Аналитический отчет по анализу данных недвижимости

Автор: [Сулейманова Зарема]

Группа: [ИСП-23В]

Введение

Цель исследования

Цель данной работы — проанализировать данные о недвижимости, полученные через API DomClick. В результате анализа необходимо выявить факторы, влияющие на стоимость квадратного метра жилья, и подготовить данные для дальнейшего использования в моделях машинного обучения.

Задачи исследования

1. Сбор и предварительная обработка данных о недвижимости.
2. Анализ числовых и категориальных переменных.
3. Заполнение пропущенных значений и подготовка данных для визуализации и корреляционного анализа.
4. Построение визуализаций для выявления ключевых закономерностей.
5. Формулирование выводов и рекомендаций для дальнейшего использования данных.

Методология и инструменты

Для достижения поставленных задач были использованы следующие инструменты и библиотеки:

1. Python — основной язык программирования для обработки данных и автоматизации запросов.
2. Pandas и NumPy — библиотеки для анализа и подготовки данных.
3. Seaborn и Matplotlib — библиотеки визуализации для построения графиков и тепловых карт корреляции.
4. KNN Imputer — метод заполнения пропущенных значений на основе значений ближайших соседей.
5. API DomClick — источник данных.

Этапы работы

1. Загрузка данных через API DomClick.

Для загрузки данных был создан класс DomClickApi, который автоматизирует процесс отправки запросов к API DomClick с заданными параметрами, такими как тип недвижимости, регион и количество комнат. На каждом этапе работы выводились промежуточные результаты для проверки корректности полученной информации.

2. Предварительная обработка данных.

После получения данных из API был выполнен следующий процесс:

1. Создан DataFrame с необходимыми столбцами: price, area, rooms, square_price, subways, monthly_payment.
2. В столбцах subways и monthly_payment были обнаружены пропущенные значения, которые были обработаны с помощью метода KNN Imputer и других подходов.
3. Выявление столбцов с пропущенными значениями.

Проверка на наличие пропущенных значений была выполнена с помощью соответствующего кода. В результате были обнаружены пропуски в столбцах subways, monthly_payment и других, которые были заполнены с помощью KNN Imputer и дополнительных методов.

4. Визуализация данных.

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

1. Диаграммы рассеяния для столбцов price, area, rooms относительно square_price.

2. Тепловая карта корреляции, демонстрирующая степень взаимосвязи между числовыми переменными (см. прилагаемый график).

Результаты и выводы

1. **Анализ корреляции.**

Тепловая карта корреляции выявила несколько ключевых зависимостей:

1. Цена за квадратный метр (square_price) наиболее сильно коррелирует с общей ценой (price) и площадью квартиры (area).
2. Количество комнат (rooms) показало слабую корреляцию с ценой за квадратный метр, что указывает на меньшее влияние этого параметра на стоимость по сравнению с общей площадью и ценой.
3. Обработка пропущенных значений.

Применение KNN Imputer позволило эффективно заполнить пропуски в столбце subways на основе схожих значений соседних объектов. Этот метод обеспечил более точное восстановление данных по сравнению с простыми статистическими подходами, такими как использование среднего значения или медианы.

Рекомендации

1. **Применение обработанных данных для создания модели ценообразования.**

Данные готовы для обучения модели машинного обучения, способной предсказывать стоимость квартиры на основе таких признаков, как price, area, rooms и subways.

2. **Регулярное обновление данных через API.**

Для поддержания актуальности модели рекомендуется периодически обновлять данные через API DomClick, чтобы учитывать изменения на рынке недвижимости.

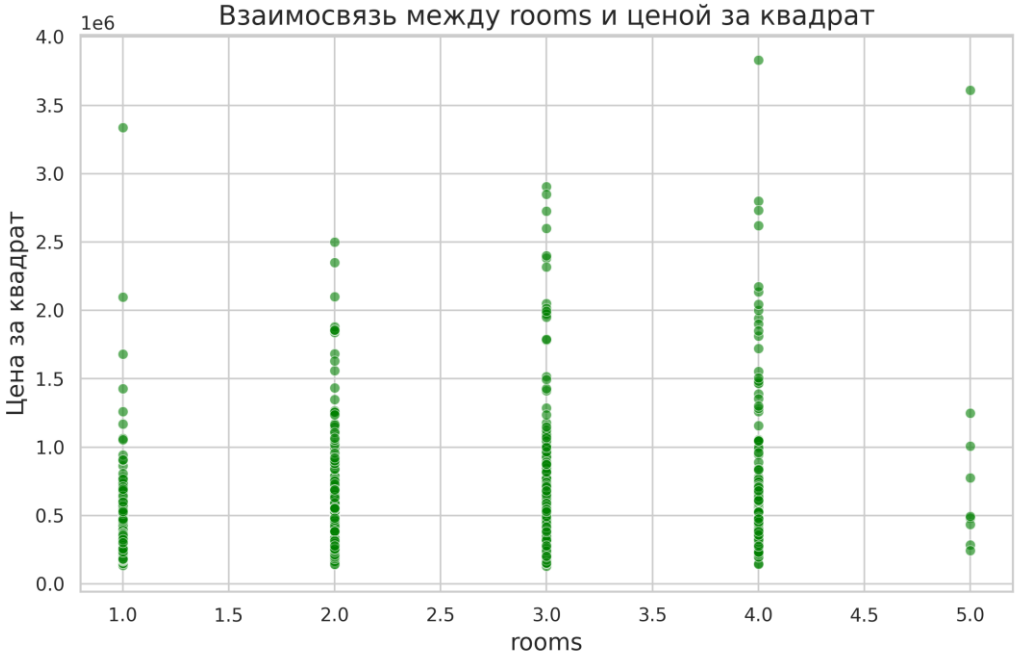
3. **Дальнейший анализ категориальных переменных.**

Рекомендуется провести более глубокое исследование влияния других категориальных признаков, таких как renovation и placement_type, которые могут существенно влиять на цену.

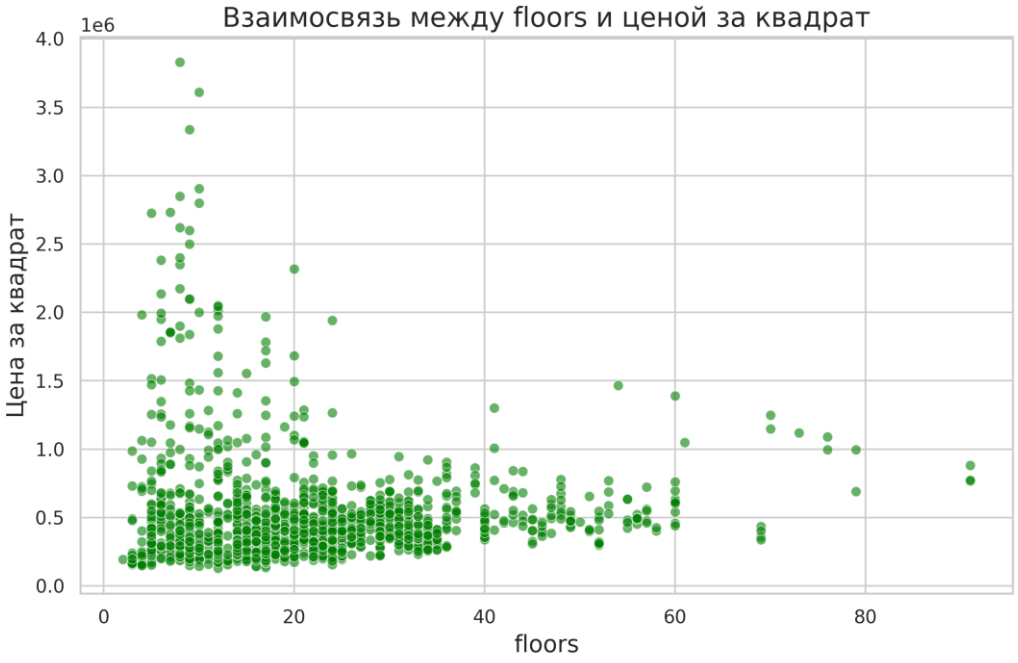
Заключение

В ходе работы был проведён анализ и очистка данных о рынке недвижимости, полученных из API DomClick. Проведённая обработка позволила выявить ключевые зависимости между параметрами объектов и подготовить данные для дальнейшего использования в моделях прогнозирования цен. Данные готовы к применению в задачах машинного обучения и для мониторинга изменений на рынке.

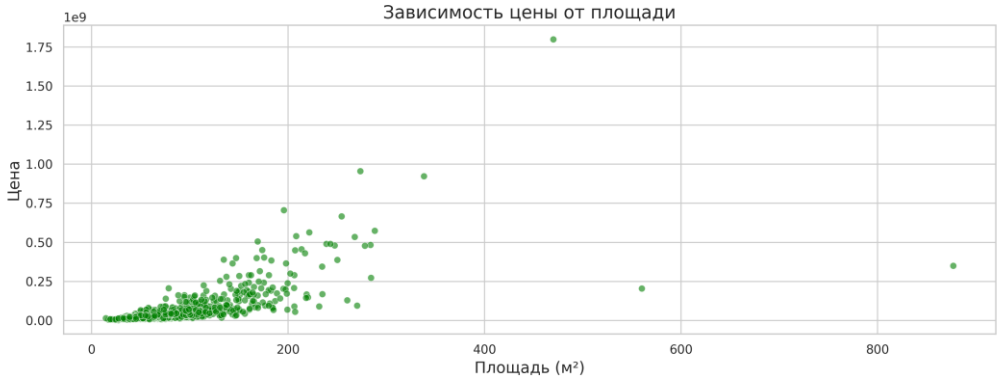

Взаимосвязь между monthly_payment и ценой за квадрат
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Цена за квадрат
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
monthly_payment

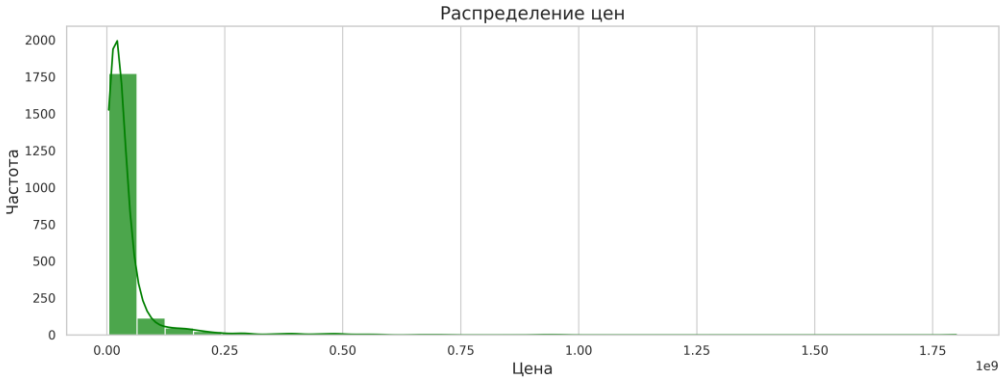

Взаимосвязь между rooms и ценой за квадрат
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Цена за квадрат
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
rooms

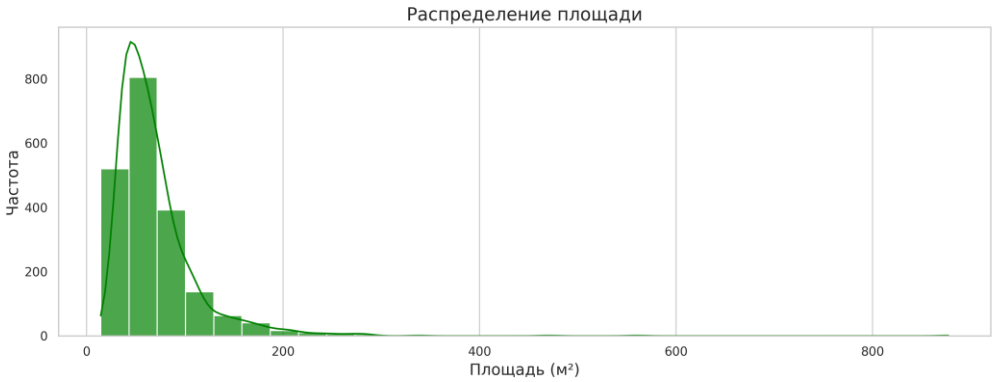

Взаимосвязь между price и ценой за квадрат
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Цена за квадрат
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
price
1e9

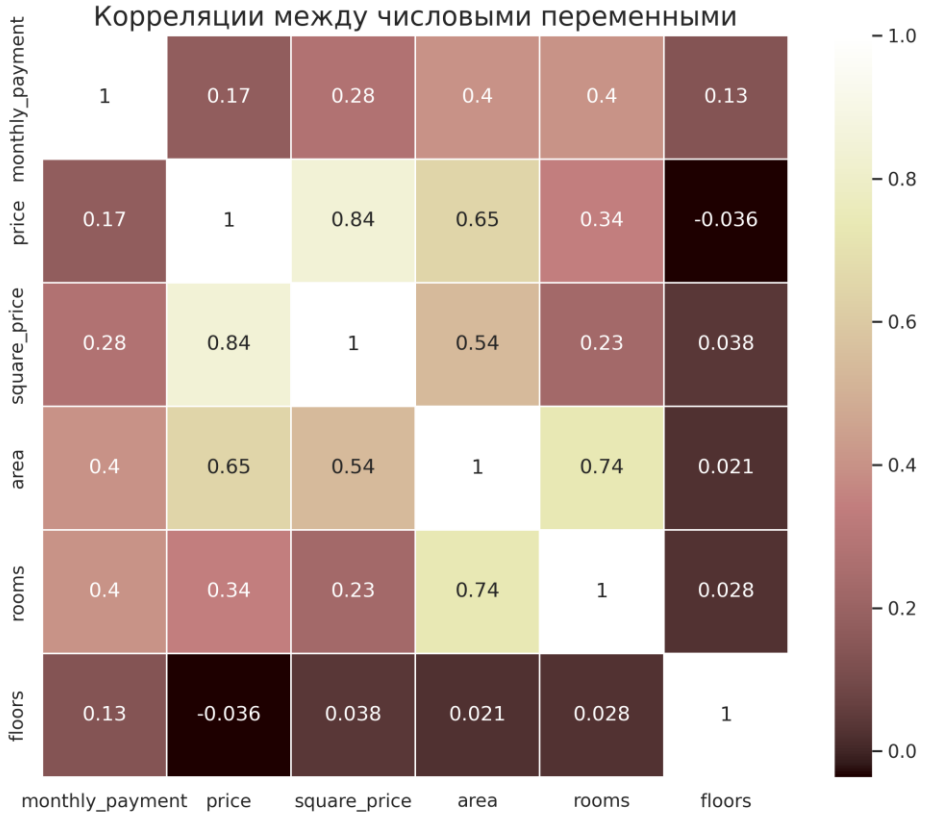

Взаимосвязь между floors и ценой за квадрат
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Цена за квадрат
0
20
40
60
80
floors


Взаимосвязь между area и ценой за квадрат
1e6
Цена за квадрат
area
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
50
100
150
200
250
300
350
400


Зависимость цены от площади
1e9
Цена
Площадь (м²)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
200
400
600
800


Распределение цен
Частота
Цена
1e9
0
250
500
750
1000
1250
1500
1750
2000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


Распределение площади
Частота
800
600
400
200
0
0
200
400
600
800
Площадь (м²)


Корреляции между числовыми переменными
monthly_payment
price
square_price
area
rooms
floors
1
0.17
0.28
0.4
0.4
0.13
0.17
1
0.84
0.65
0.34
-0.036
0.28
0.84
1
0.54
0.23
0.038
0.4
0.65
0.54
1
0.74
0.021
0.4
0.34
0.23
0.74
1
0.028
0.13
-0.036
0.038
0.021
0.028
1
monthly_payment
price
square_price
area
rooms
floors
1.0
0.8
0.6
0.4
0.2
0.0