Аналитический отчет по анализу данных недвижимости

Автор: [Пальчак Тимофей Станиславович]

Группа: [ИСП-23В]

---

1. Введение

1.1 Цель исследования

Целью данной работы является анализ данных о недвижимости, полученных с помощью API DomClick, для выявления факторов, влияющих на стоимость квадратного метра жилья, 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. 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.
- **KNN Imputer** для заполнения пропущенных значений на основе значений ближайших соседей.

Источником данных является API DomClick. Использованы запросы к API для получения информации о квартирах в Москве по различным параметрам (количество комнат, площадь, этаж и т.д.).

---

3. Этапы работы

3.1 Загрузка данных через API DomClick

Для загрузки данных был разработан класс DomClickApi, который выполняет автоматизированные запросы к API DomClick с необходимыми параметрами (тип недвижимости, регион, количество комнат и т.д.). На каждом этапе выводились промежуточные данные для проверки корректности полученной информации.

3.2 Предварительная обработка данных

После загрузки данных из API был выполнен следующий процесс:

- Создан DataFrame с нужными колонками: price, area, rooms, square_price, subways, monthly_payment.
- Обнаружены пропущенные значения в колонке subways и monthly_payment, которые были обработаны с использованием **KNN Imputer** и других методов.

3.3 Выявление столбцов с пропущенными значениями

Проверка на пропущенные значения была выполнена с помощью кода:

В результате были обнаружены пропуски в колонках subways, monthly_payment и других, которые были заполнены с помощью **KNN Imputer** и других методов.

3.4 Визуализация данных

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

- **Диаграммы рассеяния** для колонок price, area, rooms относительно square_price.
- **Тепловая карта корреляции**, показывающая степень взаимосвязи между числовыми переменными (см. приложенный график).

---

4. Результаты и выводы

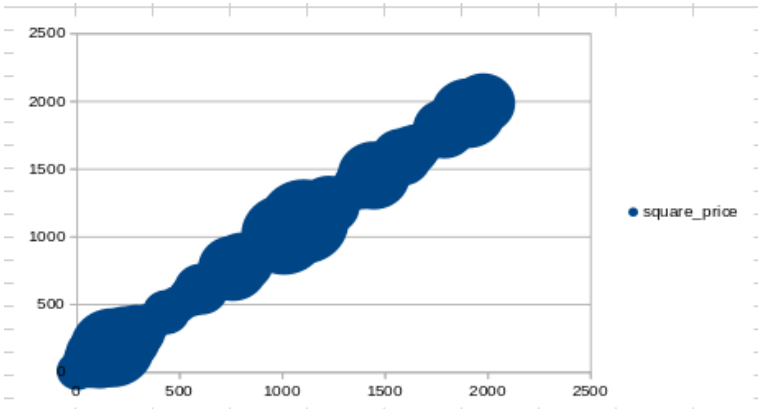
4.1 Анализ корреляции

Тепловая карта корреляции показала следующие ключевые зависимости:

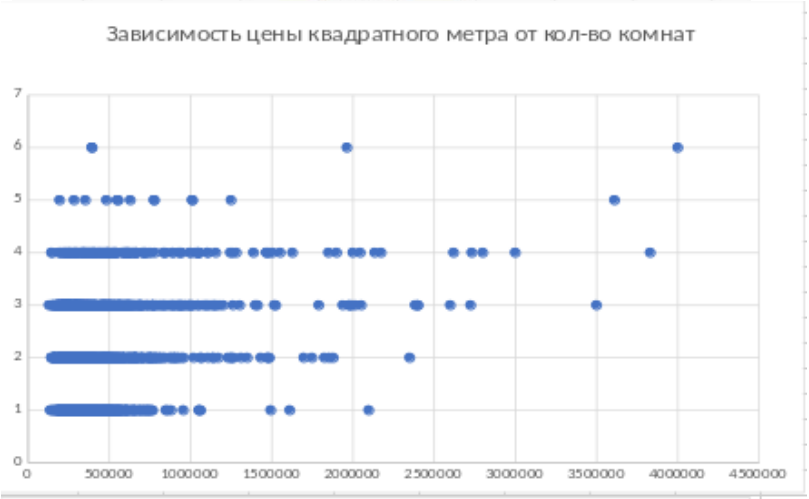
- **Цена за квадратный метр (square_price)** наиболее сильно коррелирует с общей ценой (price) и площадью квартиры (area).
- Количество комнат (rooms) показало слабую корреляцию с ценой за квадратный метр, что говорит о меньшем влиянии этого параметра на стоимость в сравнении с общей площадью и общей ценой.

4.2 Обработка пропущенных значений

Использование **KNN Imputer** позволило корректно заполнить пропуски в колонке subways, основываясь на схожих значениях соседних объектов. Данный метод обеспечил более точное восстановление данных по сравнению с простыми статистическими методами (среднее, медиана и т.д.).


2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
square_price


Зависимость цены квадратного метра от площади квар-
тиры
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
2 94 186 278 370 462 554 646 738 830 922 1014 1106 1198 1290 1382 1474 1566 1658 1750 1842 1934
square_price
area


Зависимость цены квадратного метра от кол-во комнат
7
6
5
4
3
2
1
0
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000


Зависимость цены квадратного метра от этажа
100
90
80
70
60
50
40
30
20
10
0
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000

5. Рекомендации

1. **Использование обработанных данных для построения модели ценообразования**:
    - Данные готовы для обучения модели машинного обучения, которая может предсказывать стоимость квартиры на основе признаков price, area, rooms, subways.
2. **Регулярное обновление данных через API**:
    - Для актуальности модели рекомендуется периодически обновлять данные через **API DomClick**, чтобы учесть изменения на рынке недвижимости.
3. **Дальнейший анализ категориальных переменных**:
    - Рекомендуется детально изучить влияние других категориальных признаков, таких как renovation и placement_type, которые могут оказать влияние на цену.

---

6. Заключение

В ходе работы был проведен анализ и очистка данных о рынке недвижимости, полученных из **DomClick API**. 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.

Приложения

1. Примеры кода для работы с API и очистки данных.
2. Графики зависимости и корреляции.

Отчет составлен на основе данных, полученных из DomClick API.