Государственное бюджетное профессиональное
образовательное учреждение Московской области
«Физико-технический колледж»

Аналитический отчет

Работу выполнил:
Студент группы ИСП-21
Затыка Артём
Проверил:
преподаватель информатики
Базяк Г.В.

Долгопрудный, 2024

ВВЕДЕНИЕ

Цель отчета: Собрать данные и проанализировать, какие факторы влияют на цену квартир, используя доступные данные (регион, этажность, площадь, стоимость, наличие метро и т.д.) взятые с сети Интернет.

Актуальность: Рынок недвижимости — это важная часть экономики, и понимание факторов, влияющих на цены, помогает более точно оценивать объекты недвижимости и прогнозировать тенденции.

Основные вопросы:

- Какие переменные оказывают наибольшее влияние на цену квартиры?
- Насколько важны такие факторы, как наличие метро, этажность здания и площадь кухни?

Задачи:

- Используя открытые источники и личный опыт, составить список параметров, значительно влияющих на цену квадратного метра жилой площади.
- С учётом выявленных выше факторов произвести парсинг данных по квартирам на продажу, используя парсер. Данные получаем, используя сайт с объявлениями о продаже недвижимости: Циан.
- Произвести подготовку данных для анализа: проверка на пропуски, выбросы и ошибки. Обработать выявленные аномалии (удалить / заполнить)
- Проведите Исследовательский Анализ Данных (EDA). Постройте распределение основных параметров; визуализируйте взаимосвязи между ними; определите признаки, оказывающие наиболее сильное влияние на целевую переменную.

МЕТОДОЛОГИЯ

В данном исследовании используются несколько инструментов для сбора, обработки и анализа данных, включая Python, Excel и Power BI. Эти инструменты позволяют эффективно обрабатывать данные, визуализировать ключевые зависимости и анализировать факторы, влияющие на стоимость недвижимости.

СБОР И ПОДГОТОВКА ДАННЫХ

1. **Сбор данных**: Данные были собраны с помощью библиотеки cianparser, которая позволяет автоматизировать сбор информации с платформы Циан.

```
def parsing():
    from cianparser import CianParser
    '''Парсинг данных для анализа'''
    locations = tuple('Москва', 'Балашиха', 'Лобня', 'Солнечногорск')
    for loc in locations:
        parser = CianParser(loc)
        parser.get_flats(deal_type="sale", rooms=1, with_saving_csv=True,
with_extra_data=True, additional_settings={"start_page":1, "end_page":54})
```

2. **Просмотр полученных данных:**

```
import pandas as pd

# выводим полученные данные
# выводим первые 5 строк
df = pd.read_csv('cian_data.csv', delimiter=';', encoding='utf-8')
pd.set_option('display.max_columns', None)
print(df.head())
```

```
                     author        author_type  \
0              Метражи group  real_estate_agent
1                      ЛэндЛ  real_estate_agent
2                     Stenoy          developer
3  Alliance Agency Real Estate  real_estate_agent
4                Зиля Карамова            realtor

                                  url location deal_type  \
0  https://www.cian.ru/sale/flat/308167237/   Москва      sale
1  https://www.cian.ru/sale/flat/302263383/   Москва      sale
2  https://www.cian.ru/sale/flat/300878920/   Москва      sale
3  https://www.cian.ru/sale/flat/298254403/   Москва      sale
4  https://www.cian.ru/sale/flat/263316279/   Москва      sale

  accommodation_type  floor  floors_count  rooms_count  total_meters  \
0               flat    5.0          7.0          1.0          34.6
1               flat   14.0         45.0          1.0          41.3
2               flat   10.0         12.0          1.0          34.4
3               flat    4.0         33.0          1.0          42.9
4               flat    1.0         16.0          1.0          37.7

       price  year_of_construction  object_type  house_material_type  \
0  9000000.0                1978.0         -1.0                   -1
1 12800000.0                2026.0         -1.0                   -1
2 19372498.0                2026.0         -1.0  Монолитно-кирпичный
3 23600000.0                  -1.0         -1.0                   -1
4  9450000.0                1982.0         -1.0                   -1

   heating_type                    finish_type living_meters kitchen_meters  \
0          -1.0                             -1        18 м²          8 м²
1          -1.0                   Без отделки        20 м²         11 м²
2          -1.0  Без отделки, черновая, чистовая       11,8 м²       11,7 м²
3          -1.0                             -1       22,9 м²         15 м²
4          -1.0                             -1            -1            -1

         phone              district            street house_number  \
0  7.985041e+10     Северное Измайлово     15-я Парковая           54
1  7.964559e+10          Нижегородский    Перовское шоссе          NaN
2  7.499716e+10         Преображенское  Электрозаводская           60
3  7.965188e+10           Останкинский         Годовикова         11к2
4  7.916094e+10  Чертаново Центральное   Варшавское шоссе        142К2

             underground  residential_complex
0            Щёлковская                  NaN
1          Нижегородская  Level Нижегородская
2  Преображенская площадь                ARTEL
3           Алексеевская                iLove
4              Пражская                  NaN
```

```
# выводим количество данных
print(df.shape)
```

```
(10583, 24)
```

```
# выводим типы данных
print(df.dtypes)
```

```
author                  object
author_type             object
url                     object
location                object
deal_type               object
accommodation_type      object
floor                  float64
floors_count           float64
rooms_count            float64
total_meters           float64
price                  float64
year_of_construction   float64
object_type            float64
house_material_type     object
heating_type           float64
finish_type             object
living_meters           object
kitchen_meters          object
phone                  float64
district                object
street                  object
house_number            object
underground             object
residential_complex     object
dtype: object
```

3. **Предобработка данных в Python**:

1. Удалены строки с некорректными значениями, такие как будущие годы постройки (после 2024).

```
df[df['year_of_construction'].fillna(-1).astype(int) >= 2024].shape[0]
```

```
Количество значений больше 2024 в колонке year_of_construction: 1539
```

2. Приведены типы данных для ключевых столбцов: площади (квадратные метры кухни и жилья) были преобразованы в float, этаж, количество этажей и комнат, год постройки — в int.
3. Удалены столбцы с низким уровнем заполненности, такие как информация об отделке дома (house_material_type, object_type, finish_type) и отоплении(heating_type), а также столбцы с номером телефона(phone), типом сделки(deal_type) (всегда продажа), тип жилья(accommodation_type) (всегда квартира), так как эти столбики не дают какой-то дополнительной информации для анализа.

```
print ('Процент не заполненных строк в house_material_type:',
df[df['house_material_type'] == '-1'].shape[0] / df.shape[0] * 100)
print('Процент не заполненных строк в object_type:', df[df['object_type'] == -
1].shape[0] / df.shape[0] * 100)
print('Процент не заполненных строк в finish_type:', df[df['finish_type'] == '-
1'].shape[0] / df.shape[0] * 100)
print('Процент не заполненных строк в heating_type:', df[df['heating_type'] == -
1].shape[0] / df.shape[0] * 100)
```

```
Процент не заполненных строк в house_material_type: 87.88278526335627
Процент не заполненных строк в object_type: 99.67021577310845
Процент не заполненных строк в finish_type: 83.5107886554226
Процент не заполненных строк в heating_type: 99.67021577310845
```

4. Введен дополнительный столбец price_per_sqm для отображения стоимости квадратного метра.
5. Пропуски были заменены на -1 для удобства обработки, и вручную заполнены пропущенные значения в поле location при помощи Excel.

Код:

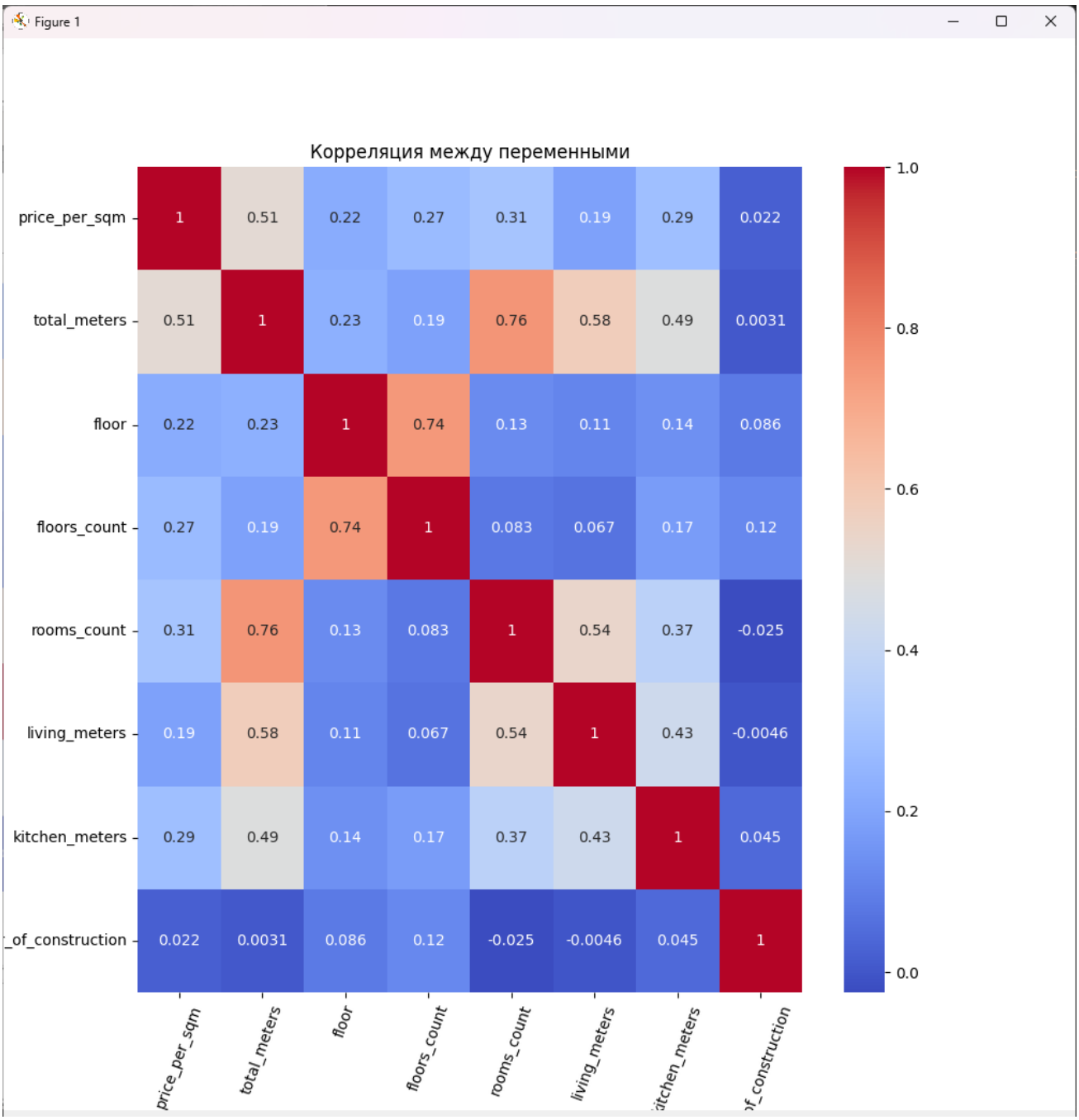
```
df = pd.read_csv('cian_data.csv', delimiter=';', encoding='utf-8')

df = df[df['year_of_construction'].fillna(-1).astype(int) <= 2024]  # удаляю
строки с годом больше 2024
df['living_meters'] = df['living_meters'].replace('м²', '').str.replace(',',
'.').astype(float)    # убираю лишние знаки и меняю тип данных
df['kitchen_meters'] = df['kitchen_meters'].replace('м²', '').str.replace(',',
'.').astype(float)
df['floor'] = df['floor'].fillna(-1).astype(int)    # заполняю пропущенными
значениями -1 и меняю тип данных на целые числа
df['floors_count'] = df['floors_count'].fillna(-1).astype(int)
df['rooms_count'] = df['rooms_count'].fillna(-1).astype(int)
df['year_of_construction'] = df['year_of_construction'].fillna(-1).astype(int)
df.drop(['phone', 'house_material_type', 'object_type', 'finish_type',
'heating_type', 'deal_type', 'accommodation_type'], axis=1, inplace=True)
df = df[df['price'].notna()]    # удалю строки с пропущенными значениями
df['price'] = df['price'].astype(int)
df['price_per_sqm'] = round(df['price'] / df['total_meters']).astype(int)   #
создаю и вычисляю цену за кв. метр

df.info()
print('Количество строк после обработки:', df.shape[0])
```

```
Index: 9536 entries, 0 to 10612
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   author                9471 non-null   object
 1   author_type           9468 non-null   object
 2   url                   9533 non-null   object
 3   location              9536 non-null   object
 4   floor                 9536 non-null   int64
 5   floors_count          9536 non-null   int64
 6   rooms_count           9536 non-null   int64
 7   total_meters          9536 non-null   float64
 8   price                 9536 non-null   int64
 9   year_of_construction  9536 non-null   int64
 10  living_meters         9536 non-null   float64
 11  kitchen_meters        9536 non-null   float64
 12  district              5619 non-null   object
 13  street                8757 non-null   object
 14  house_number          8891 non-null   object
 15  underground           5859 non-null   object
 16  residential_complex   3726 non-null   object
 17  price_per_sqm         9536 non-null   int64
dtypes: float64(3), int64(6), object(9)
memory usage: 1.4+ MB
Количество строк после обработки: 9536
```

Анализ данных

1. Корреляционный анализ:

    - С помощью Python и библиотеки pandas будет рассчитан коэффициент корреляции между переменными, такими как общая площадь (total_meters), цена (price), этажность (floor, floors_count), количество комнат (rooms_count), год постройки(year_of_construction), площадь кухни и жилья(kitchen_meters, living_meters), чтобы определить, какие факторы имеют наибольшее влияние на стоимость недвижимости.



На основе этой корреляционной матрицы можно сделать несколько выводов:

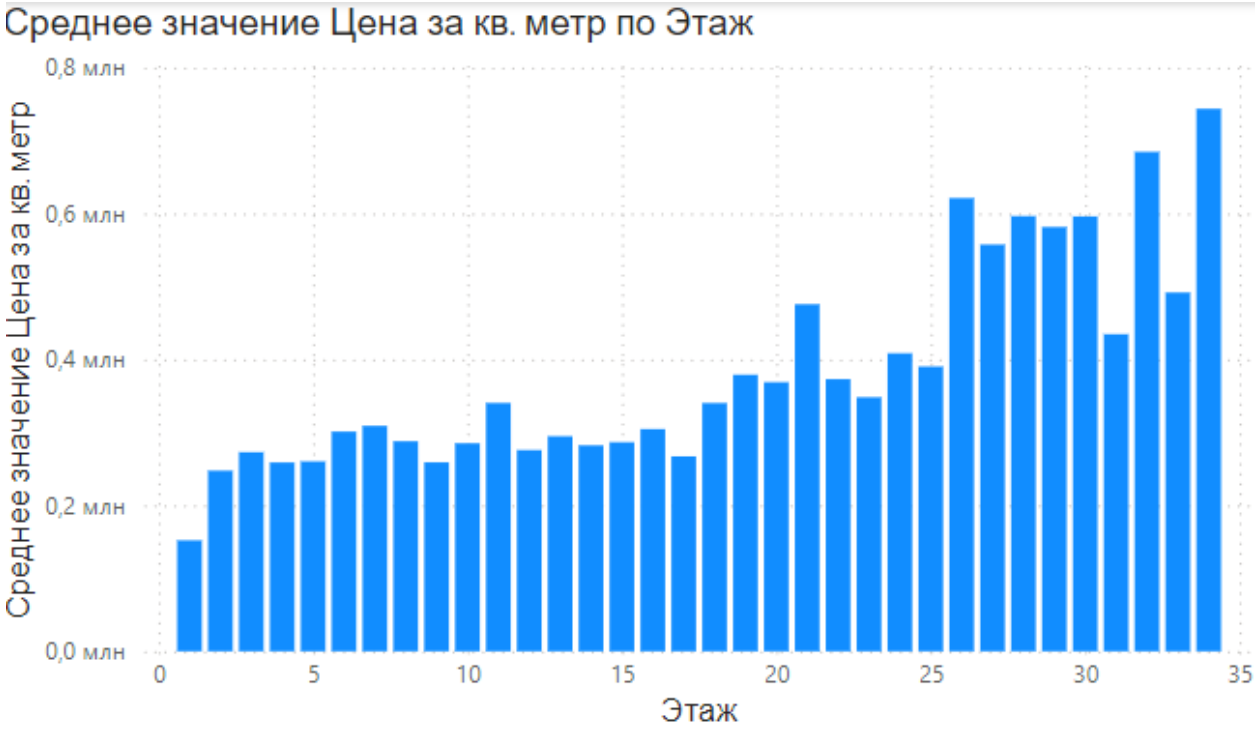
1. **Общая площадь (total_meters):**
    - Это наиболее значимый фактор, влияющий на цену за квадратный метр (корреляция 0.51). Более крупные квартиры обычно стоят дороже за квадратный метр.
2. **Этажность (floors_count):**
    - Имеет слабую положительную связь с ценой за квадратный метр (корреляция 0.27). Высотные здания могут означать более

современные комплексы или квартиры в центральных районах, что также может влиять на цену.

3. **Количество комнат (rooms_count):**
    - Корреляция с ценой за квадратный метр составляет 0.31, что указывает на слабую положительную связь. Количество комнат важно учитывать, так как потенциально влияет на спрос.
4. **Площадь кухни (kitchen_meters):**
    - Умеренная положительная корреляция с ценой за квадратный метр (0.29).
5. **Этаж (floor):**
    - Корреляция 0.22 с ценой за квадратный метр предполагает, что этаж, на котором расположена квартира, может играть роль в её стоимости, особенно если квартира находится на высоких этажах (особенно в высокоэтажных зданиях).

Общий вывод: среди всех факторов на цену за квадратный метр (price_per_sqm) заметное влияние оказывает **общая площадь(total_meters)** (коэффициент 0.51). Остальные факторы, такие как этажность и количество комнат, также оказывают влияние, но они не так ярко выражено.

2. Анализ влияния этажности и комнатности:
    - Данные будут разбиты на группы по этажности (floor, floors_count) и количеству комнат (rooms_count), чтобы выявить их влияние на стоимость за квадратный метр (price_per_sqm).



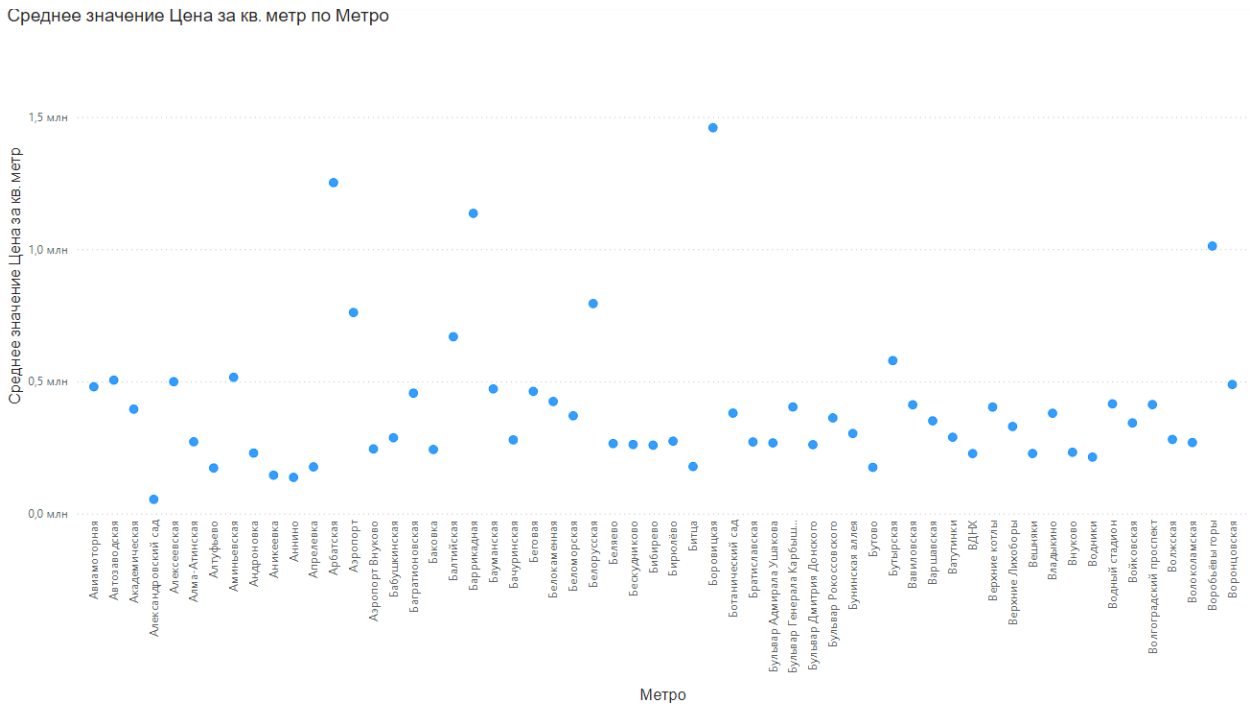
Первый график показывает, как средняя цена за квадратный метр изменяется в зависимости от количества комнат. Можно заметить, что

с увеличением количества комнат цена за квадратный метр возрастает, особенно это заметно для 4- и 5-комнатных квартир. Это может быть связано с более высоким классом жилья или более просторными квартирами с дополнительными удобствами.

Второй график демонстрирует распределение средней цены за квадратный метр по этажам. Видно, что на более высоких этажах (особенно от 20-го и выше) средняя цена за квадратный метр в целом растет.

3. Анализ влияния близости метро:

Квартиры, находящиеся рядом с метро, имеют значительно более высокую цену за квадратный метр. Это указывает на высокий спрос на такие объекты, поскольку удобный доступ к транспорту является важным фактором для покупателей. (*См. Диаграмма средняя цена по метро стр. 2*)



ЗАКЛЮЧЕНИЕ

Основываясь на проведённом анализе, можно выделить следующие ключевые выводы:

1. **Влияние транспортной доступности**: квартиры, расположенные рядом со станциями метро, демонстрируют значительно более высокую

цену за квадратный метр. Это подчёркивает важность транспортной доступности как одного из главных факторов, определяющих стоимость жилья в Москве.

2. **Этажность**: анализ зависимости стоимости от этажности показал, что квартиры на средних этажах в целом имеют более высокую стоимость. Это может быть связано с тем, что средние этажи считаются более удобными для проживания.
3. **Количество комнат и площадь**: цена за квадратный метр имеет тенденцию к увеличению в зависимости от количества комнат, особенно для квартир с четырьмя и пятью комнатами. Это говорит о том, что более крупные объекты более востребованы среди покупателей, несмотря на их высокую стоимость.
4. **Год постройки**: на стоимость квадратного метра также оказывает влияние возраст здания, однако корреляция достаточно слабая. Возможно, это связано с тем, что более новые здания часто имеют улучшенные условия или расположены в престижных районах.
5. **Другие факторы**: несмотря на выявленные зависимости, такие факторы, как наличие инфраструктуры, вид из окна, качество отделки, также могут оказывать влияние на цену, однако их необходимо исследовать отдельно, так как их вклад может быть неоднозначным и индивидуально различаться по объектам. К тому же эти данные могли бы участвовать в анализе, однако инструмент сбора данных не позволял сохранить значения в большинстве случаев.

Итог: Проведённый анализ показывает, что ключевые факторы, такие как транспортная доступность, количество комнат и этажность, играют основную роль в формировании цены на недвижимость.