

Аналитический отчет Парсинг

данных:

```
1 import cianparser
2 from time import sleep
3 a=0
4 while a < 52:
5     data = cianparser.CianParser(location="Москва").get_flats(deal_type="sale", rooms=(3,4), with_saving_csv=True, with_extra_data=True, additional_settings={"start_page":1+a,
6     "end_page":2+a})
7     sleep(120)
8     a+=2
```

Собираем все данные о квартирах с помощью библиотеки `cian`. После того как мы спарсим нужное кол-во строк, нам нужно произвести подготовку данных для анализа: проверка на пропуски, выбросы и ошибки. Обработать выявленные аномалии (удалить / заполнить)

1) Я решил удалить столбцы `author`, `author_type`, `url`, `deal_type` и `accommodation_type` из нашего CSV файла с данными о квартирах по следующим причинам:

А) Необходимость в релевантной информации: Основная цель анализа данных о квартирах заключается в оценке их стоимости и сравнении цен за квадратный метр. Столбцы, такие как `author` и `author_type`, содержат информацию о том, кто разместил объявление, что не влияет на фактическую стоимость недвижимости. Эти данные не имеют отношения к характеристикам квартиры, которые могут повлиять на цену.

Б) Фокус на ценовых показателях: Для анализа рынка недвижимости важнее сосредоточиться на факторах, которые непосредственно влияют на цену, таких как площадь квартиры, местоположение и другие характеристики. Столбцы `deal_type` и `accommodation_type` могут быть полезны в определенных контекстах, но для расчета цены за квадратный метр они не являются критически важными, так как цена может варьироваться в зависимости от других более значимых факторов.

```
3 # Загрузка данных из CSV файла
4 file_path = r'C:\Users\user\Desktop\intensiv\merged_file.csv' # Укажите путь к вашему CSV файлу
5 df = pd.read_csv(file_path)
6
7 # Удаление указанных столбцов
8 columns_to_drop = ['author', 'author_type', 'url', 'deal_type', 'accommodation_type']
9 df.drop(columns=columns_to_drop, inplace=True)
10
11 # Сохранение обновленного DataFrame обратно в CSV файл
12 df.to_csv('updated_apartments.csv', index=False)
13
```

процент пропусков в колонках.

location	0.000000
deal_type	0.000000
accommodation_type	0.000000
floor	0.000000
floors_count	0.000000
rooms_count	0.000000
total_meters	0.000000
price	0.487673
district	9.428339
street	14.020591
house_number	15.253319
underground	17.691682
residential_complex	25.995665
year_of_construction	25.020320
object_type	25.020320
house_material_type	25.020320
heating_type	25.020320
finish_type	25.020320
living_meters	25.020320
kitchen_meters	25.020320
price_per_meter	0.487673
dtype: float64	

2) Анализируем какие колонки имеют пропуски, видим, что пропусков достаточно много. Я решил удалить строки, в которых отсутствует год постройки.

```

1  # Удаление строк, где отсутствует год постройки
2  df_cleaned = df.dropna(subset=['year_of_construction'])
3
4  # Сохранение очищенного DataFrame обратно в CSV файл
5  df_cleaned.to_csv('cleaned_apartments.csv', index=False)
6

```

3) После удаления строк, снова смотрим пустые значения.

Видим, что в колонке метро есть много пропусков. Буду заполнять колонку underground. Часть процентов я заполнил модой, другую медианой. Колонку house_number заполним медианой, так как этот фактор не сильно влияет на цену квадратного метра.

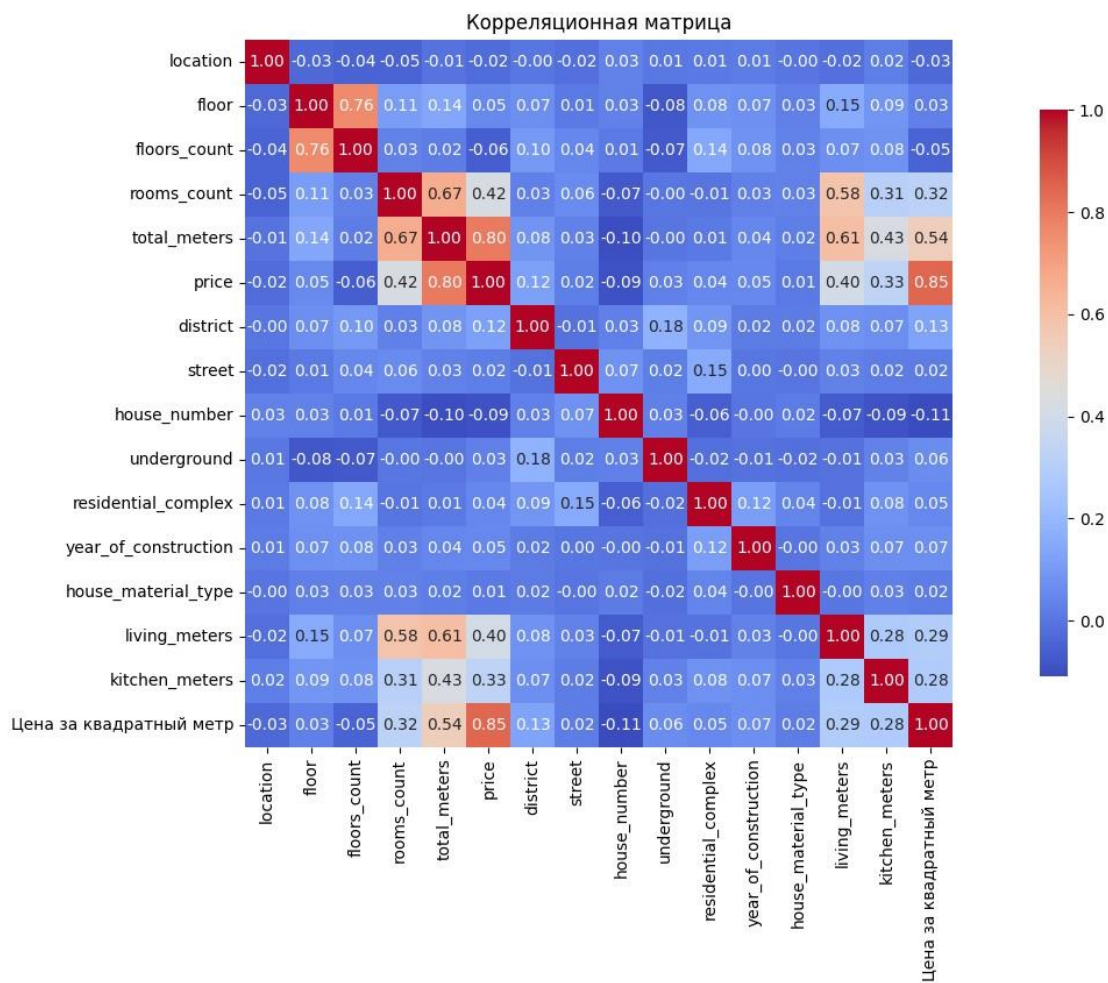
Так же нужно заполнить значения -1 в колонках. Заполнил разные колонки по разному. Колонку с годом постройки, определил медиану и по ней заполнил. Квадратные метры кухни и жилой площади заполнял по среднему. Смотрел сколько квадратных метров квартира, и в процентном соотношении заполнил аномалии

location	0.000000
deal_type	0.000000
accommodation_type	0.000000
floor	0.000000
floors_count	0.000000
rooms_count	0.000000
total_meters	0.000000
price	0.469738
district	11.833785
street	15.627823
house_number	17.127371
underground	23.197832
residential_complex	27.588076
year_of_construction	0.000000
object_type	0.000000
house_material_type	0.000000
heating_type	0.000000
finish_type	0.000000
living_meters	0.000000

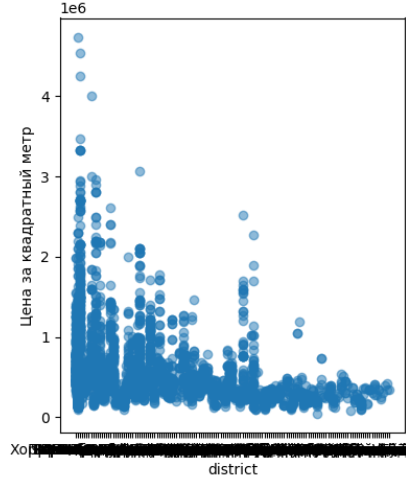
Вывод:

В процессе анализа рынка недвижимости мы пришли к важному выводу: цена за квадратный метр квартиры в значительной степени определяется тремя ключевыми факторами.

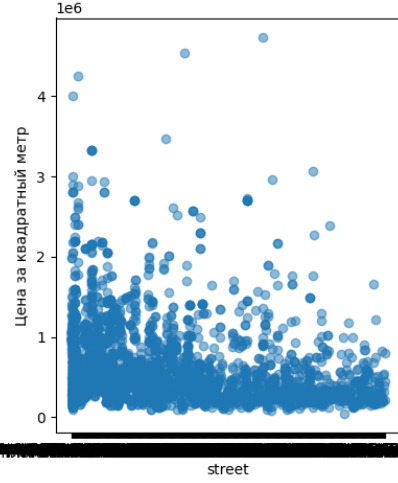
- 1) **Площадь квартиры**
- 2) **Количество комнат**
- 3) **Локация**



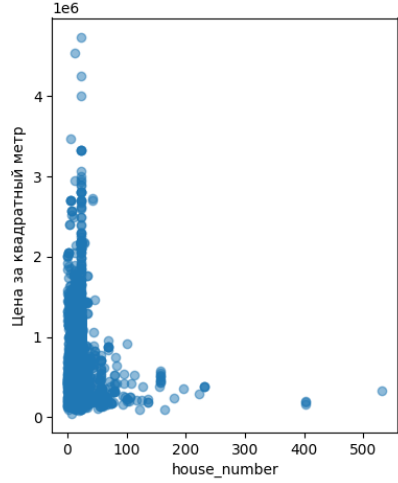
Зависимость цены за квадратный метр от district



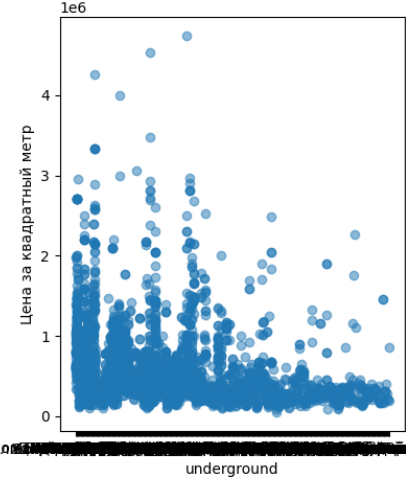
Зависимость цены за квадратный метр от street



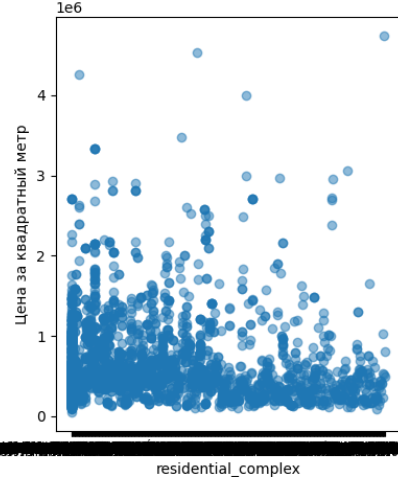
Зависимость цены за квадратный метр от house_number



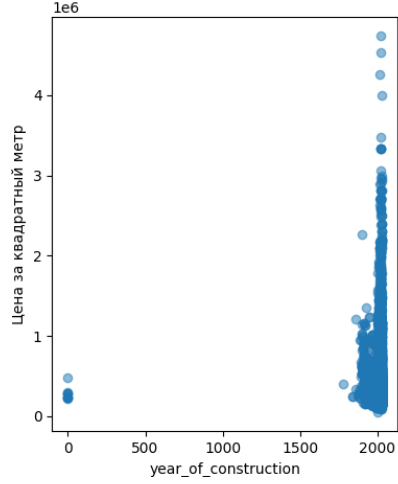
Зависимость цены за квадратный метр от underground



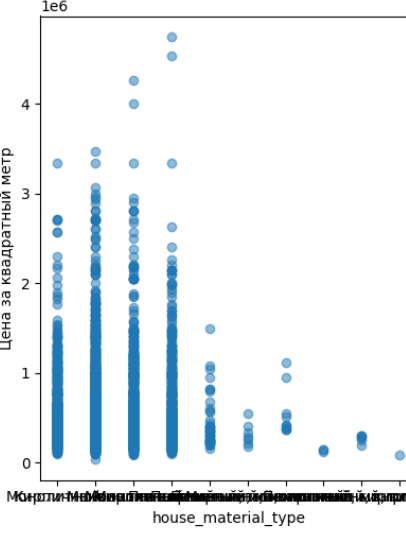
Зависимость цены за квадратный метр от residential_complex



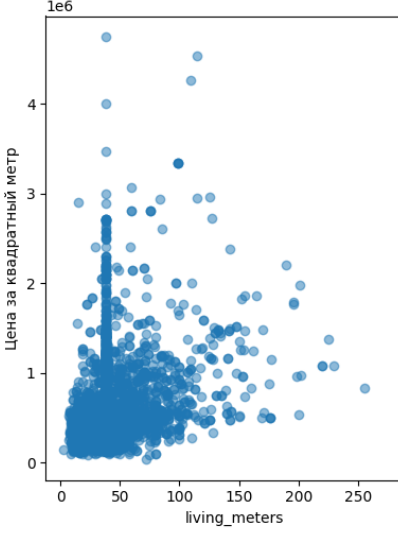
Зависимость цены за квадратный метр от year_of_construction



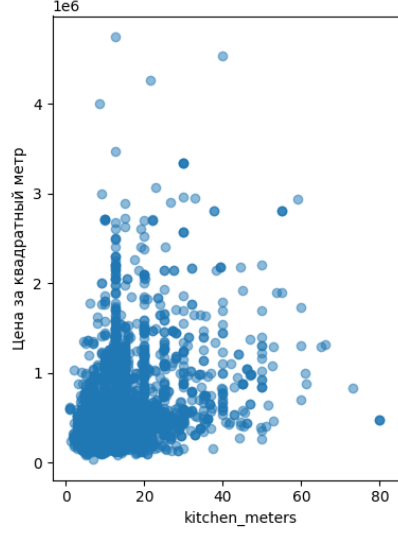
Зависимость цены за квадратный метр от house_material_type



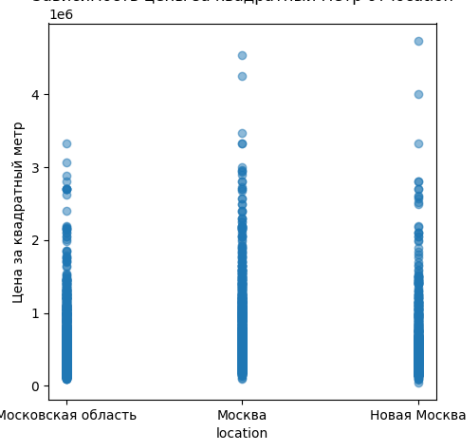
Зависимость цены за квадратный метр от living_meters



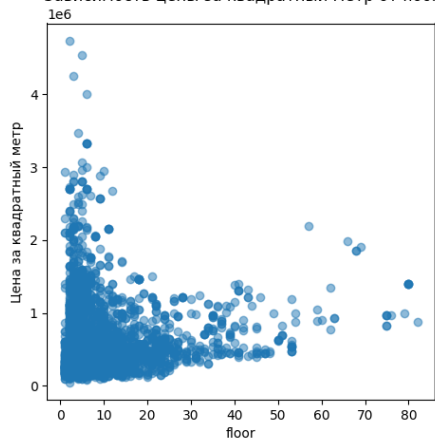
Зависимость цены за квадратный метр от kitchen_meters



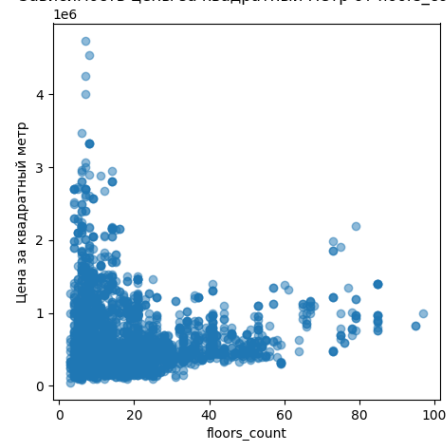
Зависимость цены за квадратный метр от location



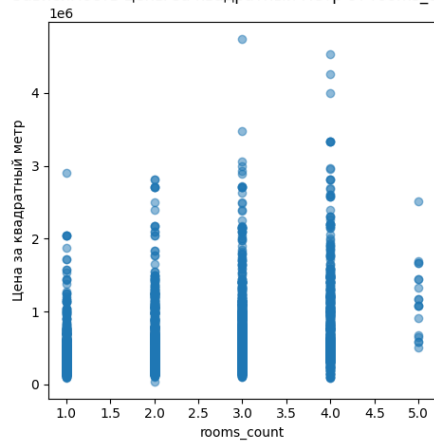
Зависимость цены за квадратный метр от floor



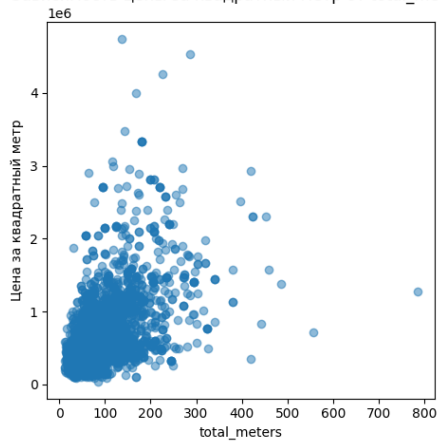
Зависимость цены за квадратный метр от floors_count



Зависимость цены за квадратный метр от rooms_count



Зависимость цены за квадратный метр от total_meters



Зависимость цены за квадратный метр от price

