**Проект: Исследовательский анализ данных**

**Исследование объявлений о продаже квартир**

Поздравляем! Вы прошли курс в тренажёре. Самое время проверить полученные знания на практике и решить аналитический кейс. Выполнять работу вы будете самостоятельно.

Как закончите работу над проектом, отправьте его на проверку ревьюеру. В течение суток вы получите комментарии. Их нужно учесть: доработать проект и вернуть ревьюеру обновлённый вариант.

Скорее всего, вы снова получите комментарии по кейсу. Это нормально — доработка может проходить в несколько этапов.

Проект завершён, когда засчитаны все исправления.

**Описание проекта**

В вашем распоряжении данные сервиса Яндекс Недвижимость — архив объявлений о продаже квартир в Санкт-Петербурге и соседних населённых пунктах за несколько лет. Нужно научиться определять рыночную стоимость объектов недвижимости. Ваша задача — провести исследовательский анализ данных, который поможет установить 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](https://code.s3.yandex.net/datasets/real_estate_data.csv)

1. Загрузите данные из csv-файла в датафрейм c помощью библиотеки pandas.
2. Изучите общую информацию о полученном датафрейме.
3. 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. Предобработка данных**

1. 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 можно заменить на число 0. Если логичную замену предложить невозможно, то оставьте пропуски. Пропущенные значения — тоже важный сигнал, который нужно учитывать.
   * В ячейке с типом markdown укажите причины, которые могли привести к пропускам в данных.
2. 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.
3. 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:**

1. 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:**

* + общая площадь;
  + жилая площадь;
  + площадь кухни;
  + цена объекта;
  + количество комнат;
  + высота потолков;
  + тип этажа квартиры («первый», «последний», «другой»);
  + общее количество этажей в доме;
  + расстояние до центра города в метрах;
  + расстояние до ближайшего парка

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

1. 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?
2. Какие факторы больше всего влияют на общую (полную) стоимость объекта?

Изучите, зависит ли цена от:

* + общей площади; total\_area
  + жилой площади; living\_area
  + площади кухни; kitchen\_area
  + количества комнат; rooms
  + этажа, на котором расположена квартира (первый, последний, другой); floor (floor\_type)
  + даты размещения (день недели, месяц, год) weekday month year.

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

1. 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.
2. 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 — 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

**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!