Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Отчёт по кейсу «Самолёт»:**

Работу выполнил:

Студент группы № ИСП-21

Бухаров Егор

Долгопрудный, 2024

# Введение

В данном отчёте рассматриваются выводы, полученные с аналитической работы над данными в области «Квартиры в Московской области, Новой Москве и Москве».

# Цель

Собрать данные и произвести аналитическую работу над ними для будущих работ, например, создание модели на основе выводов.

# Задачи

* Используя открытые источники собрать список данных.
* На основе полученной информации произвести удаление ненужных данных, дополнение необходимых, выявление аномалий и их блокировка.
* Визуализация данных при помощи, как минимум, двух инструментов для подобных задач. Нахождение взаимосвязей между данными или их полное отсутствие, усреднённых показателей для уверенного отчёта.

# Основная часть

Для выполнения основной задачи, существует небольшой выбор источников, откуда собирать данные, мною был выбрать интернет-ресурс «Циан». При помощи скриптов, написанных на языке Python и библиотеке CianParser было получено свыше десяти тысяч объявлений в нужных регионах.

К следующей задаче подходит такое начало, как соединение собранных данных в одну таблицу при помощи написанной функции с библиотекой Pandas.

После сбора всей информации воедино и уборки дубликатов, нужно узнать, какого типа наши данные (рис.1), так как отталкиваясь от типа данных, мы будем применять разные методы к их сортировке.

Теперь мы убираем -1 как базовое не собранное значение и смотрим, какие данные у нас смогли собраться(рис.2) при помощи библиотеки missingno. Как мы можем увидеть, object\_type и heating\_type практически нигде не указываются, значит мы вынуждены их удалить, так как строить анализ будет невозможно.

Далее отсеиваем ненужные данные и форматируем некоторые столбцы, чтобы их было легче анализировать с помощью сторонних инструментов(рис.3). После выполнения объёмной чистки данных, нужно проверить их состояние – смотрим внутрь файла и бегло проверяем на аномалии, в случае их отсутствия приступаем к кодовой проверке данных(рис.4-7).

После полной очистки данных вручную и программно можем приступать к сбору графиков/аналитической работе при помощи библиотеки matplotlib для вывода графических изображений. Например будет 4 графика:

1. Цена за м^2 по городам.
2. Цена за м^2 в зависимости от материала, используемого при постройке здания.
3. Количество объявлений по городам.
4. Количество объявлений по годам постройки здания.

Для первых и последних двух графиков будем использовать один метод подсчёта данных.

Первый метод – отбор цены за квадратный метр по трафаретному коду и запись в csv-файл(рис.8).

Второй метод – использование встроенной функции .value\_counts() и запись в csv-файл для дальнейшей обработки(рис.9).

В итоге получаем графики(рис.10), на основе которых уже можно проводить анализ, но мы перейдём к составлению графиков на Power BI.

Power BI – максимально удобный инструмент для составления графиков и аналитики данных.

Для первого графика мы выбираем данные price, меняем сумму на «среднее» и включаем их в график, на другую ось ставим «year\_of\_construction». Выбираем тип графика и получается примерный график со средней ценой квартиры, в зависимости от года его постройки(рис.11).

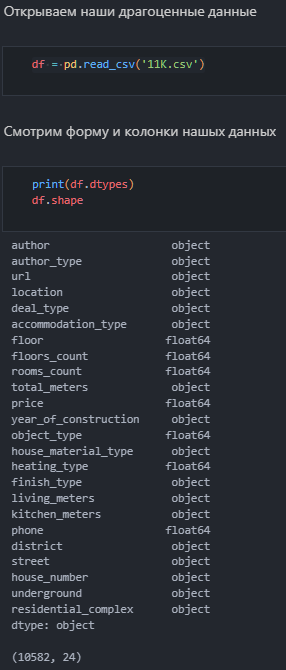
Второй график будет содержать в себе среднюю цену квартиры, в зависимости от материала здания(рис.12), просто вместо года постройки ставим тип материала. Добавлю к этим данным среднюю цену по виду отделки(рис.13). UPD: все графики были заменены на один dashboard(рис.14).

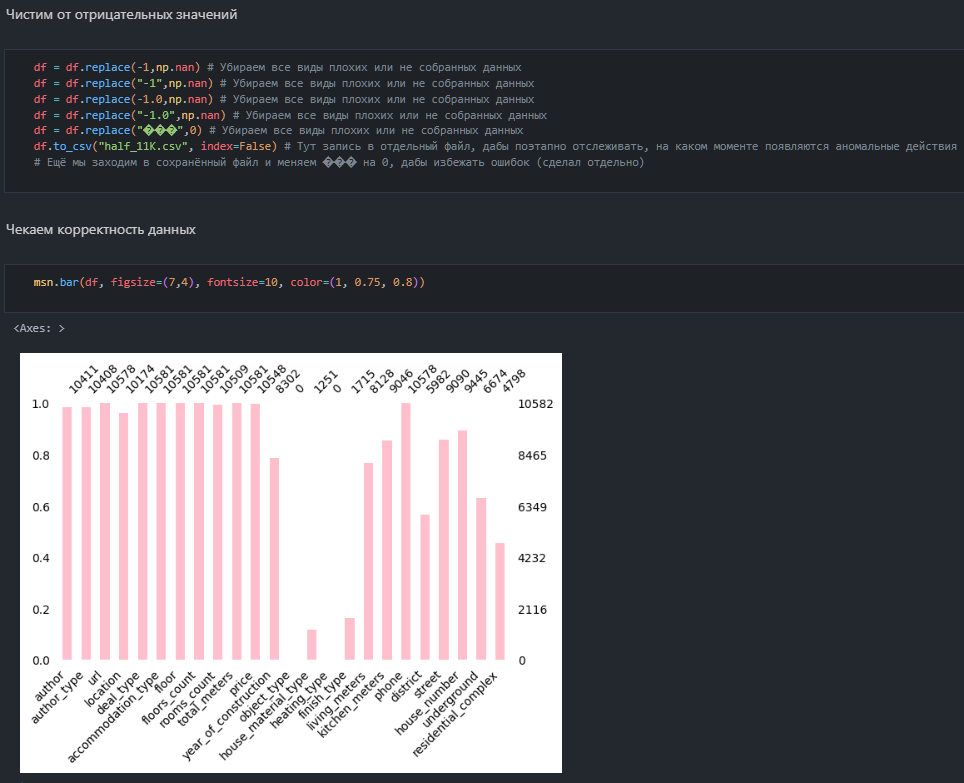
# Аналитика данных

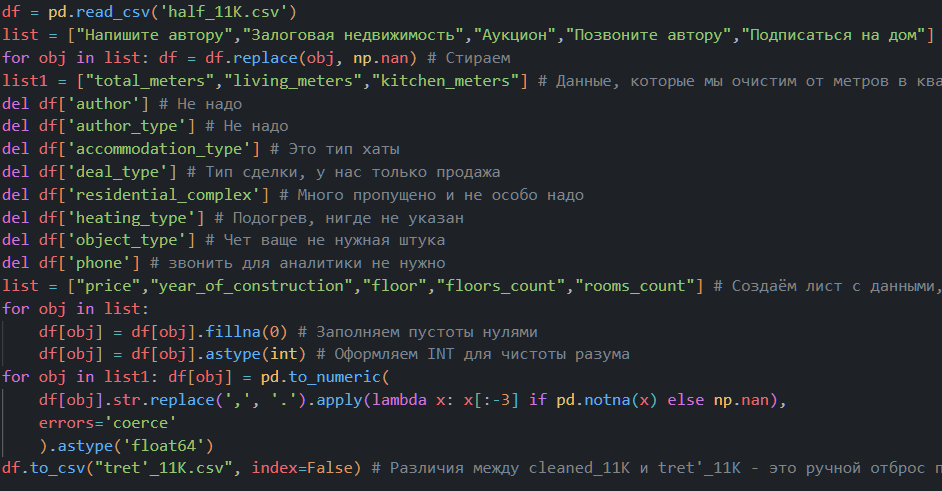
Благодаря выведенным графикам, можно сделать выводы, что цена в основном зависит от типа отделки, материала дома, города. От года постройки зданий зависит лишь их количество на рынке и количество комнат во время СССР, а на цену никак не влияет.

# Заключение

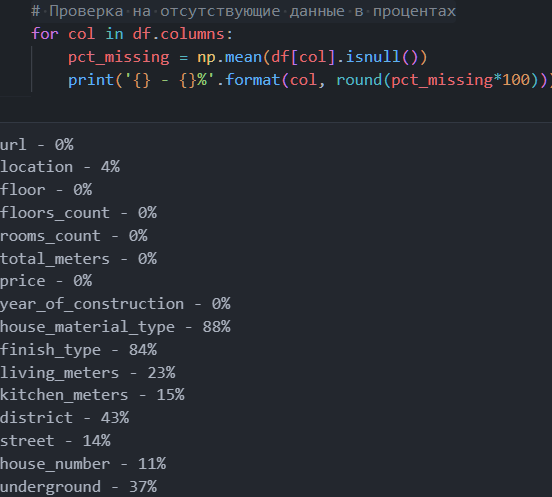
В результате аналитической работы были собраны, отсортированы, почищены данные, простроены удобные для анализа графики, благодаря которым получилось выявить не маловажные критерии в оценивании стоимости недвижимости в Московской Области, Москве и Новой Москве. Основными факторами, оказывающими влияние на стоимость, выявились тип отделки, материала здания и расположение. Полученные данные могут быть использованы для дальнейшей разработки прогностических моделей.

(рис.1)

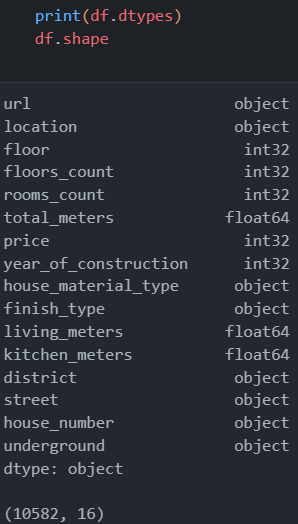
(рис.2)

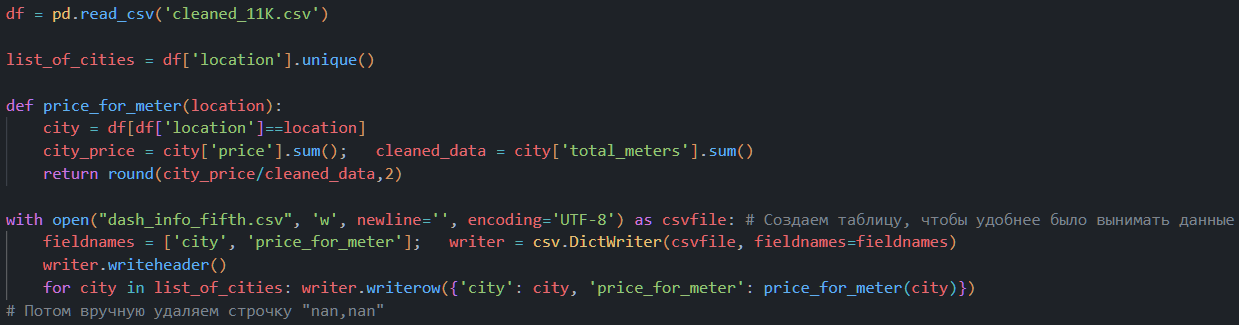
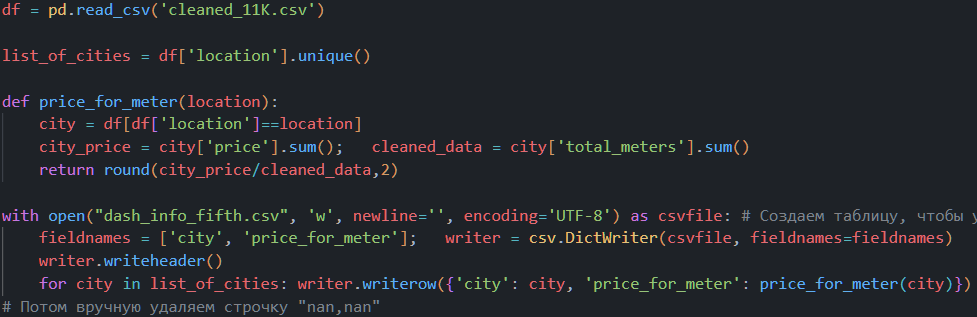
 (рис.3)

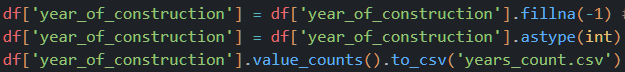
(рис.4)

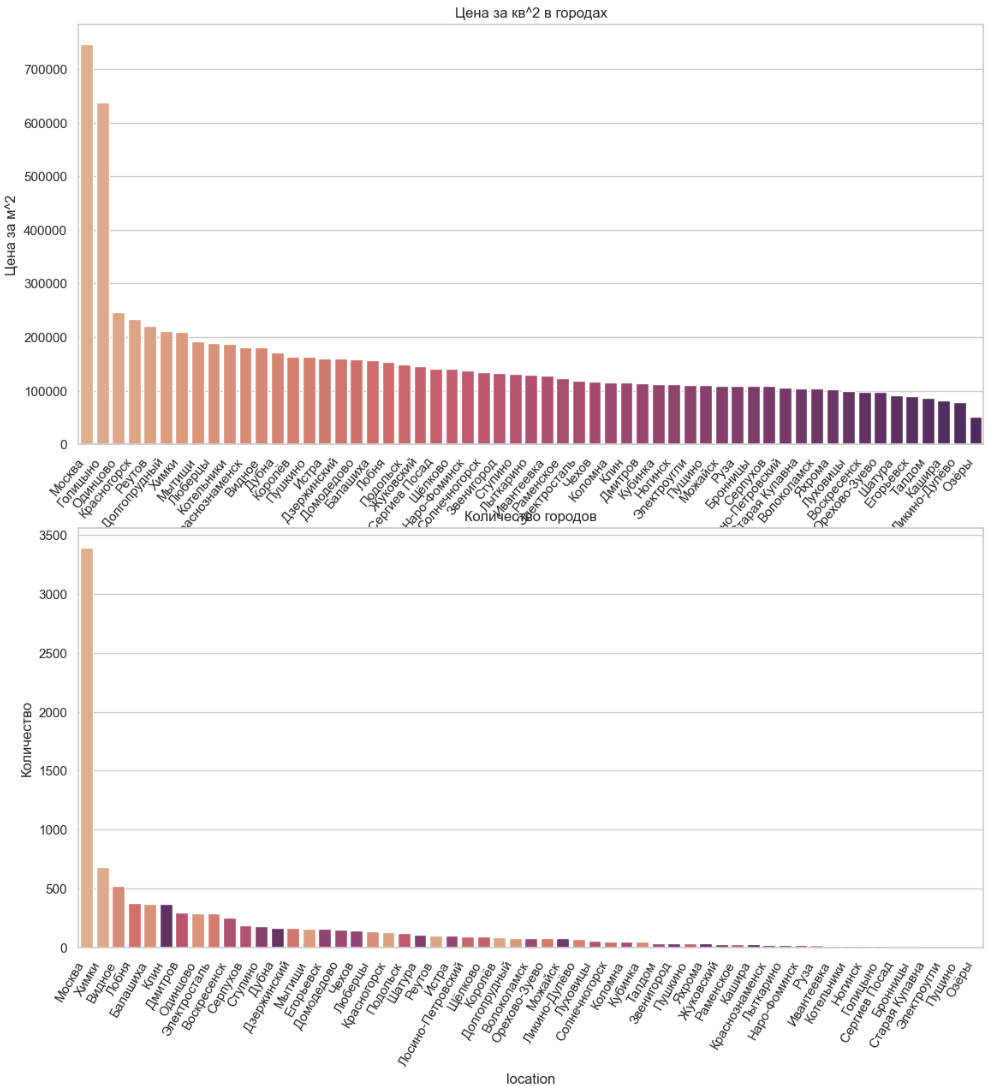
(рис.5)

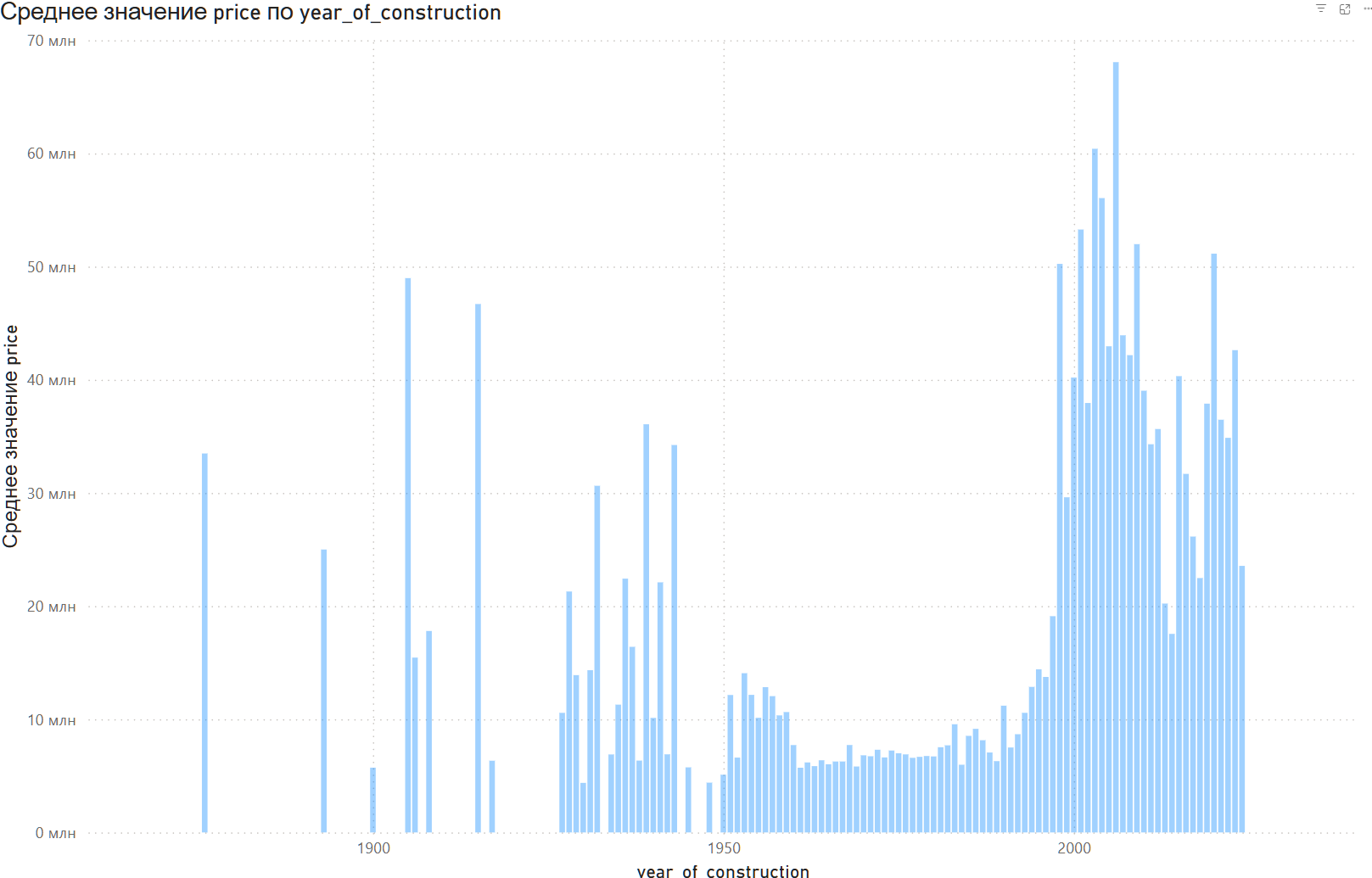
(рис.6)

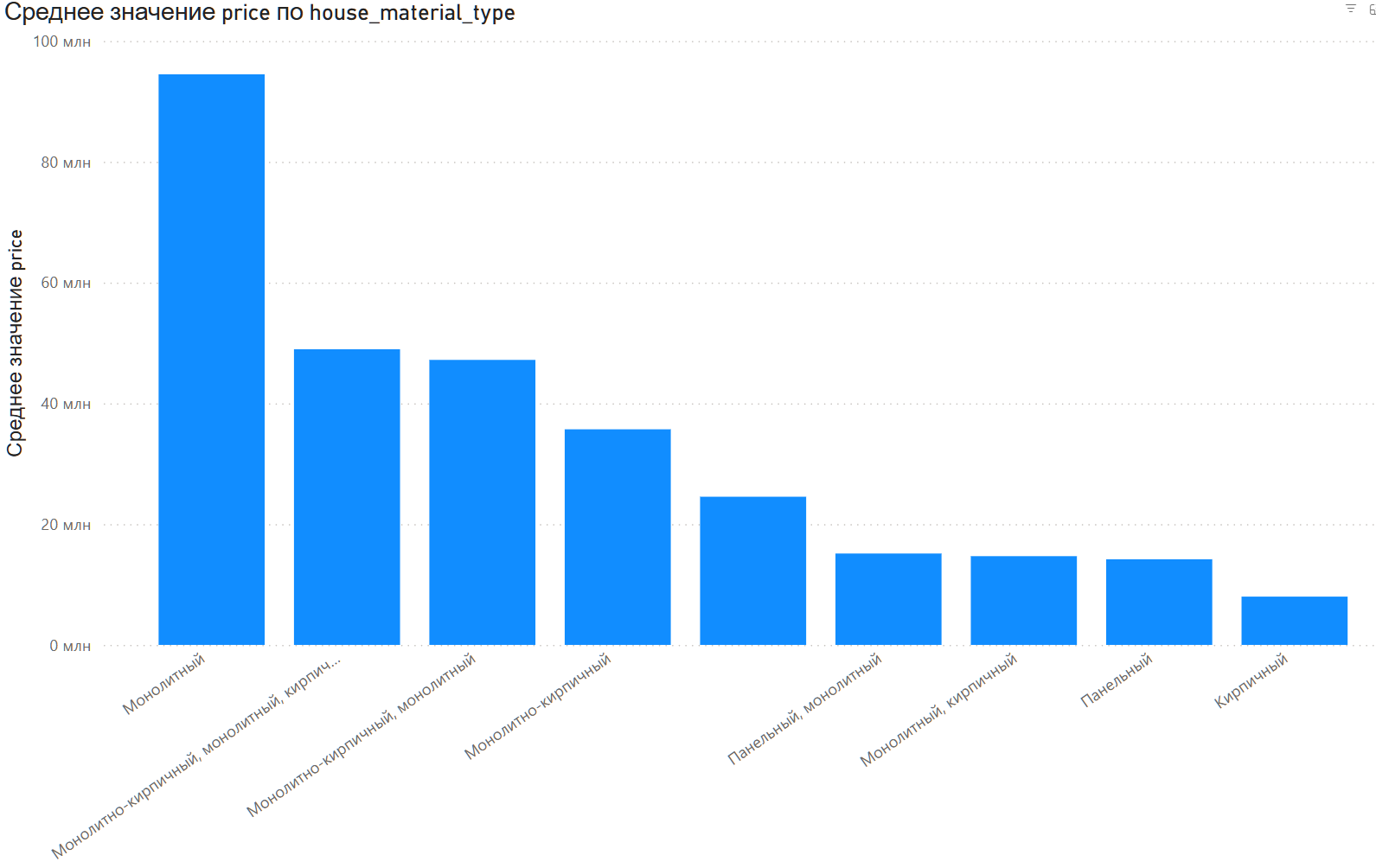
(рис.7)

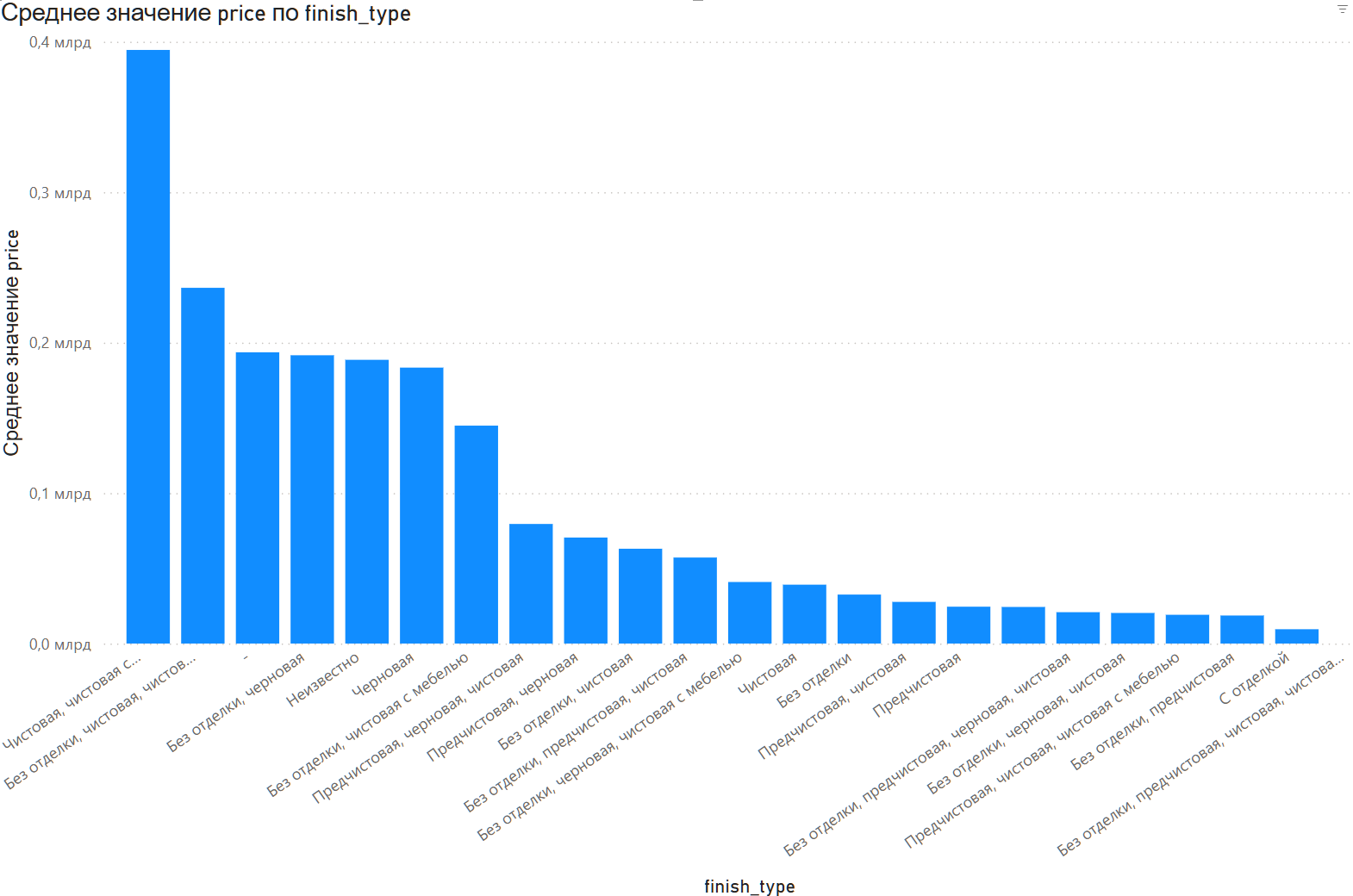
(рис.8)

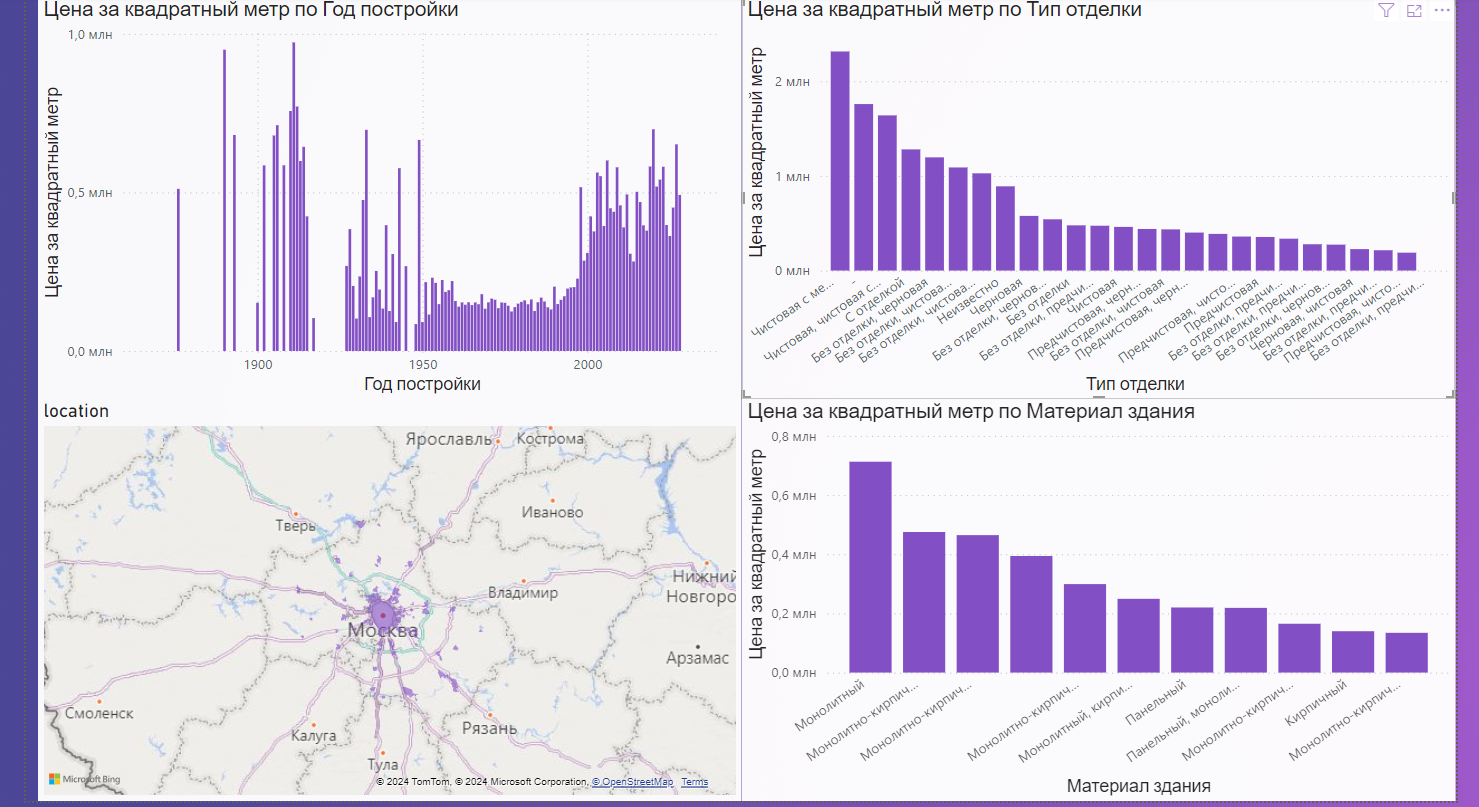
(рис.9)

(рис.10)

(рис.11)

(рис.12)

(рис.13)



(рис.14)