Проект: Исследовательский анализ данных

Поздравляем! Вы прошли курс в тренажёре. Пора применить новые знания на практике и самостоятельно решить аналитический кейс.

Когда закончите работу над проектом, отправьте его ревьюеру. Он проверит ваше решение вручную и в течение суток вышлет вам комментарии. Их нужно учесть, чтобы доработать проект. Затем вы должны вернуть ревьюеру обновлённый вариант.

Ревьюер может повторно выслать вам комментарии. Это нормально — доработка часто проходит в несколько этапов.

Проект завершён, когда засчитаны все исправления.

Описание проекта

В вашем распоряжении данные сервиса Яндекс Недвижимость — архив объявлений за несколько лет о продаже квартир в Санкт-Петербурге и соседних населённых пунктах.

Ваша задача — выполнить предобработку данных и изучить их, чтобы найти интересные особенности и зависимости, которые существуют на рынке недвижимости.

О каждой квартире в базе содержится два типа данных: добавленные пользователем и картографические. Например, к первому типу относятся площадь квартиры, её этаж и количество балконов, ко второму — расстояния до центра города, аэропорта и ближайшего парка.

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 файла в датафрейм.

Изучите общую информацию о полученном датафрейме.

Постройте общую гистограмму для всех числовых столбцов таблицы. Например, для датафрейма `data` это можно сделать командой

```
data.hist(figsize=(15, 20)) .
```

Шаг 2. Предобработка данных

Найдите и изучите пропущенные значения в столбцах:

Определите, в каких столбцах есть пропуски.

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

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

Рассмотрите типы данных в каждом столбце:

Найдите столбцы, в которых нужно изменить тип данных.

Преобразуйте тип данных в выбранных столбцах.

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.

Найдите и устраните редкие и выбивающиеся значения. 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 природа аномалии понятна и данные действительно искажены, то восстановите корректное значение.

В противном случае удалите редкие и выбивающиеся значения.

В ячейке с типом `markdown` опишите, какие особенности в данных вы обнаружили.

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

цена одного квадратного метра;

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 целых значений).

Шаг 4. Проведите исследовательский анализ данных:

Изучите следующие параметры объектов:

общая площадь;

жилая площадь;

площадь кухни;

цена объекта;

количество комнат;

высота потолков;

этаж квартиры;

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

общее количество этажей в доме;

расстояние до центра города в метрах;

расстояние до ближайшего аэропорта;

расстояние до ближайшего парка;

день и месяц публикации объявления.

Постройте отдельные гистограммы для каждого из этих параметров. 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.

Постройте гистограмму.

Посчитайте среднее и медиану.

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

Какие факторы больше всего влияют на общую (полную) стоимость объекта?

Изучите, зависит ли цена от:

общей площади;

жилой площади;

площади кухни;

количества комнат;

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

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. Выделите населённые пункты с самой высокой и низкой стоимостью квадратного метра. Эти данные можно найти по имени в столбце `locality_name`.

Ранее вы посчитали расстояние до центра в километрах. Теперь выделите квартиры в Санкт-Петербурге с помощью столбца `locality_name` и вычислите среднюю цену каждого километра. Опишите, как стоимость объектов зависит от расстояния до центра города.

Шаг 5. Напишите общий вывод

Опишите полученные результаты и зафиксируйте основной 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

Как будут проверять мой проект?

💡 Если вашу работу отправили на доработку, пожалуйста, не удаляйте в Jupyter-тетрадке комментарии ревьюера. Так ревьюеру будет проще проверить изменения.

Мы подготовили критерии оценки проекта. Прежде чем решать кейс, внимательно изучите их.

На что обращают внимание ревьюеры, когда проверяют ваш проект:

- Как вы описываете выявленные в данных проблемы?
- Какие способы обработки пропусков вы применяете?
- Как используете срезы данных?
- Решают ли ваши графики поставленные задачи?
- Какие методы построения графиков вы используете?
- Выводите ли вы финальные данные в сводных таблицах?
- Считаете ли показатели взаимосвязи в данных и как вы их объясняете?
- Соблюдаете ли вы структуру проекта и поддерживаете ли аккуратность кода?
- Какие выводы вы делаете?
- Оставляете ли вы комментарии к шагам?

Всё, что нужно для выполнения этого проекта, есть в шпаргалках и конспектах прошлых уроков.

Успехов!