Описание проекта

В вашем распоряжении данные сервиса Яндекс Недвижимость — архив объявлений о продаже квартир в Санкт-Петербурге и соседних населённых пунктах за несколько лет. Вам нужно научиться определять рыночную стоимость объектов недвижимости. Для этого проведите исследовательский анализ данных и установите параметры, влияющие на цену объектов. Это позволит построить автоматизированную систему: она отследит аномалии и мошенническую деятельность.

По каждой квартире на продажу доступны два вида данных. Первые вписаны пользователем, вторые — получены автоматически на основе картографических данных. Например, расстояние до центра, аэропорта и других объектов — эти данные автоматически получены из геосервисов. Количество парков и водоёмов также заполняется без участия пользователя.

Инструкция по выполнению проекта

Шаг 1. Откройте файл с данными и изучите общую информацию

Путь к файлу: */datasets/real_estate_data.csv*

Скачать датасет

. Загрузите данные из csv-файла в датафрейм с помощью библиотеки pandas.

. Изучите общую информацию о полученном датафрейме.

  . Постройте гистограмму для всех числовых столбцов таблицы на одном графике. Например, для датафрейма `data` можно построить такую гистограмму командой `data.hist(figsize=(15, 20))`. Напомним, что параметр `figsize` задаёт размер графика.

Шаг 2. Выполните предобработку данных

. Найдите и изучите пропущенные значения в столбцах:

.1. Определите, в каких столбцах есть пропуски.

.2. Заполните пропущенные значения там, где это возможно. Например, если продавец не указал число балконов, то, скорее всего, в его квартире их нет. Такие пропуски можно заменить на число 0. Если логичную замену предложить невозможно, то оставьте пропуски. Пропущенные значения — тоже важный сигнал, который нужно учитывать.

.3. В ячейке с типом `markdown` укажите причины, которые могли привести к пропускам в данных.

. Рассмотрите типы данных в каждом столбце:

.1. Найдите столбцы, в которых нужно изменить тип данных.

.2. Преобразуйте тип данных в выбранных столбцах.

.3. В ячейке с типом `markdown` поясните, почему нужно изменить тип данных.

. Изучите уникальные значения в столбце с названиями и устраните неявные дубликаты. Например, «поселок Рябово» и «поселок городского типа Рябово», «поселок Тельмана» и «посёлок Тельмана» — это обозначения одних и тех же населённых пунктов. Вы можете заменить названия в существующем столбце или создать новый с названиями без дубликатов.

Подсказка

Шаг 3. Добавьте в таблицу новые столбцы со следующими параметрами:

. цена одного квадратного метра (нужно поделить стоимость объекта на его общую площадь, а затем округлить до двух знаков после запятой);

. день недели публикации объявления (0 — понедельник, 1 — вторник и так далее);

. месяц публикации объявления;

. год публикации объявления;

. тип этажа квартиры (значения — «первый», «последний», «другой»);

. расстояние до центра города в километрах (переведите из *м* в *км* и округлите до ближайших целых значений).

Шаг 4. Проведите исследовательский анализ данных:

Изучите перечисленные ниже параметры объектов и постройте отдельные гистограммы для каждого из этих параметров. В некоторых параметрах встречаются редкие и выбивающиеся значения. При построении гистограмм удалите их. Например, в столбце `ceiling_height` может быть указана высота потолков 25 м и 32 м. Логично предположить, что на самом деле это вещественные значения: 2.5 м и 3.2 м. Попробуйте обработать аномалии в этом и других столбцах, если они есть. Если природа аномалии понятна и данные действительно искажены, то восстановите корректное значение. В противном случае удалите редкие и выбивающиеся значения.

Список параметров:

.1. общая площадь;
.2. жилая площадь;
.3. площадь кухни;
.4. цена объекта;
.5. количество комнат;
.6. высота потолков;
.7. тип этажа квартиры («первый», «последний», «другой»);
.8. общее количество этажей в доме;
.9. расстояние до центра города в метрах;
.10. расстояние до ближайшего парка

Опишите все ваши наблюдения по параметрам в ячейке с типом `markdown`.

. Изучите, как быстро продавались квартиры (столбец `days_exposition`). Этот параметр показывает, сколько дней было размещено каждое объявление.

.1. Постройте гистограмму.

.2. Посчитайте среднее и медиану.

.3. В ячейке типа `markdown` опишите, сколько времени обычно занимает продажа. Какие продажи можно считать быстрыми, а какие — необычно долгими?

Определите факторы, которые больше всего влияют на общую (полную) стоимость объекта.

Изучите, зависит ли цена от:

.4. общей площади;

.5. жилой площади;

.6. площади кухни;

.7. количества комнат;

.8. этажа, на котором расположена квартира (первый, последний, другой);

.9. даты размещения (день недели, месяц, год).

Постройте графики, которые покажут зависимость цены от указанных выше параметров. Для подготовки данных перед визуализацией вы можете использовать сводные таблицы.

. Посчитайте среднюю цену одного квадратного метра в 10 населённых пунктах с наибольшим числом объявлений — постройте сводную таблицу с количеством объявлений и средней ценой квадратного метра для этих населенных пунктов. Выделите населённые пункты с самой высокой и низкой стоимостью квадратного метра.

. Ранее вы посчитали расстояние до центра в километрах. Теперь выделите квартиры в Санкт-Петербурге с помощью столбца `locality_name` и вычислите их среднюю стоимость на разном удалении от центра. Учитывайте каждый километр расстояния: узнайте среднюю цену квартир в одном километре от центра, в двух и так далее. Опишите, как стоимость объектов зависит от расстояния до центра города — постройте график изменения средней цены для каждого километра от центра Петербурга.

Шаг 5. Напишите общий вывод

Опишите полученные результаты и зафиксируйте итоговый вывод проведённого исследования.

Оформление

Выполните задание в Jupyter Notebook. Заполните программный код в ячейках типа `code`, текстовые пояснения — в ячейках типа `markdown`. Примените форматирование и заголовки.

Описание данных

- `airports_nearest` — расстояние до ближайшего аэропорта в метрах (м)
- `balcony` — число балконов
- `ceiling_height` — высота потолков (м)
- `cityCenters_nearest` — расстояние до центра города (м)
- `days_exposition` — сколько дней было размещено объявление (от публикации до снятия)
- `first_day_exposition` — дата публикации
- `floor` — этаж
- `floors_total` — всего этажей в доме
- `is_apartment` — апартаменты (булев тип)
- `kitchen_area` — площадь кухни в квадратных метрах (м²)
- `last_price` — цена на момент снятия с публикации
- `living_area` — жилая площадь в квадратных метрах (м²)
- `locality_name` — название населённого пункта
- `open_plan` — свободная планировка (булев тип)
- `parks_around3000` — число парков в радиусе 3 км
- `parks_nearest` — расстояние до ближайшего парка (м)
- `ponds_around3000` — число водоёмов в радиусе 3 км
- `ponds_nearest` — расстояние до ближайшего водоёма (м)
- `rooms` — число комнат

- `studio` – квартира-студия (булев тип)
- `total_area` – общая площадь квартиры в квадратных метрах (м²)
- `total_images` – число фотографий квартиры в объявлении