Аналитический отчет по анализу данных

Автор: Грибов Виталий Владимирович

Группа: ИСП-23В

1. Введение

1.1 Цель исследования

Целью данной работы является анализ данных полученных с помощью парсинга данных с сайта Циан , для выявления факторов, влияющих на стоимость квадратного метра, и
подготовки данных для использования в построении моделей машинного обучения.

1.2 Задачи:

1. Получить и очистить данные.
2. Провести анализ числовых и категориальных переменных.
3. Заполнить пропущенные данные и подготовить DataSet для визуализации и корреляционного анализа.
4. Построить визуализации.
5. Сформировать выводы и рекомендации.

---

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 сайт Циан.

---

3. Этапы работы

3.1 Загрузка данных через Cianparser

Для загрузки данных был использован cianparser, который выполняет автоматизированные запросы к сайту Циан с необходимыми параметрами

3.2 Предварительная обработка данных

После загрузки данных был выполнен следующий процесс:

- Создан DataFrame с нужными колонками: price, total_meters, rooms, price_per_metr, underground, district.
- Обнаружены пропущенные значения в колонке district и underground, которые были обработаны.

3.3 Выявление столбцов с пропущенными значениями

Проверка на пропущенные значения была выполнена с помощью кода:

В результате были обнаружены пропуски в колонках underground, district , которые были заполнены.

3.4 Визуализация данных

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

- **Гистограммы** для колонок floors, total_meters, rooms относительно price_per_metr.
- **Тепловая карта корреляции**, показывающая степень взаимосвязи между числовыми переменными.

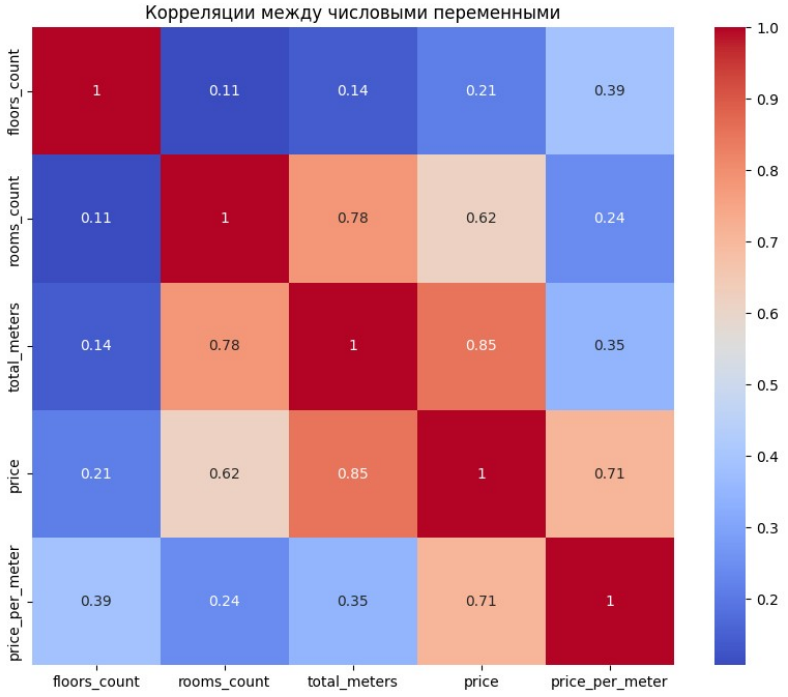
---

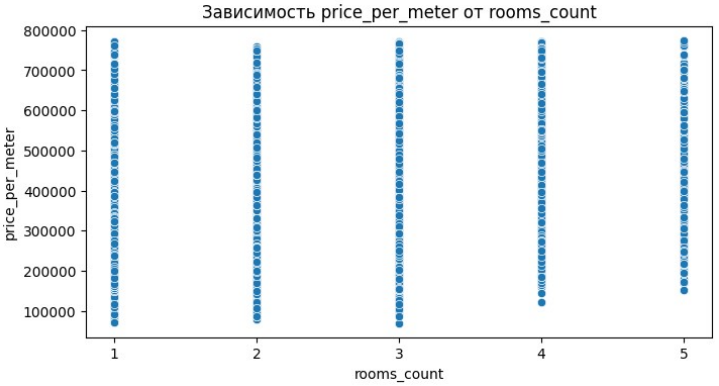
4. Результаты и выводы

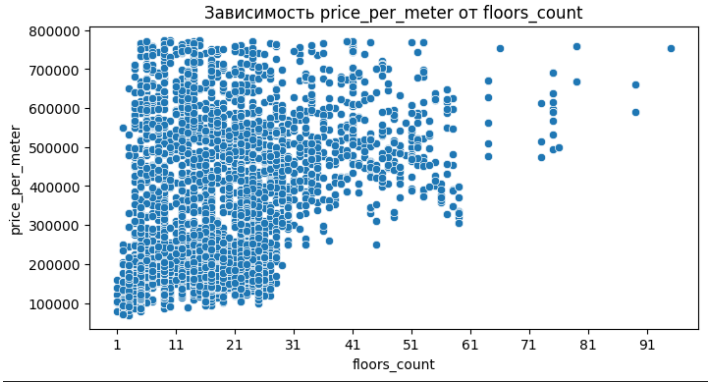
4.1 Анализ корреляции

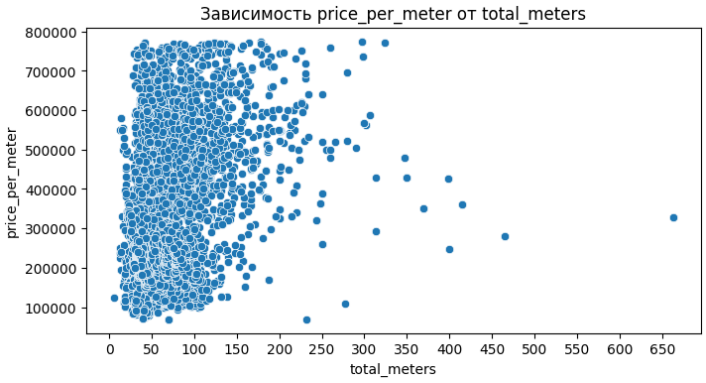
Тепловая карта корреляции показала следующие ключевые зависимости:

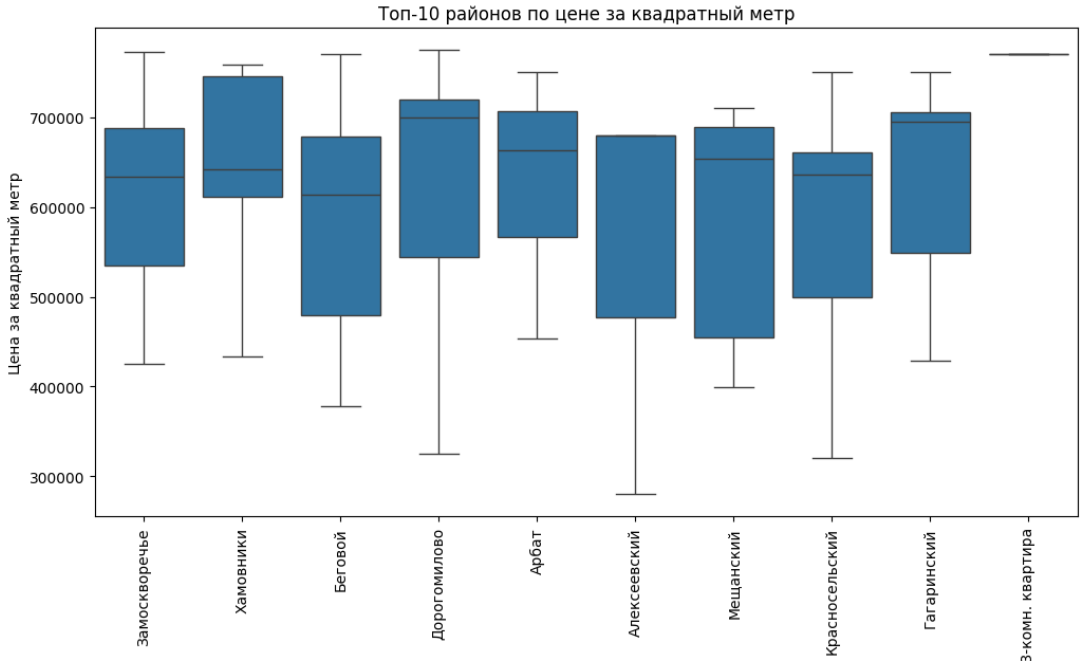
- **Цена за квадратный метр (price_per_metr)** наиболее сильно коррелирует с общей ценой (price) и площадью квартиры (total_meters).
- Количество этажей (floors_count) показало слабую корреляцию с ценой за квадратный метр, что говорит о меньшем влиянии этого параметра на стоимость.


Корреляции между числовыми переменными
floors_count
rooms_count
total_meters
price
price_per_meter
1
0.11
0.14
0.21
0.39
0.11
1
0.78
0.62
0.24
0.14
0.78
1
0.85
0.35
0.21
0.62
0.85
1
0.71
0.39
0.24
0.35
0.71
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Зависимость price_per_meter от rooms_count
price_per_meter
rooms_count
800000
700000
600000
500000
400000
300000
200000
100000
1
2
3
4
5


Зависимость price_per_meter от floors_count
price_per_meter
floors_count
800000
700000
600000
500000
400000
300000
200000
100000
1
11
21
31
41
51
61
71
81
91


Зависимость price_per_meter от total_meters
price_per_meter
total_meters
800000
700000
600000
500000
400000
300000
200000
100000
0
50
100
150
200
250
300
350
400
450
500
550
600
650


Топ-10 районов по цене за квадратный метр
Цена за квадратный метр
700000
600000
500000
400000
300000
Замоскворечье
Хамовники
Беговой
Дорогомилово
Арбат
Алексеевский
Мещанский
Красносельский
Гагаринский
3-комн. квартира

5. Рекомендации

1. **Использование обработанных данных для построения модели**:

   Данные готовы для обучения модели машинного обучения, которая может предсказывать стоимость квартиры на основе признаков price, total_meters, rooms, District, floors.

2. **Регулярное обновление данных**:

   Для актуальности данных рекомендуется периодически обновлять их через cianparser, чтобы учесть изменения на рынке.

3. **Дальнейший анализ категориальных переменных**:

   Рекомендуется изучить влияние других категориальных признаков которые могут оказать влияние на цену.

---

6. Заключение

В ходе работы был проведен анализ и очистка данных.
Выполненная обработка позволила выявить ключевые зависимости между параметрами объектов и подготовить данные для дальнейшего использования в построении моделей предсказания цен.
