Аналитический отчет по анализу данных недвижимости

Автор: [Матвеева Алина Алексеевна]

Группа: [ИСП-23В]

1. Введение

1.1 Цель исследования

Данная работа направлена на анализ данных о недвижимости, собранных с помощью API DomClick. Основной целью является выявление факторов, влияющих на стоимость квадратного метра жилья, а также подготовка данных для последующего применения в моделях машинного обучения.

1.2 Задачи

1. Сбор и предварительная очистка данных о недвижимости.
2. Провести анализ числовых и категориальных переменных.
3. Заполнение пропущенных значений и подготовка набора данных для визуализации и корреляционного анализа.
4. Построить визуализации для выявления ключевых закономерностей.
5. Формулирование выводов и рекомендаций для дальнейшего использования данных.

2. Методология и инструменты

Для достижения поставленных задач были использованы следующие инструменты и библиотеки:

- **Python**: основной язык программирования для обработки данных и автоматизации запросов.
- **Pandas и NumPy**: библиотеки для анализа и подготовки данных.
- **Seaborn и Matplotlib**: библиотеки визуализации для построения графиков и тепловых карт корреляции.
- **KNN Imputer**: метод 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 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.

3.2 Предварительная обработка данных

После получения данных из API был выполнен следующий процесс:

- Создан DataFrame с необходимыми столбцами: **price**, **area**, **rooms**, **square_price**, **subways**, **monthly_payment**.
- В столбцах **subways** и **monthly_payment** были обнаружены пропущенные значения, которые были обработаны с помощью метода KNN Imputer и других подходов.

3.3 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.

3.4 Визуализация данных

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

- Диаграммы рассеяния для столбцов **`price`**, **`area`**, **`rooms`** относительно **`square_price`**.
- Тепловая карта корреляции, демонстрирующая степень взаимосвязи между числовыми переменными (см. прилагаемый график).

4. Результаты и выводы

4.1 Анализ корреляции

Тепловая карта корреляции выявила несколько ключевых зависимостей:

- 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 указывает на меньшее влияние этого параметра на стоимость по сравнению с общей площадью и ценой.

4.2 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.

5. Рекомендации

1. **Применение обработанных данных для создания модели ценообразования**:
    a. 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**:
    a. Для поддержания актуальности модели рекомендуется периодически обновлять данные через API DomClick, чтобы учитывать изменения на рынке недвижимости.
3. **Дальнейший анализ категориальных переменных**:
    a. Рекомендуется провести более глубокое исследование влияния других категориальных признаков, таких как **renovation** и **placement_type**, которые могут существенно влиять на цену.

6. 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.