Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Аналитический отчёт в рамках работы с кейсом компании «Самолёт»**

***«Модель оценки цены квартиры на вторичном рынке по Московскому региону: Москва, Новая Москва, Московская область»***

Работу выполнил:

Студент группы ИСП-21

Канин Илья

Проверил:

преподаватель дисциплины

Базяк Г.В.

Долгопрудный, 2024

**СОДЕРЖАНИЕ**

[НАЧАЛЬНЫЙ ЭТАП 3](#_Toc181173968)

[ОЧИСТКА ДАННЫХ 4](#_Toc181173969)

[1. Подготовительный этап 4](#_Toc181173970)

[2. Удаление ненужных колонок 5](#_Toc181173971)

[3. Работа с пропущенными значениями 5](#_Toc181173972)

[3.1. Удаление очевидных значений 5](#_Toc181173973)

[3.2. Промежуточное планирование 5](#_Toc181173974)

[3.3. Проработка оставшихся пропущенных значений 6](#_Toc181173975)

[4. Изменение типов данных 8](#_Toc181173976)

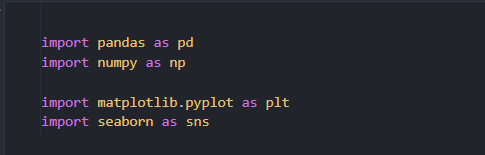
[ДОБАВЛЕНИЕ НОВОЙ ПЕРЕМЕННОЙ 9](#_Toc181173977)

[ВИЗУАЛИЗАЦИЯ ДАННЫХ, ВЫВОДЫ 9](#_Toc181173978)

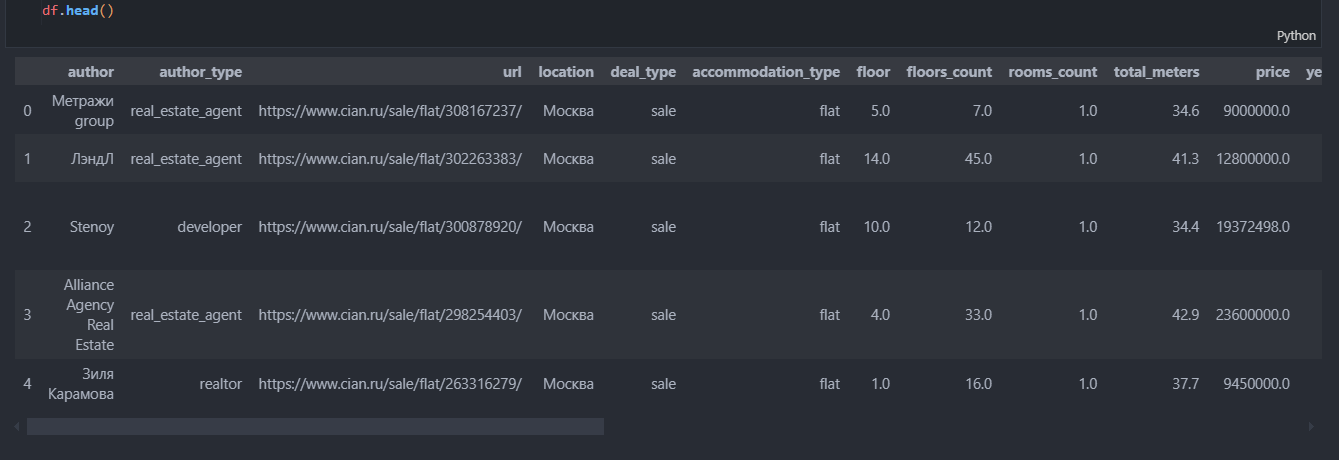
[ВЫВОД 12](#_Toc181173979)

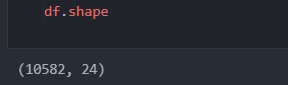
# НАЧАЛЬНЫЙ ЭТАП

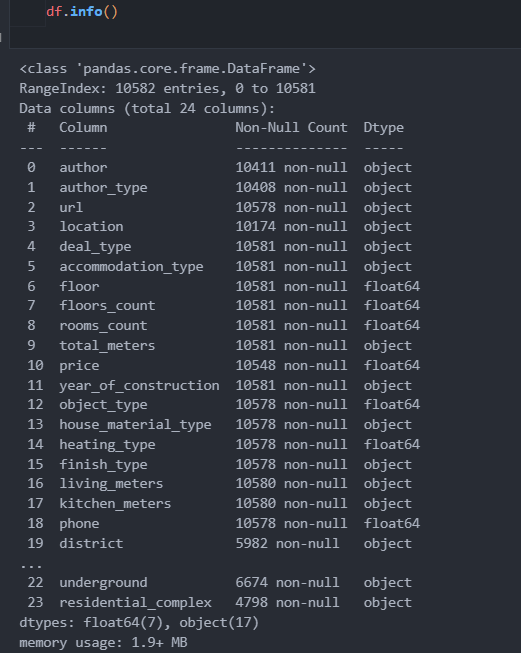
Импорт библиотек для работы с датасетом и создания графиков:



Для ознакомления с датасетом я вывел срез по первым данным, размерность и краткую информацию по колонкам:



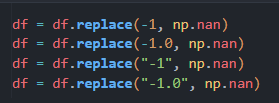




# ОЧИСТКА ДАННЫХ

## 1. Подготовительный этап

В процессе парсинга пустые значения могли быть заполнены или nan, или -1. Чтобы корректно понимать количество пропущенных значений заменяю значение -1 на NaN по всему датасету (и т.к. значения -1 принимают разную форму, для надежности заменяю её в различных вариациях)



После этого повторно вывожу информацию, немного более полную чем в начальном этапе, и составляю для себя план очистки датасета:

1. Cразу принимаю за ненужные колонки, которые не будут влиять на анализ и которые можно удалить сразу, следующие:

- author (информация о том, как зовут человека, который продаёт квартиру, или название компании, которая этим занимается не будет влиять на цену)

- url (ссылки на обьявление никаким образом не будут влиять на цену)

- deal\_type (по всему датасету один тип сделки - продажа)

- accomodation\_type (по всему датасету один тип жилья - квартира)

- phone (телефон для связи никак не будет влиять на цену)

- object\_type (все значения отсутсвуют)

- heating\_type (все значения отсутствуют)

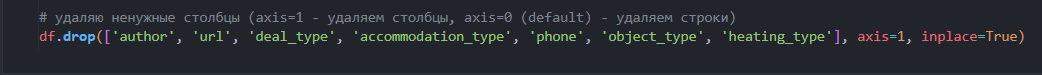
2. Необходимо разобраться с пропущеными значениями: удалить строки, где пропущены ключевые значения и колонки, где пропущено слишком много значений. также заполнить колонки данными, где это возможно сделать, чтобы не потерять большое количество файлов;

3. Понимаю, что нужно изменить типы данных колонок:

- из float в int - floor, floors\_count, rooms\_count, price

- из object в int или float - total\_meters, year\_of\_construction, living\_meters

## 2. Удаление ненужных колонок



## 3. Работа с пропущенными значениями

### 3.1. Удаление очевидных значений

Для начала удаляю строки, в которых отсутствуют значения по ключевым переменным:



После этого удаляю колонки, в которых количество пропущенных значений критично:



### 3.2. Промежуточное планирование

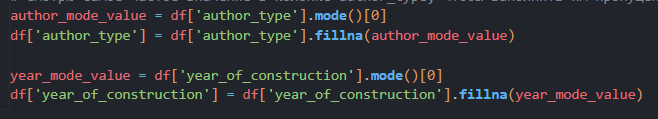
Работа с пропущенными значениями в остальных колонках будет не такой простой, так что для удобства подвожу промежуточные итоги по избавлению от пропущенных значений и составляю небольшой список предстоящих действий с колонками, чтобы было легче ориентироваться и работать с данными в дальнейшем:

* autor\_type - можно заполнить самым частым значением;
* year\_of\_construction - также можно заполнить самым частым значением, это не будет особо критично, но зато не придется терять 21% от датасета;
* living\_meters - можно попробовать сделать функцию для рассчёта жилых метров, основываясь на пропорции общего метража с жилым;
* kitchen\_meters - аналогично с living\_meters;
* district - если обьявление не из москвы ставить туда город, если москва - метро (если оно есть) или просто значение москва (если значение метро - nan);
* street - пропущенных значений не очень много, теоретически можно их удалить, либо удалить колонку;
* house number - возможно стоит снести всю колонку, потому что сомнительно что эта переменная будет влиять на цену;
* underground – метро, важный фактор, можно заполнить городом, что не хотелось бы делать, либо удалить строки с пропущенными значениями, хотя и удалится достаточно много;
* residential\_complex - слишком много пропущенных значений, заполнить проблематично, влиять на цену должно не сильно существенно. Лучшим решением будет удалить колонку.

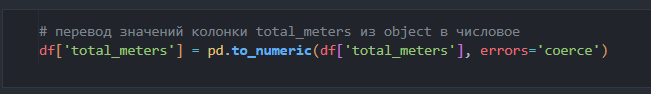
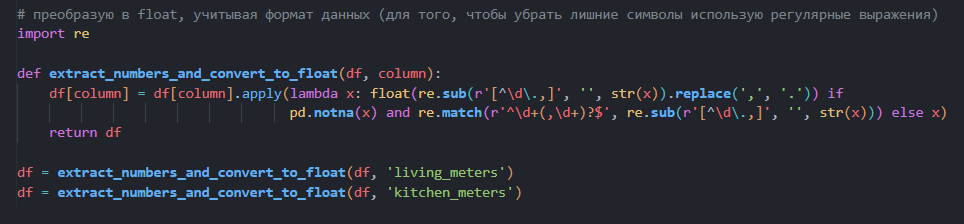
### 3.3. Проработка оставшихся пропущенных значений

Далее по этому плану заполняю колонки **author\_type и year\_of\_construction**

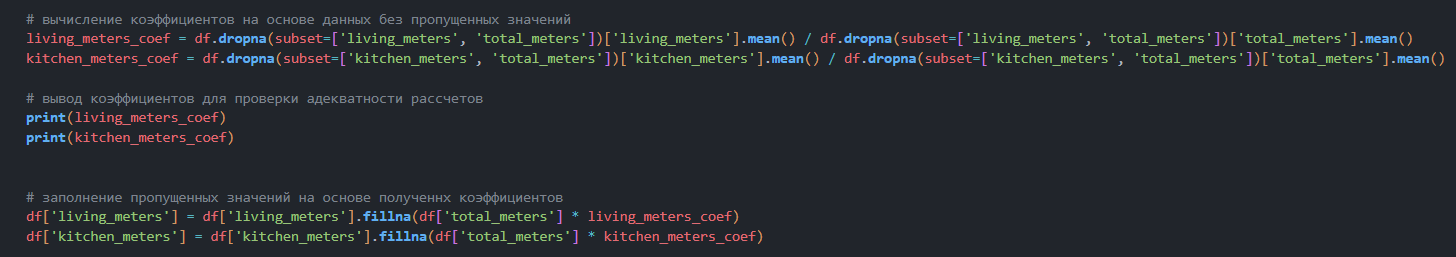
модой:



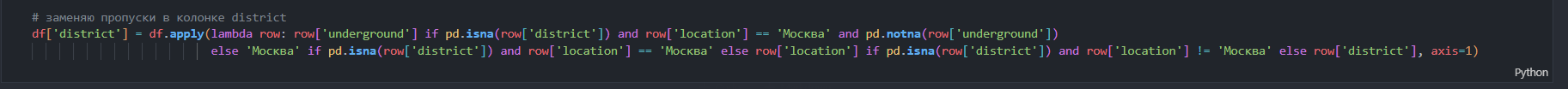
Для заполнения колонок kitchen\_meters и living\_meters сначала перевожу их и колонку total\_meters к правильному типу данных:



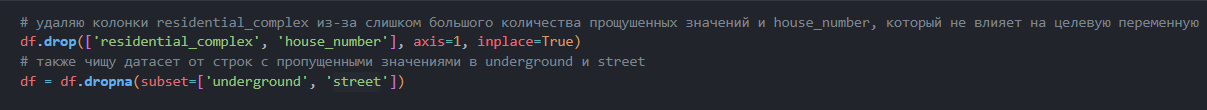
После чего заполняю значениями:



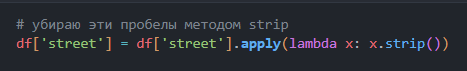
Колонку district заполняю по вышеописанному принципу:



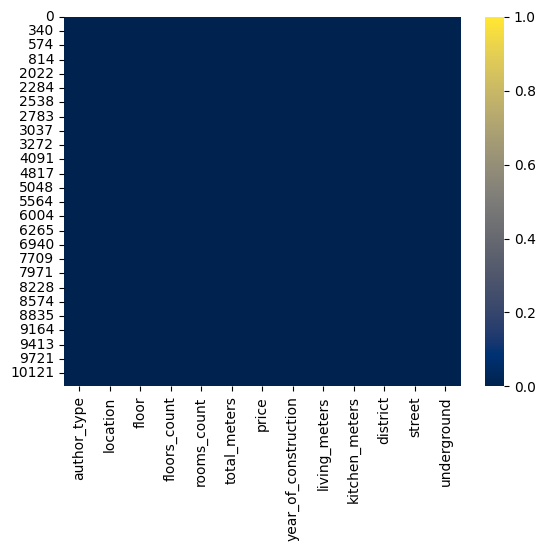
Далее удаляю колонки house\_nubmer и residental\_complex, а также строки с пропущенными значениями в street и underground:

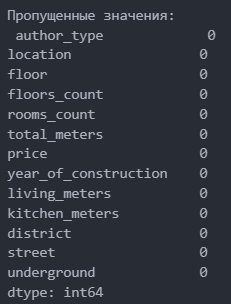


Также в процессе удаления строк нахожу ошибку в формате данных в колонке street, сразу убираю её:



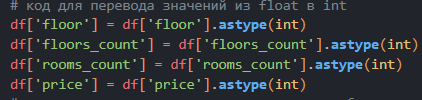
После чего вывожу тепловую карту, которая отображает количество пропусков, а также количество пропусков в числовых значениях, чтобы удостоверится в том, что их не осталось:



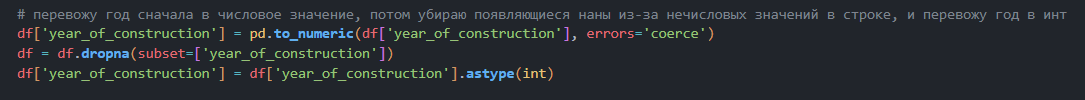


## 4. Изменение типов данных

Преобразую типы данных колонок floor, floors\_count, rooms\_count и price из float в int



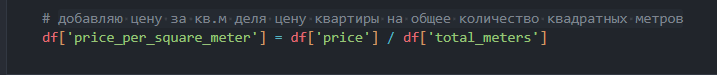
Преобразую тип данных колонки year of construcrion в int из object



На этом чистку можно считать завершенной.

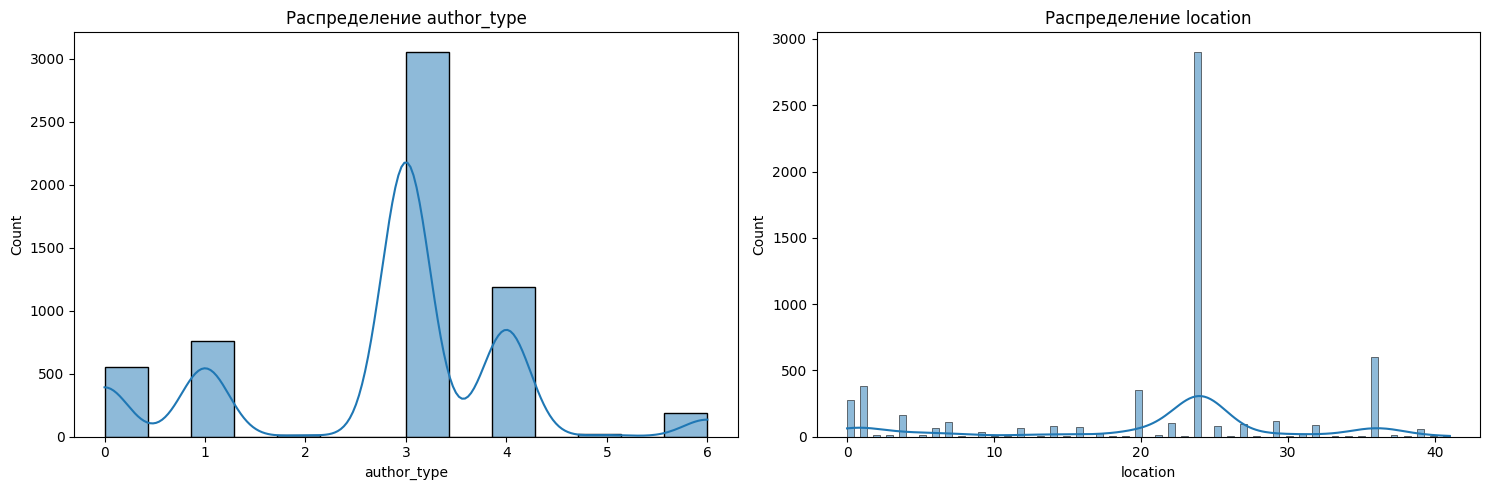
## ДОБАВЛЕНИЕ НОВОЙ ПЕРЕМЕННОЙ

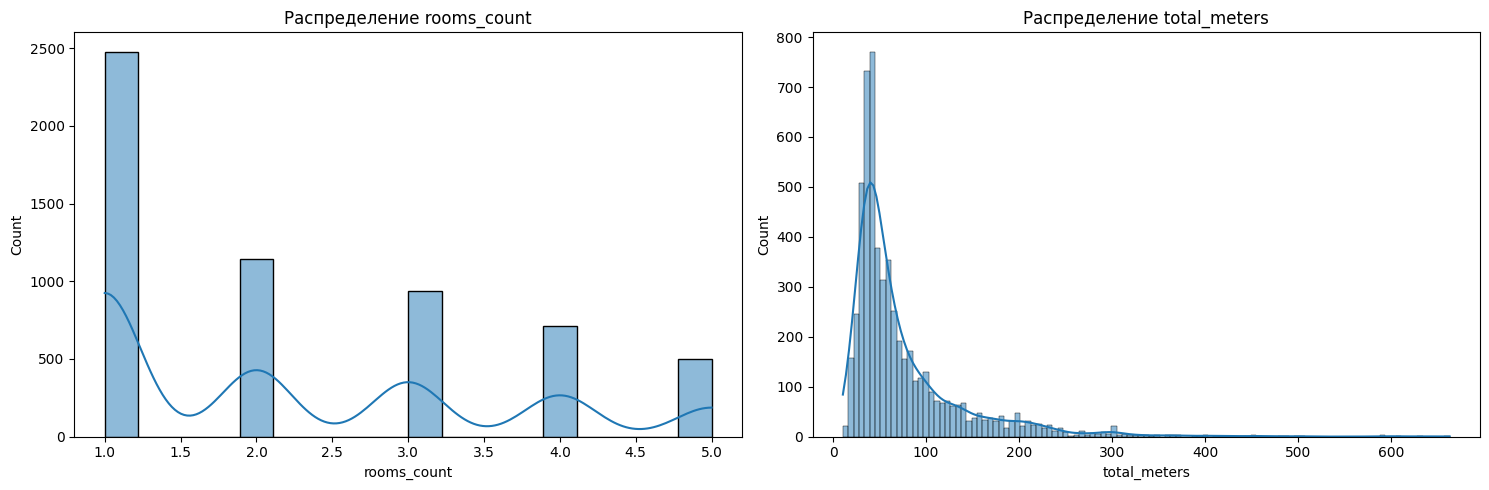
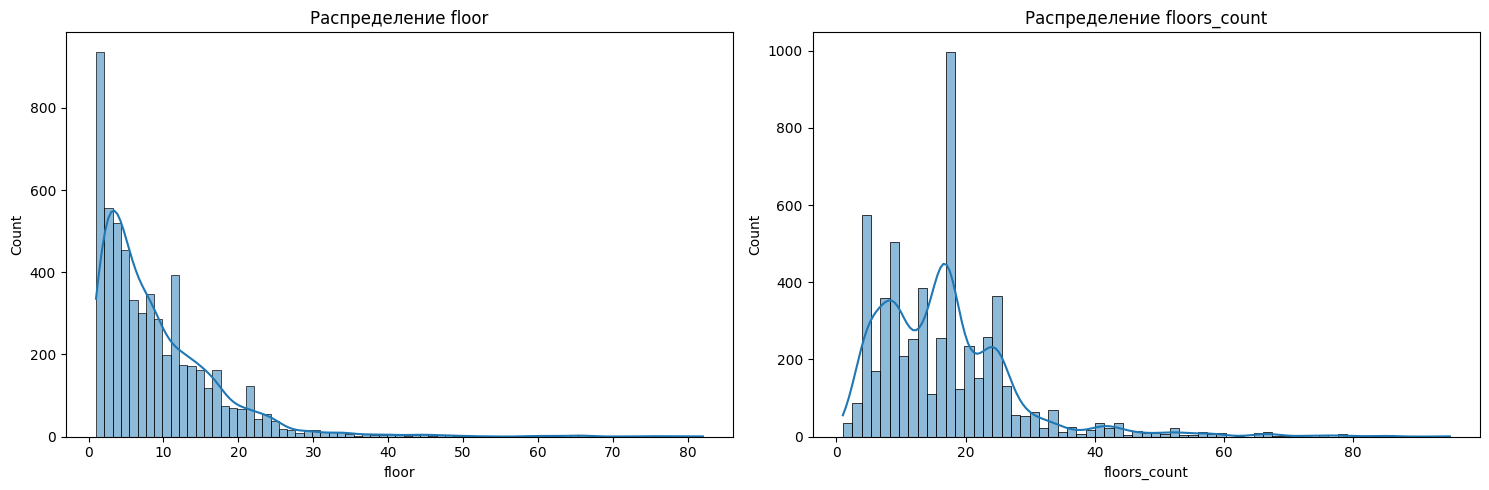
Добавляю цену за кв.м, деля цену квартиры на общее количество квадратных метров:

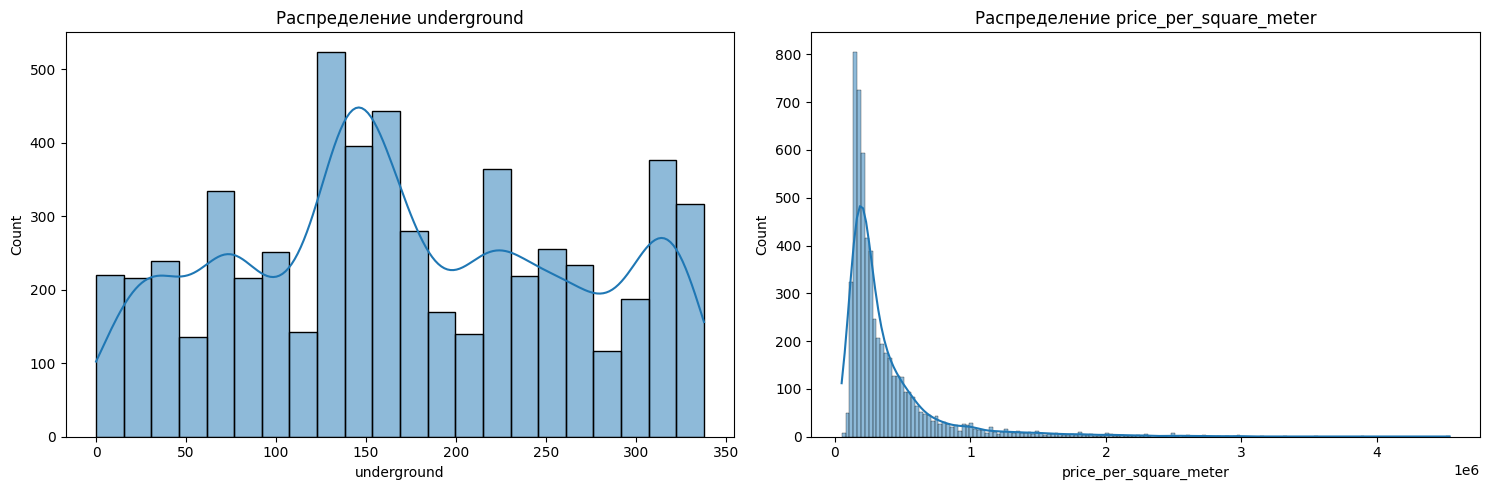
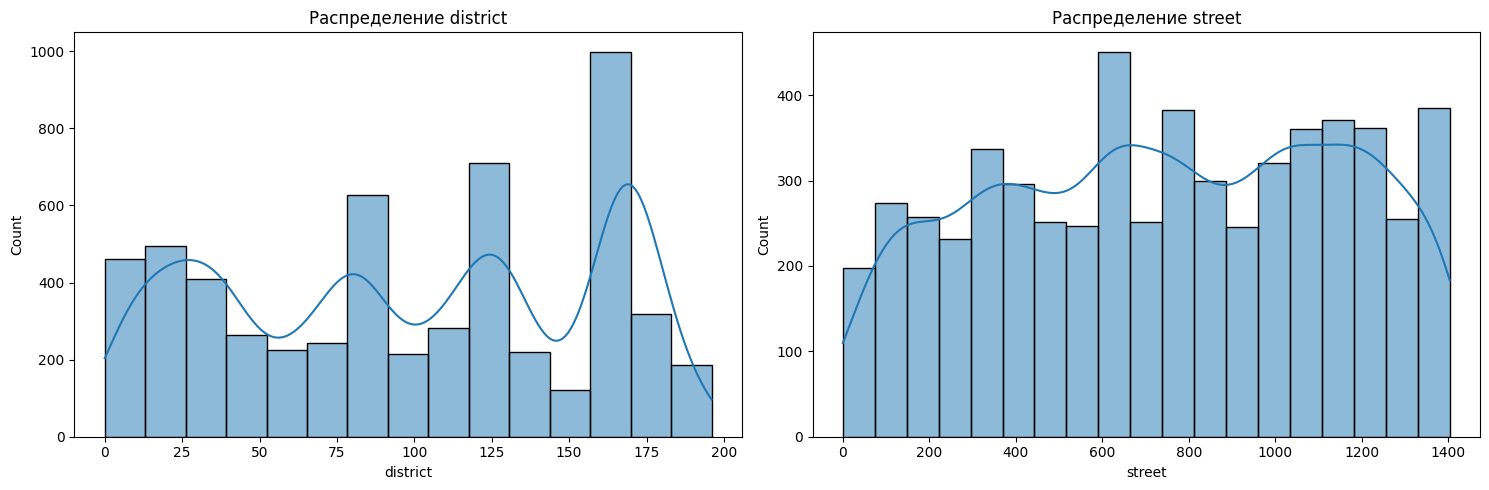
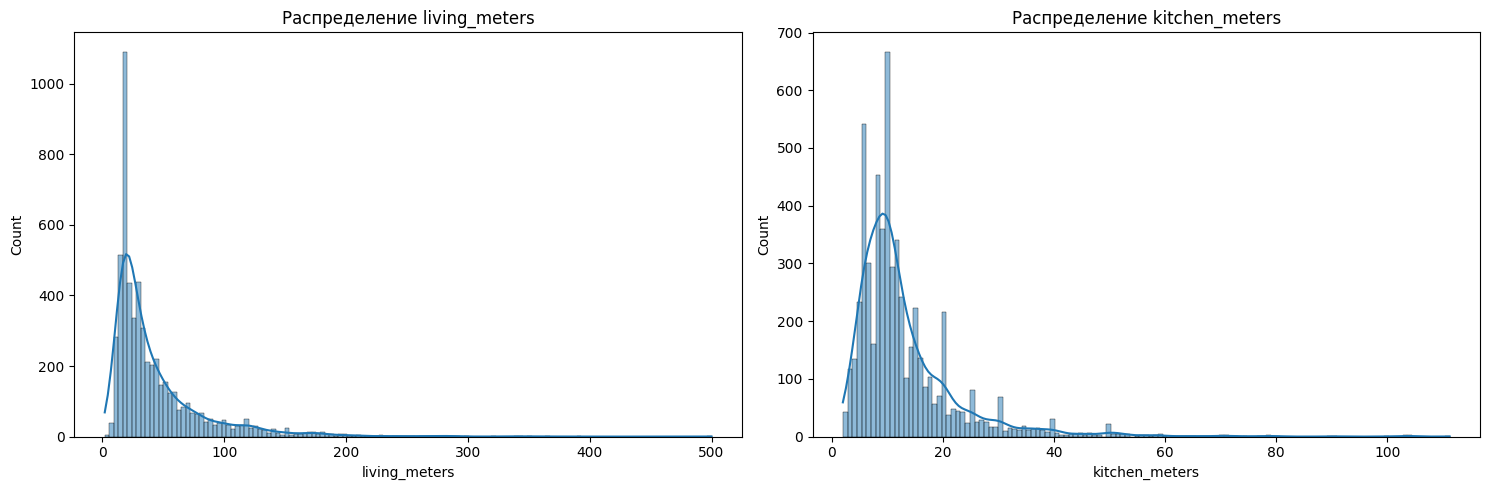
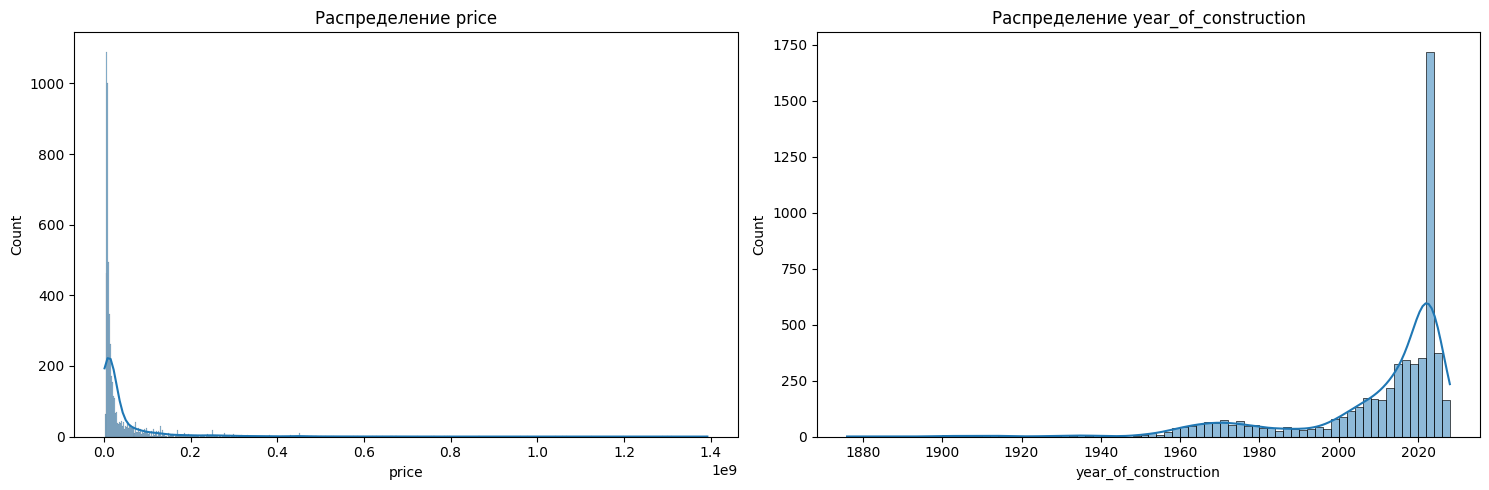


# ВИЗУАЛИЗАЦИЯ ДАННЫХ, ВЫВОДЫ

После очистки датасета перекодирую категориальные данные для дальнейшего удобства работы с ними и вывожу распределения по всем оставшимся переменным:

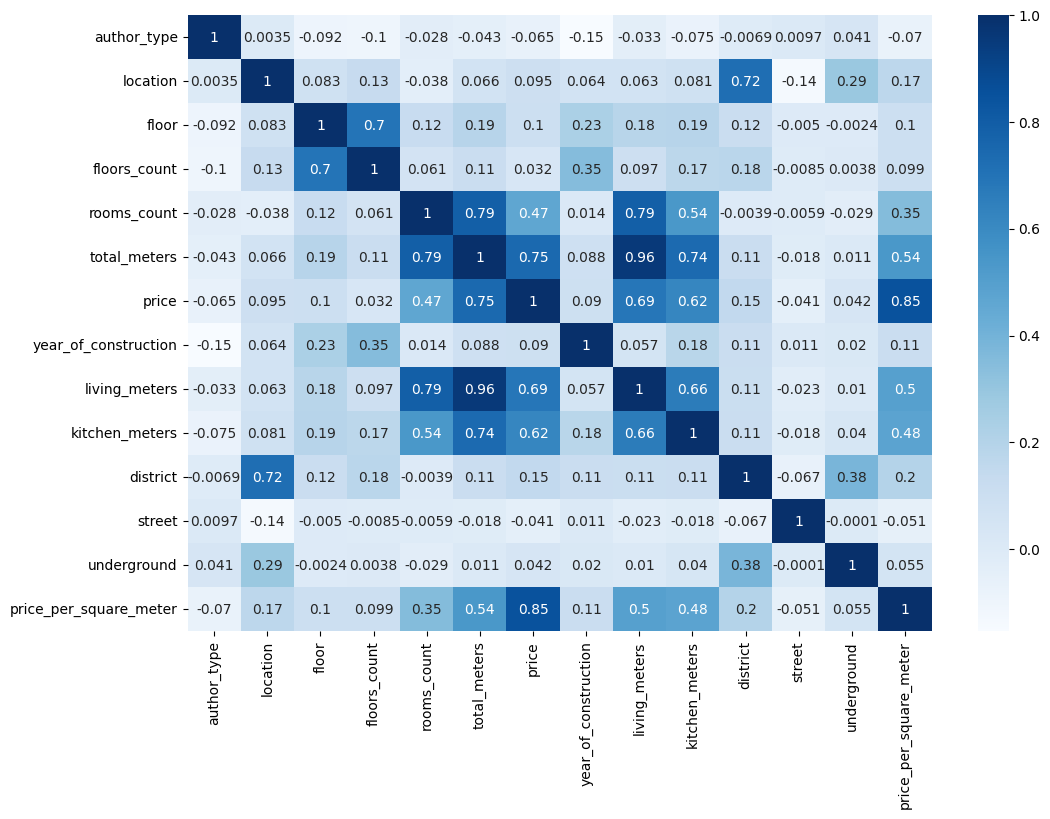




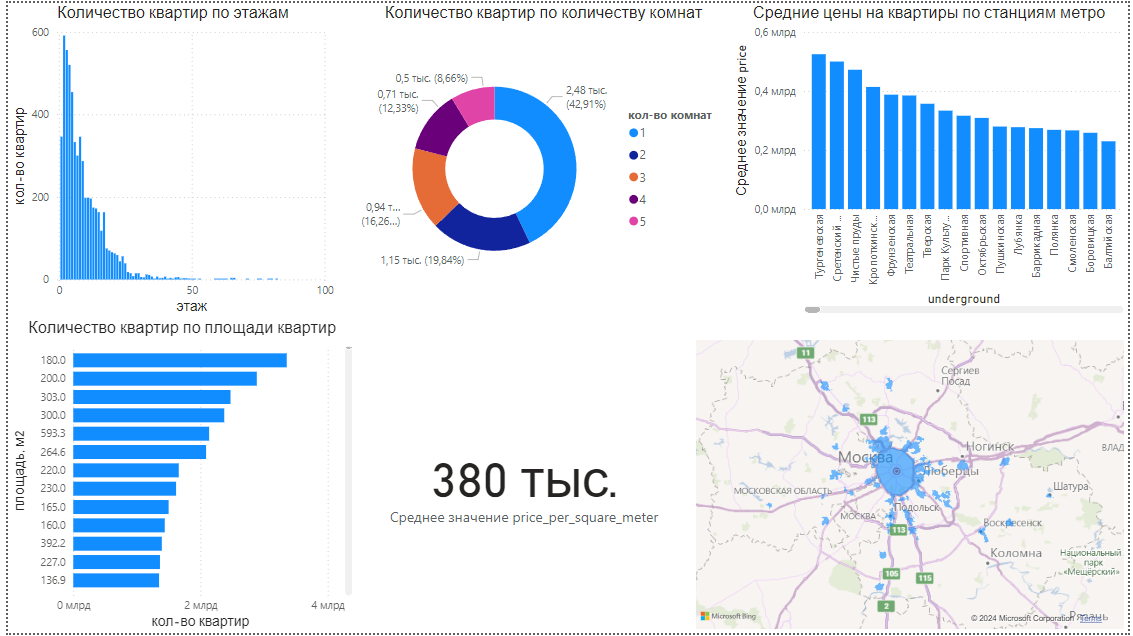


Посмотрев на эти графики видно, что распределения приближены к нормальным, но имеются выбросы, с которыми нужно будет что-то делать.

Также вывожу тепловую карту, которая будет показывать корреляцию между переменными:



Дополнительно к этой визуализации создал дашборд в приложении PowerBI, где можно интерактивно взаимодействовать с графиками и получить более детальную информацию в зависимости от города, метро, площади, комнат и этажа:



# ВЫВОД

Можно сделать вывод о корреляции колонок: больше всего на цену на квадратный метр влияет метраж квартиры (общий, жилой и кухня), количество комнат, район, город, этаж, количество этажей и год постройки.

Улица, метро и кто продаёт квартиру меньше всего влияют на целевую переменную.