

Аналитический отчет по данным недвижимости

Введение

Цель данного проекта — анализ данных о недвижимости для получения полезной информации и проведения дальнейших прогнозов, определения стоимости недвижимости, анализа рынка или прогнозирования спроса. Я работал с данными, сталкиваемся с отсутствующими значениями и некорректными данными, которые требуют очистки и предобработки.

Шаг 1: Очистка данных

Перед началом анализа я провёл очистку данных, так как исходный набор содержал некорректные и пропущенные значения:

1. **Замена значений -1 на NaN:** В наборе данных встречались значения -1, которые обозначают отсутствие информации. Эти значения мы заменили на NaN, чтобы более корректно обрабатывать пропуски.
2. **Обработка NaN значений:**
 - a. Для **числовых столбцов** (таких как `living_meters`, `kitchen_meters`, `ceiling_height` и т.д.) пропущенные значения были заполнены средними значениями по соответствующему столбцу.
 - b. Для **категориальных столбцов** (например, `object_type`, `parking_type`, `district`) пропуски заполнили наиболее частым значением (модой) для каждого столбца.
3. **Приведение типов данных:** Мы конвертировали числовые столбцы в тип `int` или `float` для удобства последующего анализа.

Шаг 2: Исследовательский анализ данных

После очистки данных мы провели первичный анализ:

1. **Распределение числовых признаков:**
 - a. Параметры, такие как `living_meters`, `kitchen_meters` и `ceiling_height`, показывают разброс значений, который может помочь в определении типов объектов и их уникальных характеристик.

- b. Показатель `price` имеет большой разброс, что говорит о высокой разнице в стоимости недвижимости в зависимости от других характеристик.

2. Анализ категориальных признаков:

- a. **Тип недвижимости:** Основные категории включают «Вторичка» и «Новостройка». Это может быть полезным для дальнейшего анализа.
- b. **Тип парковки:** Параметр `parking_type` показывает, есть ли у объекта парковка и какого она типа. Большинство объектов имеют категорию «Наземная» или «Открытая» парковка.
- c. **Расположение и район:** `district` и `street` содержат информацию о месте расположения объектов, что является важным фактором для анализа стоимости и спроса на жилье.

Шаг 3: Выявление зависимости между параметрами

1. **Зависимость цены от общей площади (`total_meters`):** Существует положительная корреляция между общей площадью и ценой — чем больше площадь, тем выше цена объекта.
2. **Влияние высоты потолков (`ceiling_height`) на цену:** Высота потолков также может играть роль в формировании стоимости жилья, особенно для элитных объектов, где высокие потолки считаются преимуществом.
3. **Анализ района и типа недвижимости:** Объекты в определенных районах и новостройки, как правило, имеют более высокую стоимость. Этот параметр важен для определения уровня доступности жилья и его инвестиционной привлекательности.

Заключение

Данный проект включает в себя важные шаги предобработки данных, что позволило подготовить их для дальнейшего анализа и моделирования. Очистка данных и замена пропущенных значений были необходимы для получения корректных результатов и проведения статистического анализа.

В будущем, на основе этих данных можно строить модели машинного обучения для прогнозирования стоимости недвижимости или определения параметров, наиболее влияющих на цену. Такие модели помогут точнее оценить объекты недвижимости и будут полезны для агентов, покупателей и инвесторов на рынке жилья.

Рекомендации

1. **Дополнить данные:** Добавить больше данных о конкретных характеристиках районов (например, близость к метро, наличие школ и больниц), что улучшит качество моделей.
2. **Регулярно обновлять данные:** Рынок недвижимости динамичен, и важно поддерживать актуальность данных, чтобы модели могли учитывать последние изменения на рынке.