Аналитический отчет по анализу данных недвижимости

Автор: Грушковский Захар Романович
Группа: ИСП-23в

1 ноября 2024

1. Введение

1.1 Цели исследования

Целью данной работы является анализ данных о недвижимости, полученных с Циан, для выявления факторов, влияющих на стоимость квадратного метра квартиры, и подготовки данных для дальнейшего использования в построении моделей машинного обучения.

1.2 Задачи:

1. Получить и очистить данные о недвижимости.
2. Провести анализ числовых и категориальных переменных.
3. Заполнить пропущенные данные и подготовить датасет для визуализации и корреляционного анализа.
4. Построить визуализации для выявления ключевых закономерностей.
5. Создать модель машинного обучения для предсказания ключевого параметра
6. Сформировать выводы и рекомендации для использования данных в дальнейшем.

2. Методология и инструменты

Для выполнения поставленных задач использовались следующие инструменты и библиотеки:

- Python для обработки данных и автоматизации запросов.
- Библиотеки pandas, numpy для анализа и подготовки данных.
- Визуализационные библиотеки: seaborn и matplotlib для построения графиков и тепловой карты корреляции.

Источником данных является Циан.

3. Этапы работы

3.1 Парсинг через Python

Для загрузки данных был создан парсер на основе cianparser.

3.2 Предварительная обработка данных

После парсинга был выполнен следующий процесс:

- Создан DataFrame с нужными колонками:

```
floor = этаж
floors_count = кол-во этажей
rooms_count = кол-во комнат
total_meters = кол-во м²
underground = наличие метро
price_per_meter = цена за м²
```

- Удаление дубликатов:

```python
data = pd.read_csv(base)
data = data.drop_duplicates()
data.to_csv(base, index=False)
```

3.3 Выявление столбцов с пропущенными значениями

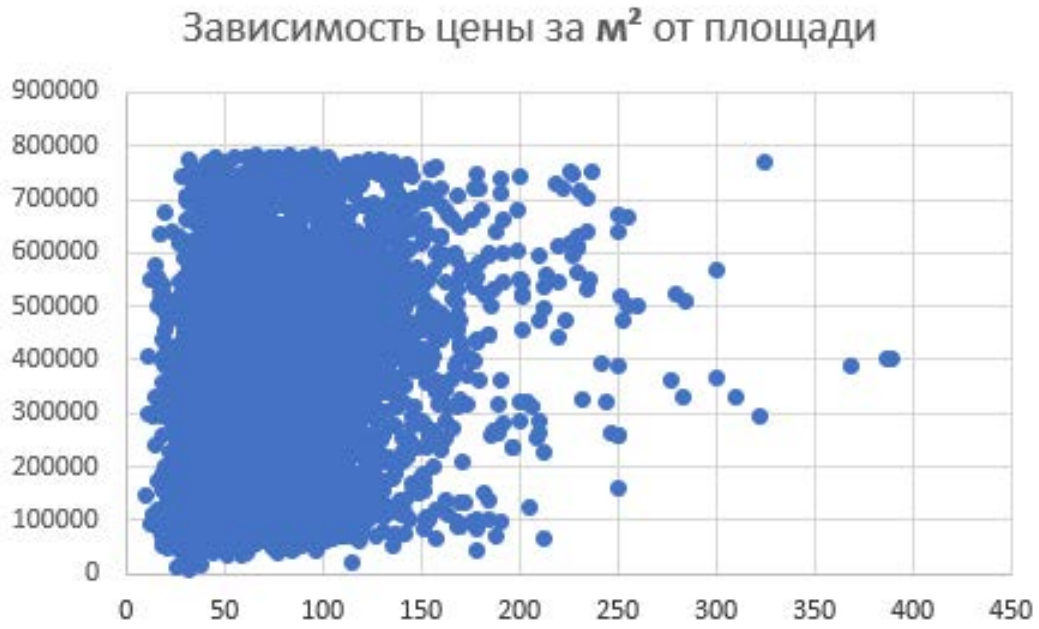
Проверка на пропущенные значения и их удаление было выполнено с помощью кода:

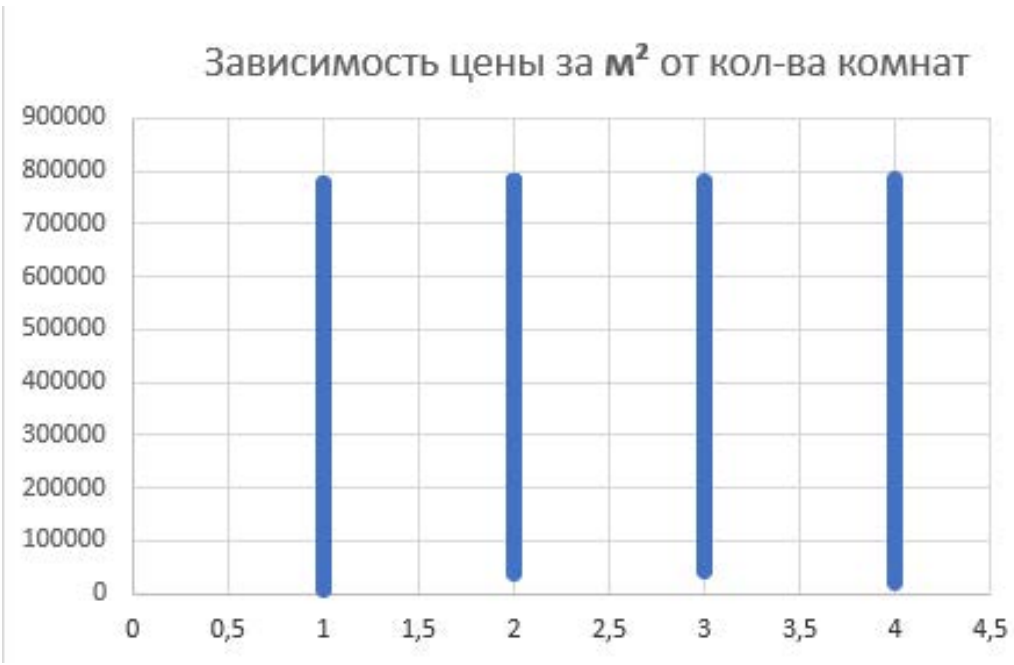
```python
df = pd.read_csv(base)
df = df.dropna(subset=['floor','floors_count','rooms_count','total_meters','price'])
df.to_csv(base, index=False)
```

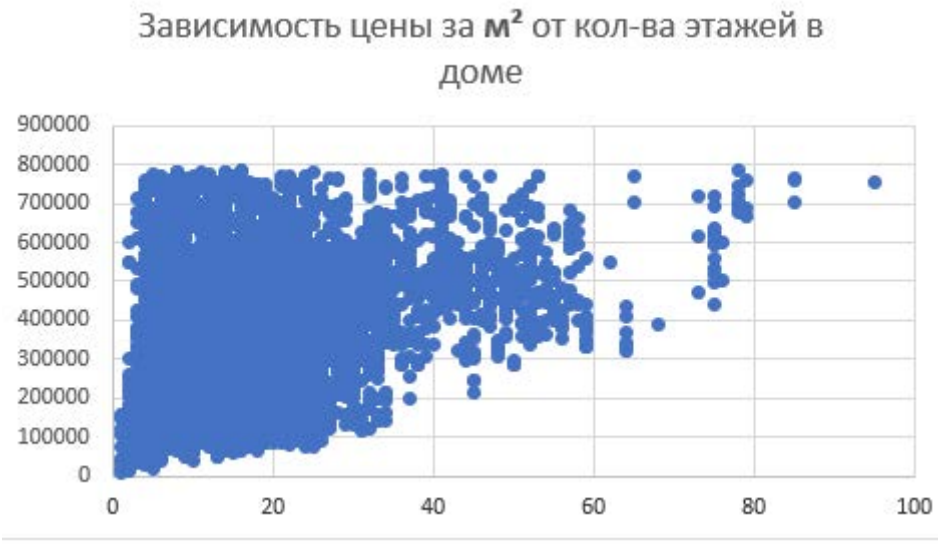
3.4 Визуализация данных

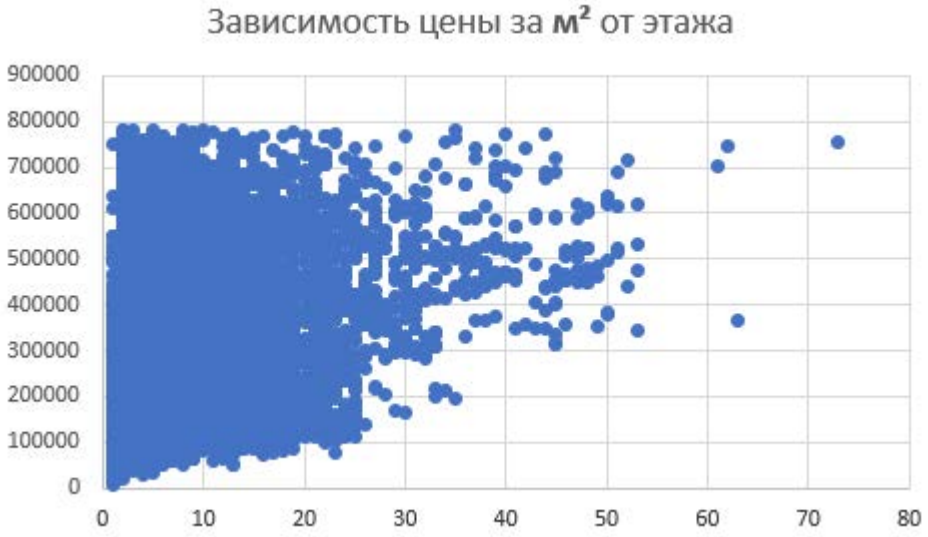
Для анализа взаимосвязи между ценой за квадратный метр и другими признаками были построены следующие графики:

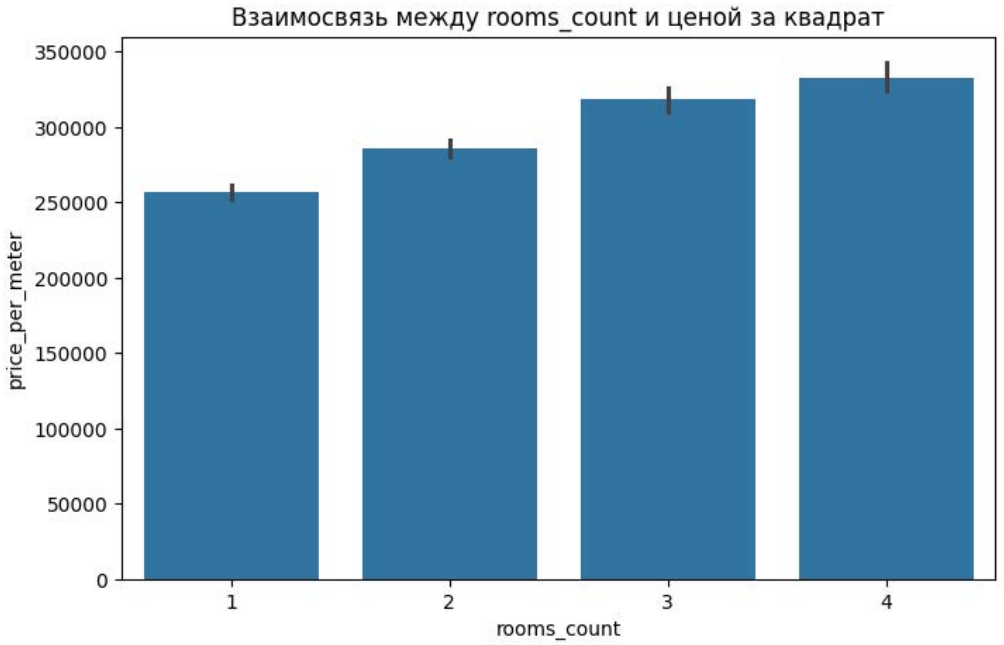
- Диаграммы рассеяния:

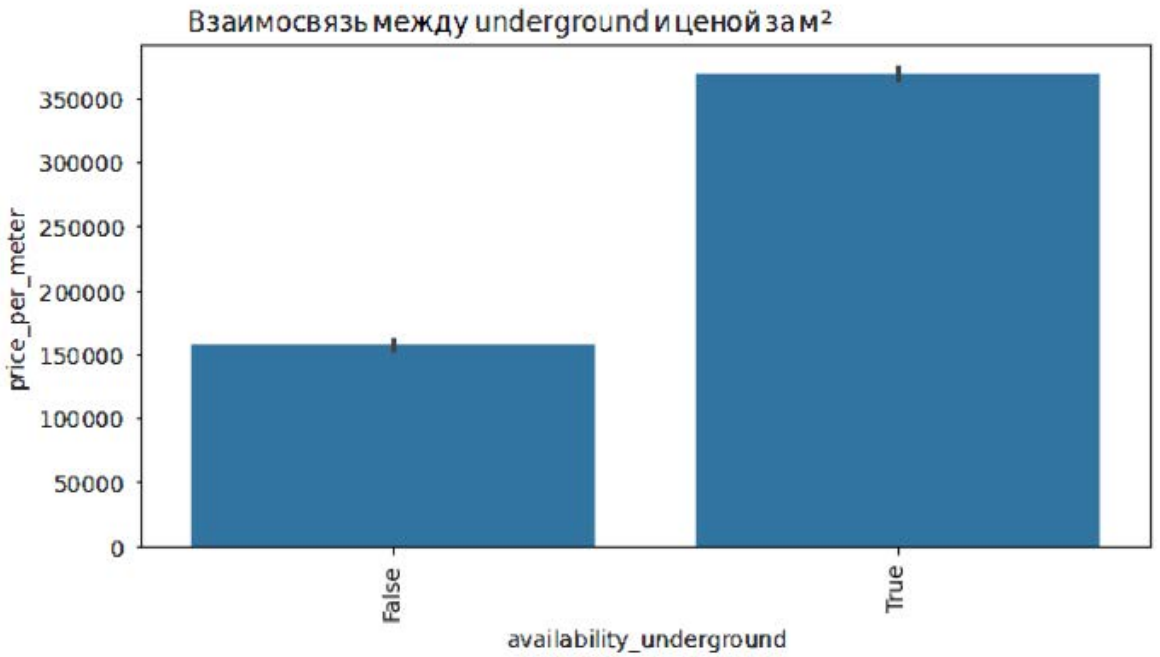




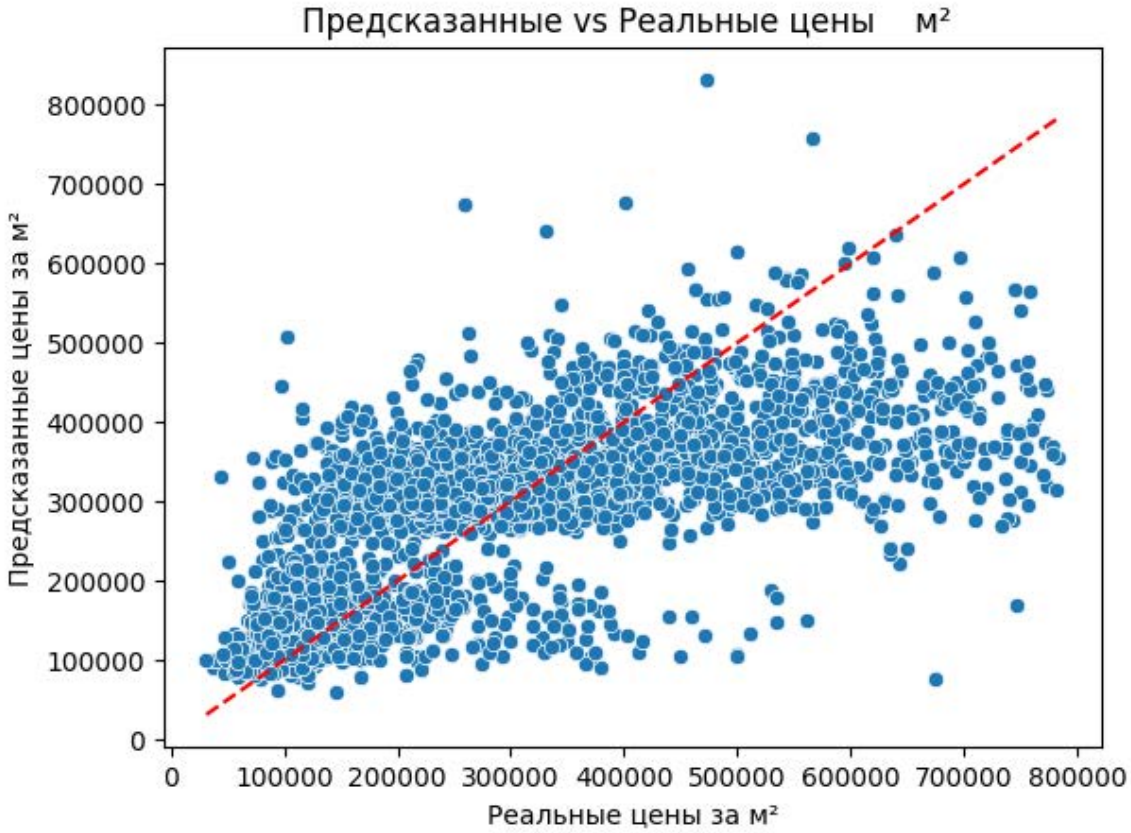


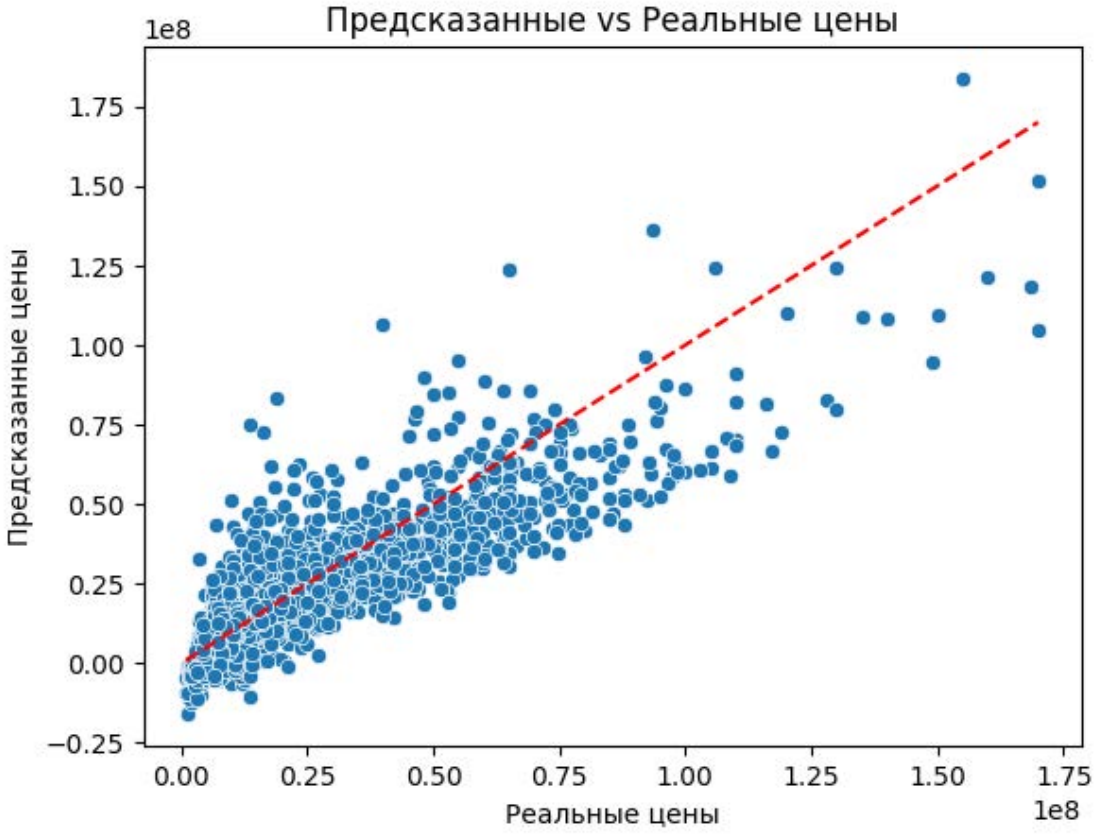

Зависимость цены за м² от этажа
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
0
10
20
30
40
50
60
70
80


Взаимосвязь между rooms_count и ценой за квадрат
price_per_meter
350000
300000
250000
200000
150000
100000
50000
0
1
2
3
4
rooms_count

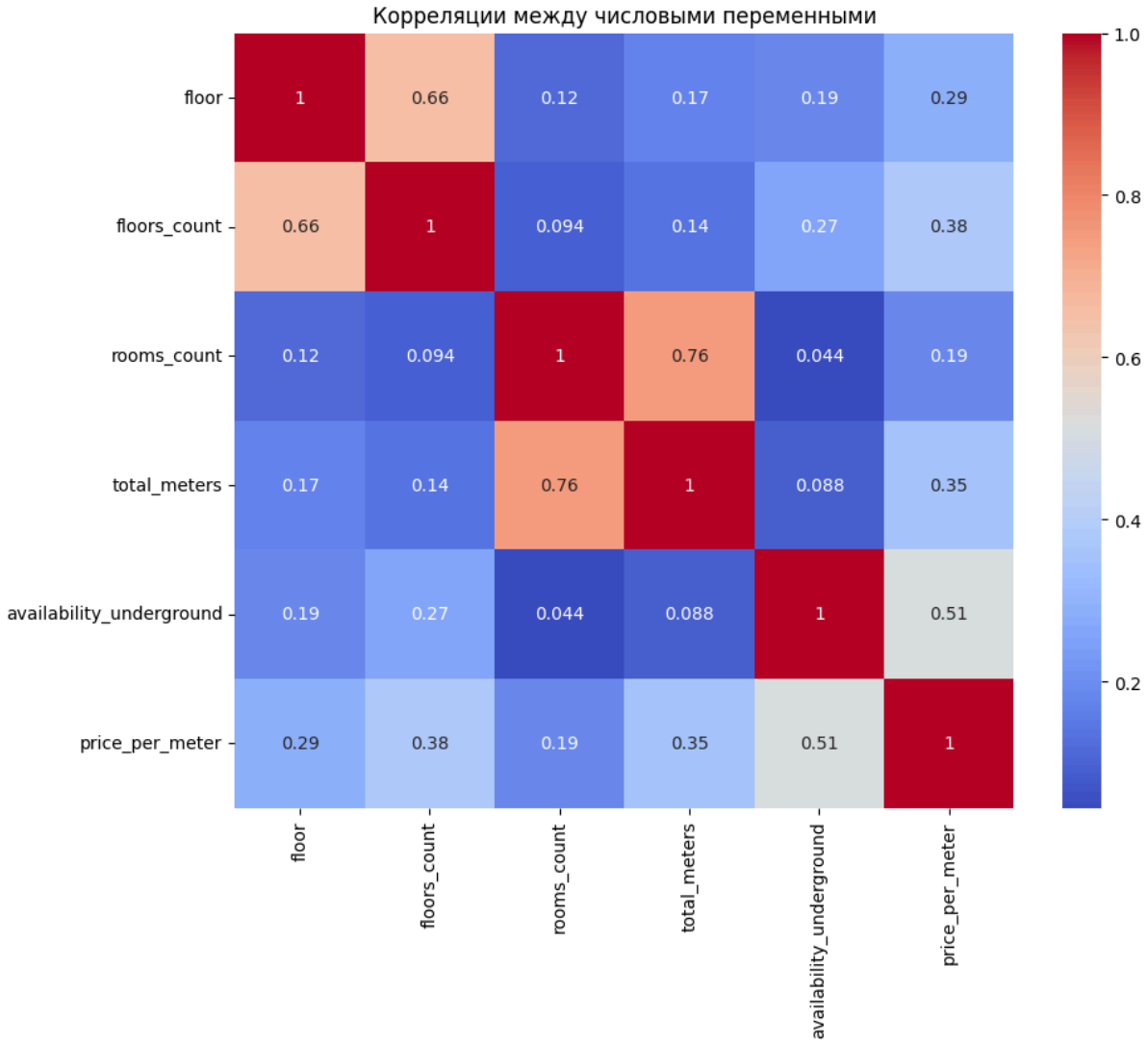

Взаимосвязь между underground и ценой за м²
price_per_meter
350000
300000
250000
200000
150000
100000
50000
0
False
True
availability_underground

Графическое представление предсказаний модели машинного обучения:





- Тепловая карта корреляции, показывающая степень взаимосвязи:



4. Результаты и выводы

4.1 Анализ корреляции

Тепловая карта корреляции показала следующие ключевые зависимости:

- Цена за квадратный метр (`price_per_meter`) наиболее сильно коррелирует с наличием метро (`underground`) и кол-во этажей (`floors_count`).

- Количество комнат (`rooms_count`) показало слабую корреляцию с ценой за квадратный метр, что говорит о меньшем влиянии этого параметра на стоимость в сравнении с общей площадью и общей ценой.

6. Заключение

В ходе работы был проведен анализ и очистка данных о рынке недвижимости, полученных из Циан. Выполненная обработка позволила выявить ключевые зависимости между параметрами объектов и подготовить данные для дальнейшего использования в построении моделей предсказания цен. Данные готовы к применению для задач машинного обучения и мониторинга изменений рынка.

В последующих работах хотелось бы больше внимания уделить построению моделей машинного обучения