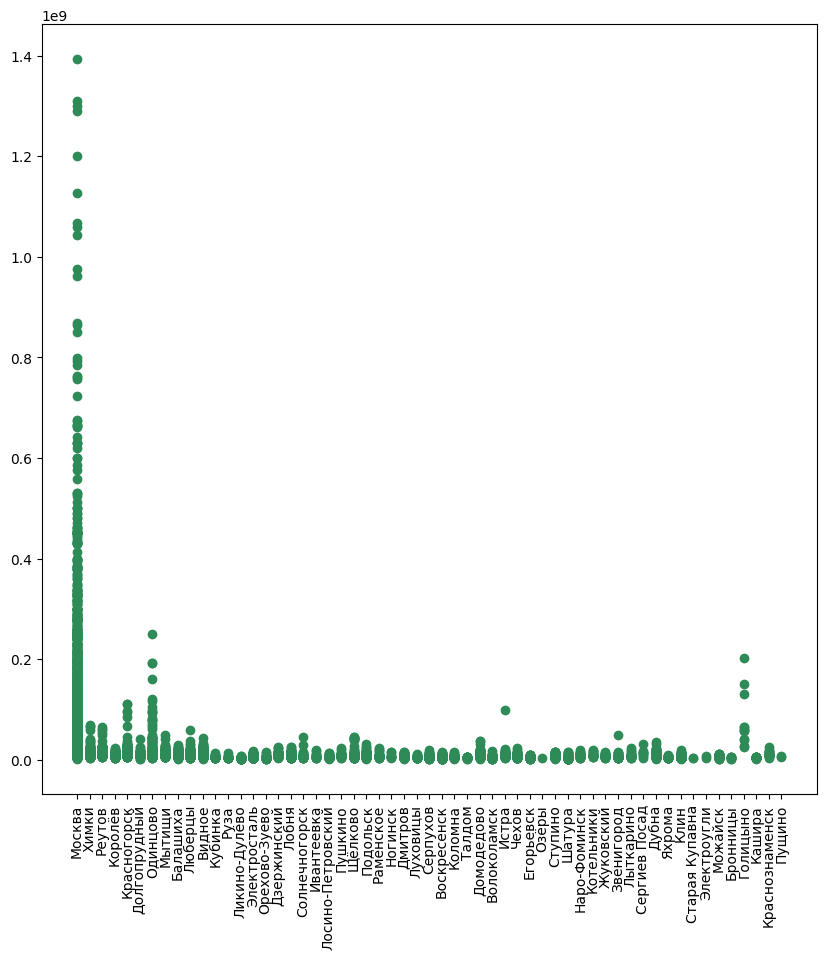
Смотря на графики, делаем выводы, что цена в основном зависит от количества комнат и метража. Меньше всего на цену влияют год постройки и количество этажей.

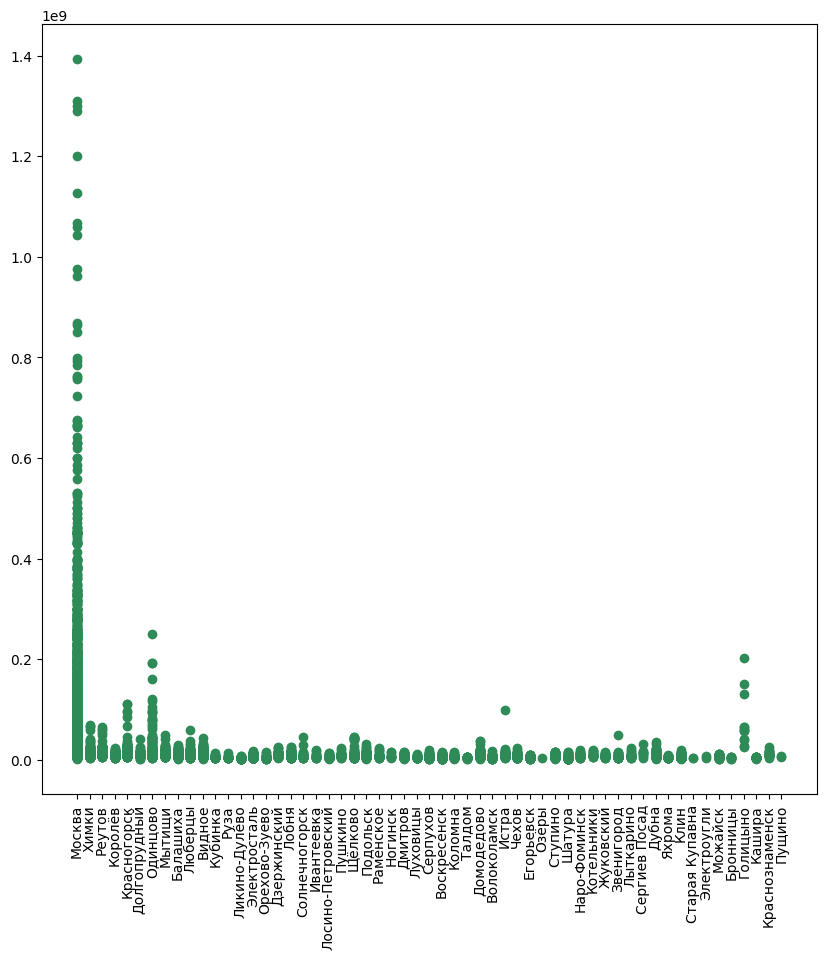
# Заключение

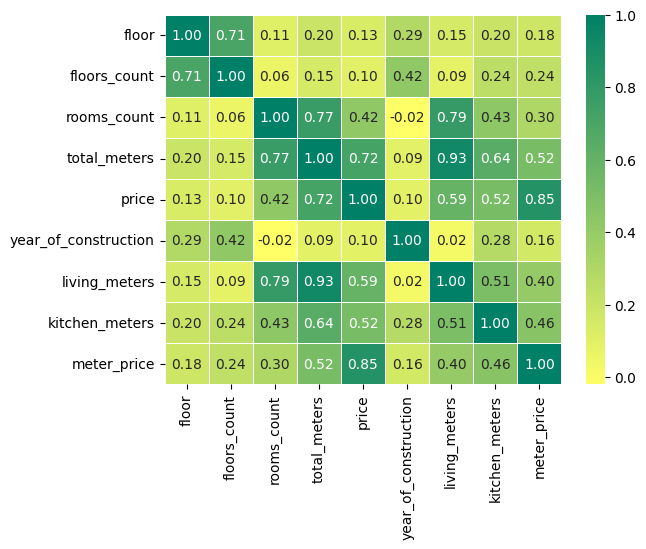
В результате работы были собраны около одиннадцати тысяч квартир, отсортированы, почищены данные, после чего, на удивление осталось почти десять тысяч, простроены облегчающие анализ графики, которые помогли понять главные критерии в оценивании стоимости недвижимости в Москве и Московской Области. Основными факторами, влияющие на стоимость, выявились метраж и количество комнат.

В теории можно было бы и закодировать категориальные данные, на что мне не хватило времени, знаний и сил, и тщательней прочесать данные, ибо тот факт, что в годе постройки были поля “Напишите Автору” и “Аукцион” показывает качество данных в негативном ключе, и скорее всего было что-то пропущено, ещё, наверно, можно было бы и использовать второй инструмент для анализа, что б точно закрепить выводы о данных.

рис.1

рис.2 рис.3 рис.4 рис.5

рис.6

рис.7