GeekBrains

Исследование объявлений о продаже квартир в г.Ульяновск на основе объявлений с сайта AVITO

Специальность: Аналитик больших данных, Geekbrains

Выполнил: Бурдасов А. В.

2023

Оглавление

Введение 3

Глава 1. Основы анализа данных с использованием Python 4

1.1 Анализ данных. 4

1.2 Инструменты для проведения анализа данных. 5

1.3 Основные библиотеки Python для анализа данных. 11

1.4 Jupyter Notebook и преимущества его использования. 17

1.5 Этапы анализа данных и использование Jupyter Notebook. 21

Глава 2. Подготовка датасета объявлений по продаже квартир в г.Ульяновск за август 2023г. на сайте Avito к анализу 23

2.1 Постановка задачи 23

2.2 Извлечение данных 23

2.3 Подготовка данных. 24

2.4 Исследование и визуализация данных 36

2.5 Интерпретация результатов. 49

Заключение 50

Список используемой литературы 51

Приложения 52

Введение

Тема проекта:

Исследование объявлений о продаже квартир в г.Ульяновск на основе объявлений с сайта Avito.

Цель:

Изучить инструменты для анализа данных, преимуществе использования Python, основных библиотеках Python, применяемых в анализе, преимуществах и особенностях работы в Jupyter Notebook. Используя Python и Jupyter Notebook, выполнить предобработку данных объявлений по продаже квартир в г.Ульяновск за август 2023г. на сайте Avito и изучить их, чтобы найти интересные особенности и зависимости, которые существуют на рынке недвижимости.

Задачи:

- Изучить литературу, касающуюся темы исследования;
- Рассмотреть основные виды и методы анализа;
- Получить и провести предварительную обработку данных, сформировать датасет для анализа;
- Провести анализ датасета;
- Сделать выводы по содержанию анализа и по качеству предоставленных данных.

Инструменты:

Jupyter Notebook, Python с библиотеками pandas, matplotlib, seaborn

Состав команды:

Бурдасов А.В.- аналитик

Глава 1. Основы анализа данных с использованием Python.

1.1 Анализ данных.

Данные — это информация, которая собирается, организуется и анализируется для получения знаний и принятия решений. Это могут быть числа, текст, изображения или любой другой формат информации. С развитием технологий в мире генерируется все больше и больше данных каждый день и эти данные необходимо анализировать, находить в них природную зависимость, делать выводы на основе этих данных.

Анализ данных — это широкий термин, охватывающий множество различных типов анализа данных. Любой тип информации может быть подвергнут методам анализа данных для получения информации, которая может быть использована для улучшения ситуации. Методы анализа данных позволяют выявить тенденции и показатели, которые в противном случае были бы утеряны в массе информации. Затем эта информация может быть использована для оптимизации процессов с целью повышения общей эффективности бизнеса или системы.

Необходимость анализа данных обусловлена тем, что он позволяет принимать более обоснованные решения в различных сферах жизни и бизнеса. Анализ данных необходим для того, чтобы принимать обоснованные решения на основе фактических данных, а не интуиции или предположений. Он позволяет выявить тенденции, определить причины и следствия различных событий, а также прогнозировать будущие результаты. Например, в маркетинге анализ данных помогает определить наиболее эффективные каналы рекламы и способы привлечения клиентов. В экономике анализ данных используется для прогнозирования инфляции, безработицы и других макроэкономических показателей.

Анализ данных играет ключевую роль в принятии решений на всех уровнях - от индивидуального до корпоративного, и является неотъемлемой частью успешного бизнеса и управления. В целом, анализ данных позволяет

организациям и индивидуумам принимать более информированные решения, улучшать свои продукты и услуги, а также оптимизировать процессы и ресурсы.

1.2 Инструменты для проведения анализа данных.

Работа с данными не осуществляется вручную, она предполагает использование специальных инструментов и состоит из нескольких этапов: сбора, анализа, визуализации и прогнозирования данных.

Чтобы решать аналитические задачи, специалисты используют разное программное обеспечение и приложения. Все инструменты аналитика делятся на несколько типов в зависимости от того, для какого этапа решения задачи они предназначены.

- Инструменты для сбора и хранения данных. В любой компании есть своя база данных. В одной это могут быть таблицы Excel, в другой — серьёзные решения типа Oracle или MySQL. Задача этих инструментов бизнес-анализа — хранить большие объёмы данных и быстро извлекать их.
- Для анализа данных. Чтобы собранные данные не лежали мёртвым грузом, а работали, их нужно доставать из базы данных и анализировать по определённым критериям с помощью различных программ. Один из самых популярных инструментов для аналитики данных — Jupyter Notebook.
- Для визуализации данных. Информацию, которую получили после анализа данных, нужно представить в удобном и понятном виде. Чтобы создавать наглядные графики и отчёты, используют программы и сервисы для визуализации. К простым относятся Power Point или Miro. Более сложные инструменты работы с аналитикой — Tableau, Power BI.
- Программы и сервисы для визуализации. Информацию, которую получили в ходе анализа данных, удобно изучать на дашбордах — интерактивных панелях с графическим интерфейсом.
- Для прогнозирования данных. Такие инструменты нужны, чтобы на основании прошлого опыта компании могли принимать успешные решения в будущем, создавать модели поведения клиентов, составлять прогнозы

ежедневного спроса определённой группы товаров и т. д. Чтобы создавать достоверные прогнозы, специалисты используют ключевые инструменты аналитиков: языки программирования Python, R и другие.

Инструменты и программы для аналитики данных бывают бесплатные и коммерческие.

- Бесплатные инструменты анализа данных. Имеют открытый исходный код, а апгрейд до платных версий не обязателен. Это значит, что любой специалист может расширять возможности инструмента, изменяя исходный код. В роли службы поддержки обычно выступает сообщество пользователей. Инструменты с открытым исходным кодом используют и стартапы, и крупные компании, потому что по уровню возможностей эти программы часто не уступают платным продуктам.
- Коммерческие инструменты бизнес-аналитики. Это программное обеспечение с закрытым исходным кодом. Эти инструменты нельзя изменить, и обычно они дорого стоят. Зато вся поддержка, обучение и устранение неполадок целиком лежит на разработчике программного продукта.

Основные инструменты аналитика помогают ему собирать, обрабатывать, анализировать и интерпретировать данные. Несмотря на большое количество сервисов и программного обеспечения, на практике специалист использует в работе 3–4 ключевых инструмента. Их выбор зависит не только от знаний и опыта аналитика, но и от того, с чем уже работает компания. Например, если бизнес использует Tableau, аналитику придётся работать с ним, даже если он привык работать в Power BI. А вот в плане написания кода специалист свободен в выборе и может использовать любой язык программирования.

Существует множество инструментов, технологий и приложений и важно понимать, что можно сделать с помощью той или иной технологии и программы.

Кратко опишем некоторые из них:

- Microsoft Excel:

Microsoft Excel — программа для работы с электронными таблицами, созданная корпорацией Microsoft. Это базовый инструмент, которым должен владеть каждый, кто хочет работать с данными. Это не только таблицы и формулы: Excel даёт большие возможности для обработки данных и помогает решать задачи разного масштаба, вплоть до обработки большого массива данных с помощью плагинов. Помимо базовых функций, условного форматирования, сводных таблиц и диаграмм аналитику важно овладеть надстройкой Power Query: она позволяет интегрировать в Excel и обрабатывать данные из внешних источников.

- R:

R — это язык программирования, который используется для анализа данных и статистической обработки. Он был разработан в 1996 году Россом Ихакой (Ross Ihaka) и Робертом Джентльменом (Robert Gentleman). R имеет широкий спектр функций для работы с данными, включая визуализацию, моделирование, статистику и многое другое. Он также имеет большое сообщество пользователей и разработчиков, что делает его очень популярным инструментом для анализа данных.

- Scala:

Scala — это язык программирования, предназначенный для создания масштабируемых, безопасных и производительных приложений. Он был разработан в компании Martin Odersky и распространяется под лицензией Apache Software License. Scala является мультипарадигменным языком, то есть он поддерживает несколько стилей программирования, включая функциональный, императивный, объектно-ориентированный и аспектно-ориентированный стили. Это позволяет разработчикам использовать тот стиль, который наиболее подходит для решения конкретной задачи. Одной из особенностей Scala является его производительность. Язык спроектирован так, чтобы обеспечивать высокую производительность на всех уровнях: от компиляции до выполнения. Это достигается за счет автоматического управления памятью, оптимизации кода и других механизмов.

Кроме того, Scala имеет обширную стандартную библиотеку, которая включает в себя множество инструментов для работы с коллекциями, параллельным выполнением, обработкой ошибок и другими аспектами программирования.

- Julia:

Julia — это высокопроизводительный язык программирования с открытым исходным кодом, который предназначен для быстрой разработки и высокопараллельных вычислений. Он был создан Джеффом Безансоном (Jeff Bezanson) и Айвором Якобсоном (Ivor Jacobson) в 2012 году. Julia отличается от других языков программирования своей производительностью, удобством использования и гибкостью. Она может использоваться для решения широкого спектра задач, включая научные вычисления, машинное обучение, обработку данных и многое другое.

- SAS:

SAS (Statistical Analysis System) — это программное обеспечение для статистического анализа данных, разработанное компанией SAS Institute. SAS используется для проведения различных видов статистического анализа, таких как анализ взаимосвязи, прогнозирование, анализ выживаемости и другие. SAS является одним из наиболее популярных программных продуктов в области статистического анализа и используется в различных отраслях.

- Stata:

Stata — это статистический пакет от компании StataCorp предназначеный для статистических исследований над разнообразными выборками данных из различных предметных областей и дисциплин. Система предоставляет сотни статистических инструментов для управления данными, статистического анализа и прочих задач анализа данных. Stata распространяется более чем в 180 странах и используется сотнями тысяч профессиональных исследователей и аналитиков.

- SQL:

SQL (Structured Query Language) — это стандартный язык запросов, используемый для работы с реляционными базами данных. SQL позволяет пользователям создавать, изменять и удалять данные, а также выполнять различные операции с ними, такие как выборка, объединение, группировка и сортировка. У SQL есть разновидности. Например, система управления базами данных MySQL, в которой можно хранить любые данные: контакты клиентов, карточки товаров, информацию о дате публикации материалов и т. д. PostgreSQL — более сложная система, которая подходит для управления большими базами данных и обработки сложных запросов, например в финансовой сфере, промышленности, крупном ретейле. SQL и его разновидности — это инструменты с открытым исходным кодом, поэтому доступны бизнесу любой сферы и формата.

SQL является одним из основных инструментов работы с данными в современных информационных системах.

- Power BI:

Power BI — это мощный инструмент для бизнес-аналитики, который позволяет пользователям визуализировать и анализировать данные. Он предоставляет различные возможности для создания отчетов, графиков и дашбордов, а также позволяет работать с большими объемами данных. Power BI также имеет возможность интеграции с другими сервисами, такими как Excel, SQL Server и облачными хранилищами данных. Технически система Power BI состоит из нескольких сервисов, которые взаимодействуют между собой, создавая платформу для полного цикла работы с данными: от сбора и обработки до визуализации и распространения. Power BI Gateway отвечает за установку безопасного соединения между локальными данными и облачным сервисом Power BI Service. Создавать отчёты и дашборды можно в приложении Power BI Desktop, инструменты Power BI Embedded помогают встроить эти отчёты в веб-приложения и встроенные системы, а Power BI Mobile предоставляет доступ к данным и отчётам из любой точки мира.

- Tableau

Tableau — это программное обеспечение для бизнес-аналитики и визуализации данных. Оно позволяет пользователям создавать наглядные отчеты, диаграммы и графики на основе различных источников данных, таких как базы данных, файлы Excel и API. Tableau помогает анализировать данные, выявлять тенденции и закономерности для принятия обоснованных бизнес-решений. Tableau имеет интуитивно понятный интерфейс, который позволяет пользователям легко создавать визуализации без необходимости написания кода. Программа предлагает широкий спектр инструментов и опций для настройки внешнего вида и функциональности диаграмм. Она также поддерживает совместную работу, позволяя нескольким пользователям одновременно работать над проектом и просматривать изменения в реальном времени.

- Talend

Talend — ETL-инструмент, который упрощает и оптимизирует процесс интеграции данных. ETL-технологии (Extract, Transform, Load) — «извлечение, преобразование и загрузка» — используют, когда нужно быстро объединить данные из нескольких источников. Например, сеть магазинов продаёт одежду онлайн и офлайн. Чтобы оценить эффективность продаж по двум источникам, нужно подгрузить данные из нескольких баз. Информацию можно скачивать по очереди из CRM, систем аналитики веб-трафика и других. А можно сделать это одновременно — с помощью Talend. Данные интегрируются, и их можно использовать для дальнейшего анализа. Talend — инструмент с открытым исходным кодом, а значит, базовую версию программы можно использовать бесплатно.

- Python и библиотеки для обработки и анализа данных:

Python — это язык программирования и универсальный инструмент для работы с данными. Как язык программирования Python имеет простой синтаксис, поэтому писать код на нём получается быстрее, чем на других языках. Это интерпретируемый язык — он не требует предварительной компиляции перед выполнением кода, что значительно ускоряет процесс разработки и отладки.

Достаточно написать скрипт или программу, чтобы выгрузить данные, создать machine learning модель, построить нейронную сеть или собрать статистику. Для каждой задачи Python имеет свою библиотеку. В отличие от Excel, Python абсолютно бесплатен для скачивания и использования. Python является одним из самых простых и популярных языков программирования, что делает его идеальным выбором для начинающих специалистов по анализу данных. Его универсальность позволяет использовать его в различных сферах, включая научные исследования, веб-разработку, машинное обучение и многое другое.

В дальнейшем будем рассматривать анализ данных с помощью Python, и он будет являться основным инструментов для работы с нашими данными.

1.3 Основные библиотеки Python для анализа данных.

Для анализа данных в Python существует множество инструментов, таких как Pandas, Numpy, Scipy, Matplotlib и другие. Каждый из них имеет свои особенности и предназначен для выполнения определенных задач. Например, Pandas используется для работы с табличными данными, Numpy - для работы с многомерными массивами, Scipy — для выполнения научных вычислений, а Matplotlib — для визуализации данных.

Подробнее рассмотрим основные библиотеки.

- Pandas:

Выпущенный в 2008 году, Pandas является расширением программной библиотеки Python. Он работает с данными, хранящимися в Python, для манипулирования и анализа данных. Позволяет манипулировать данными, выполнять различные статистические операции. Pandas работает прямо на задней панели Python. В результате получается чрезвычайно быстро и эффективно. Так, например MS Excel, как только вы превышаете 10 000 строк, начинает значительно замедляться. С другой стороны, Pandas не имеет реальных ограничений и легко обрабатывает миллионы точек данных. С точки зрения чистого пространства, Excel ограничивает одну электронную таблицу ровно 1

048 576 строками. На этом этапе ваши вычисления заняли бы целую вечность. Более вероятно, что Excel просто выйдет из строя. Миллион строк может показаться большим объемом данных, но для специалистов по обработке данных это всего лишь капля в море. Однако у Pandas нет ограничений на количество точек данных, которые вы можете иметь в наборе данных. Он ограничен только вычислительной мощностью и памятью компьютера, на котором он запущен. Также проще создавать и использовать сложные уравнения и вычисления на основе ваших данных. С помощью Pandas вы можете мгновенно применить сотни вычислений к миллионам точек данных.

Особенности использования:

– Pandas предоставляет широкий спектр методов для работы с данными, такими как чтение и запись данных из различных источников, манипулирование данными (добавление, удаление, изменение), группировка и агрегация данных.
– Является высокопроизводительным и масштабируемым, что позволяет работать с большими объемами данных.
– Pandas имеет простой и понятный API, который делает его удобным для начинающих пользователей.
– Поддерживает большое количество типов данных, таких как массивы, серии, датафреймы и временные ряды.
– Pandas может обрабатывать более 15 различных форматов (в том числе самыми популярными, такими как CSV, JSON, Excel и SQL) и легко переключаться между ними.
– Имеет встроенную поддержку для работы с временными рядами, что позволяет анализировать временные ряды и прогнозировать будущие значения.
– Pandas поддерживает многоядерные процессоры и параллельные вычисления, что позволяет ускорить процесс обработки данных.

Ограничения использования Pandas:

– Он не подходит для работы с необработанными данными в реальном времени, так как он предназначен для анализа уже имеющихся данных.

– Pandas может быть сложным для понимания новичками, так как имеет много функций и методов.

– Может быть медленным с большими наборами данных, так как требует много памяти. , так как может не справиться с нагрузкой на сервер.

– Кроме того, он не очень эффективен при работе с неструктурированными данными, такими как JSON или XML.

– Также стоит отметить, что Pandas может быть не лучшим выбором для некоторых видов анализа данных, таких как машинное обучение или статистический анализ.

– Pandas требует установки дополнительных библиотек для некоторых функций, таких как Seaborn для построения графиков.

Pandas — это молниеносный инструмент, который позволяет легко выполнять задачи с большими данными. Эта библиотека позволяет провести очистку данных, заполнение недостающих значений, нормализацию данных, статистический анализ и многое другое.

Для визуализации анализируемых данных необходимо использование дополнительных библиотек Matplotlib или Seaborn.

- Matplotlib:

Matplotlib — это популярная библиотека для визуализации данных в Python путем создания 2D-графиков и диаграмм. Она предлагает широкий набор функций и возможностей для создания различных типов графиков и диаграмм.

Вот некоторые особенности Matplotlib:

– Богатый набор функций: Matplotlib предлагает множество функций для создания разнообразных графиков и диаграмм, таких как гистограммы, столбчатые диаграммы, круговые диаграммы и т. д.

– Простота использования: Matplotlib имеет простой и интуитивно понятный API, что делает его легким для изучения и использования.

– Настраиваемость: Matplotlib позволяет пользователям настраивать многие аспекты графиков, такие как цвета, шрифты, размеры и т. д., что позволяет создавать уникальные и привлекательные визуализации.

– Поддержка разных типов данных: Matplotlib может работать с разными типами данных, включая числовые, текстовые и даты.

– Кросс-платформенность: Matplotlib работает на разных операционных системах, таких как Windows, macOS и Linux.

– Открытый исходный код: Matplotlib является проектом с открытым исходным кодом, что означает, что его можно использовать бесплатно и модифицировать по своему усмотрению.

Однако у Matplotlib есть и некоторые ограничения, особенно для более продвинутых пользователей. Вот некоторые из них:

– Сложность настройки. Хотя Matplotlib предлагает широкие возможности настройки, некоторые пользователи могут найти процесс настройки сложным и трудоемким.

– Ограниченная интерактивность. Matplotlib не имеет встроенной поддержки для создания интерактивных графиков, что может быть важно для некоторых пользователей.

– Производительность. При работе с большими объемами данных производительность Matplotlib может снижаться.

– Зависимость от сторонних библиотек. Для создания некоторых видов графиков может потребоваться использование сторонних библиотек, таких как NumPy и SciPy.

– Ограничения по работе с 3D. Хотя в Matplotlib есть некоторые возможности для работы с 3D, они могут быть ограничены по сравнению с специализированными библиотеками, такими как Mayavi или PyOpenGL.

В качестве дополнение к библиотеке Matplotlib, для расширения ее возможностей, существуют дополни тельные библиотеки, например Seaborn,

которое предоставляет более удобные и красивые способы создавать более профессиональные и стильные графики.

- Seaborn:

Seaborn — это библиотека визуализации данных для языка программирования Python, которая предоставляет более удобные и визуально привлекательные способы визуализации данных по сравнению с библиотекой Matplotlib. Вот несколько особенностей использования Seaborn:

– Стилизованные графики. Seaborn предоставляет набор стилей для графиков, которые делают их более привлекательными и понятными.
– Быстрые и эффективные графики. Seaborn оптимизирован для быстрого создания качественных графиков, что особенно полезно при работе с большими наборами данных.
– Поддержка тем. Seaborn позволяет пользователям выбирать темы для своих графиков, что делает их более приятными для глаз.
– Встроенные функции: Seaborn содержит множество встроенных функций для обработки и визуализации данных, что упрощает работу с библиотекой.
– Совместная работа с Matplotlib. Seaborn является дополнением к библиотеке Matplotlib и позволяет использовать ее функции для создания более сложных визуализаций.

Несмотря на эти преимущества, Seaborn также имеет несколько ограничений в использовании:

– Сложность настройки. Seaborn предлагает несколько готовых стилей и тем для графиков, но некоторые пользователи могут хотеть настроить их больше. В этом случае, настройка может быть сложной из-за отсутствия подробной документации.
– Ограниченное количество поддерживаемых типов данных. Как и Matplotlib, Seaborn поддерживает только несколько типов данных, что может ограничить его использование в определенных случаях.

– Зависимость от Matplotlib. Несмотря на то, что Seaborn расширяет функциональность Matplotlib, он все еще зависит от этой библиотеки. Это может привести к некоторым проблемам, если пользователь хочет использовать только Seaborn или Matplotlib.

– Ограничения по работе с 3D. Seaborn не имеет встроенной поддержки для работы с 3D графикой, что может быть ограничением для некоторых пользователей.

Для проведения анализа данных необходимо использование математических расчетов, поэтому в языке Python существуют специальные математические библиотеки, используемые самостоятельно или как дополнение к другим библиотекам:

- Встроенный модуль math в Python предоставляет доступ к различным математическим функциям, таким как вычисление логарифмов, тригонометрических функций и т.д. Вы можете использовать его, чтобы выполнять математические операции без необходимости использовать другие библиотеки.
- Numpy — это библиотека для работы с многомерными массивами (от англ. Numerical Python). Это одна из основных библиотек для научных вычислений на языке Python. Numpy предоставляет множество функций для работы с массивами, индексирование, математические операции, численное дифференцирование и интегрирование, работа с комплексными числами. Кроме того, Numpy интегрирован с другими популярными библиотеками, такими как SciPy, Pandas и Matplotlib.
- SciPy (Scientific Python) — это библиотека для языка Python, предназначенная для научных и математических вычислений. Она содержит множество функций для решения задач линейной алгебры, оптимизации, обработки сигналов и др. SciPy также интегрирован с Numpy, что позволяет использовать его для работы с массивами.

Есть еще много других библиотек для анализа данных и машинного обучения: PyTorch — библиотека для машинного обучения и глубокого обучения, которая использует графы вычислений на языке Python.

Scikit-learn — библиотека которая предлагает широкий набор алгоритмов для классификации, регрессии и кластеризации данных.

TensorFlow — библиотека, которая используется для создания нейронных сетей и других сложных моделей.

Keras — высокоуровневый API для TensorFlow и PyTorch, который упрощает создание и обучение нейронных сетей.

Применение Python с библиотеками упрощают и ускоряют работу со структурированными данными и поэтому является отличным инструментом, когда дело касается анализа данных, в том числе многомерных массивов, когда необходимо выполнить большую задачу за короткое время.

1.4 Jupyter Notebook и преимущества его использования.

Классический подход в разработке на языке Python, когда с использованием среды разработки, например Visual Studio Code или pyCharm, программный код помещается в файл с расширением .py и отправляется интерпретатору для выполнения не очень удобен при анализе данных. Поэтому для Python (и не только) существует другой способ взаимодействия с интерпретатором — интерактивные блокноты Jupyter (Jupyter Notebook), сохраняющие промежуточное состояние программы между выполнением различных блоков кода, которые могут быть выполнены в произвольном порядке. Основной элемент системы - блокнот (notebook) объединяет код и его вывод в единый документ, который объединяет визуализацию, повествовательный текст, математические уравнения и другие мультимедиа. Этот интуитивно понятный рабочий процесс способствует итеративной и быстрой разработке, что делает ноутбуки все более популярным выбором для представления в данных и их анализа.

Проект Jupyter является преемником более раннего проекта IPython, который впервые был опубликован в качестве прототипа в 2010 году. Jupyter стал самостоятельным проектом, ориентированным на работу со множеством сред выполнения («расчётных ядер») — не только Python, но и R, Julia, Scala и ряда других. Первая самостоятельная версия вышла в 2014 году

Хотя в Jupyter Notebooks можно использовать с многими разными языками программирования, наиболее распространенный вариантом использования остается использование Python.

Jupyter Notebook – это веб-приложение с открытым исходным кодом, которое позволяет создавать и совместно использовать документы, содержащие живой код на Python, визуализации, текст и уравнения. В Jupyter можно использовать необходимые для анализа данных библиотеки Python, такие как NumPy, Pandas, и другие. Он также поддерживает интерактивные вычисления и визуализацию данных с помощью Matplotlib, Seaborn и других инструментов. Все используемые в проекте элементы, образуют файл или как он называется в Jupyter Notebook блокнот, и могут быть сохранены на компьютере или в облаке для дальнейшего использования.

Вот некоторые ключевые преимущества использования Jupyter Notebook:

– Гибкость и простота использования: Jupyter предлагает простой и понятный интерфейс, который позволяет пользователям быстро начать работу и легко переключаться между различными языками программирования и инструментами визуализации.

Jupyter Notebook позволяет пользователям вводить код непосредственно в ячейки документа, и результаты отображаются мгновенно, что делает процесс разработки и обучения более интерактивным. Кроме ячеек с выполняемым кодом, в Jupyter Notebook для записи любой текстовой информации, которую требет проводимая работа, существуют специальные ячейки Markdown. В данных ячейках текст может быть отформатирован с использованием языков

разметки markdown или LaTeX. Тексты с использованием markdown легко писать и читать, а поддержка LaTeX позволяет использовать в них математические формулы и символы. Эти тексты в дальнейшемп можно без труда сконвертировать в HTML. Большинство программистов предпочитают Markdown для написания документации, описаний своих проектов, написания блогов и так далее.

Визуализация данных: Jupyter поддерживает множество языков программирования, включая Python, R и Julia, что позволяет легко визуализировать и анализировать данные. Jupyter позволяет легко визуализировать данные с помощью различных библиотек, таких как Matplotlib, Seaborn, ggplot2 (для R) и других, предоставляя пользователям возможность быстро и наглядно представлять результаты аналитической работы.

Совместное использование и обмен: Jupyter обеспечивает возможность совместного использования документов, которые могут быть открыты и отредактированы другими пользователями, что облегчает сотрудничество и обмен идеями. А такая популярная платформа для размещения кода и для контроля версий и совместной работы как GitHub имеет встроенную поддержку рендеринга файлов .ipynb непосредственно как в репозиториях, так и в списках на своем веб-сайте.

Портативность: Документы Jupyter Notebook могут быть сохранены как файлы в формате ipynb, которые можно открывать и редактировать на любом устройстве с установленным Jupyter. Особенность этого формата в том, что каждый файл .ipynb представляет собой текстовый файл, который описывает содержимое вашего документа в формате JSON. Каждая ячейка и ее содержимое, включая вложения изображений, которые были преобразованы в строки текста, перечислены в нем вместе с некоторыми метаданными.

Открытый исходный код: Jupyter является проектом с открытым исходным кодом и активно поддерживается сообществом разработчиков, что гарантирует его долгосрочную поддержку и развитие.

Jupyter Notebook доступен на разных платформах, включая Windows, macOS и Linux. Он может быть использован как локально, так и в облаке, например, на платформе Amazon Web Services (AWS).

– Интеграция с облачными сервисами: Jupyter интегрируется с различными облачными сервисами, такими как Google Colab, Amazon SageMaker и Microsoft Azure, что упрощает процесс работы с данными и ускоряет разработку.

– Безопасность и конфиденциальность: Jupyter использует протокол HTTPS для шифрования данных и соединения с серверами, что обеспечивает безопасность передачи данных и защиту от возможных атак.

Но есть и недостатки Jupyter Notebook:

Из-за интерпретируемости языка Python, Jupyter Notebook может страдать от снижения производительности при выполнении сложных или объемных скриптов.

Jupyter Notebook может загружаться довольно медленно, особенно если пользователь работает с большим количеством данных.

По умолчанию Jupyter Notebook не поддерживает системы контроля версий, что означает, что изменения в файлах могут быть потеряны, если не принять дополнительные меры предосторожности.

В целом, Jupyter Notebook — универсальный инструмент аналитика данных, он представляет собой мощное средство для разработки, анализа данных и обучения, который предлагает широкий спектр возможностей и преимуществ для пользователей. Jupyter Notebook. Это идеальный инструмент для аналитиков данных, исследователей и разработчиков, которые хотят работать с данными и визуализацией в одном месте.

1.5 Этапы анализа данных и использование Jupyter Notebook.

Анализ данных можно описать как процесс, состоящий из нескольких последовательных, связанных между собой шагов, в которых сырые данные преобразуются и обрабатываются с целью сделать выводы, создать визуализации или на основе математической модели построить предсказания дальнейшего развития.

Основные этапа анализа данных их можно сгруппировать в основные блоки:

– Постановка задачи.

На этом этапе определяем, что и для решения какой задачи нужно исследовать.

– Извлечение данных.

Это процесс формирования структурированного набора данных в цифровой форме. В некоторых случаях процесс сбора данных может включать также этап оцифровки. Как правило, оцифрованные данные бывают представлены в виде:

1. электронных таблиц в форматах XLS либо ODS;
2. текстовых файлов в формате CSV;
3. веб-страниц в формате HTML;
4. файлов в формате XML;
5. базы данных с доступом по технологии JSON либо через специализированный интерфейс (API).

– Подготовка данных

Этап очистки и преобразования данных. Зачастую наборы данных могут иметь разные особенности:

1. отличную от табличной форму представления;
2. пропуски отдельных данных;
3. некорректные значения;
4. текстовые данные.

Перечисленные особенности могут либо привести к затруднениям в процессе дальнейшей обработки данных, либо сделать её невозможной для чего и проводятся работы по их устранению или исправлению.

– Исследование и визуализация данных.

На этом этапе производится исследование данных, используя различные методы статистического анализа и визуализации. Процесс анализа может быть направлен на поиск связей между данными, предсказания будущих значений или определение важных факторов, влияющих на результат.

– Интерпретация результатов.

На этом этапе аналитик должен объяснить результаты анализа и проанализировать их значимость. Он должен также принять решение, основанное на этих результатах.

На этапах «Подготовка данных» и «Исследование и визуализация данных» Jupyter Notebook становиться удобных инструментом для работы. Процесс анализа данных начинается с загрузки и импортирования необходимых библиотек. В дальнейшем из источника данных, таких как CSV-файл, база данных или API, данные загружаются в Jupyter Notebook и в нем формируется датафрейм (датасет, набор данных) для дальнейшей работы. Данные просматриваются и подготавливаются к дальнейшему исследованию: удаляются ненужные столбцы, преобразуются типы данных и проводятся другие операции для подготовки данных к анализу. Используя разнообразные библиотеки, проводится анализ и визуализация данных. Используя разнообразные библиотеки, проводится анализ и визуализация данных. Формируется вывод из анализа и при необходимости даются рекомендации на основе результатов анализа данных.

В качестве завершения работы можно экспортировать результаты анализа данных из Jupyter Notebook, в формат HTML или PDF.

Глава 2. Подготовка датасета объявлений по продаже квартир в г.Ульяновск за август 2023г. на сайте Avito к анализу

2.1 Постановка задачи

В качестве задачи по анализу было решено провести анализ объявлений по продаже квартир на рынке недвижимости города Ульяновск и установить:

- Распределение объявлений по районам города;
- Зависимость цен на недвижимость от района города;
- Зависимость цен от количества комнат и площади недвижимости;
- Зависимость цен на недвижимость от этажности.

Данный анализ провести средствами Python.

2.2 Извлечение данных

Для проведения данного анализа методам социальной инженерии (через знакомства в среде риэлторов г. Ульяновск, имеющих собственную систему парсинга сайтов с объявлениями) были получены данные по объявлениям за август месяц. Выбор месяца ничем не был обусловлен – просто по дате обращения.

В качестве источника был выбран сайт Avito. Данный сайт по информации риэлторов является наиболее часто используемым и имеет более большую базу объявлении по городу Ульяновску. Были выбраны только объявления о продаже квартир так как остальные объявления (дома, коттеджи, дачи) сложнее группируются и имеют много индивидуальных различий.

2.3 Подготовка данных.

Очистка данных.

Дальнейшая работа будет проводится в Jupyter Notebook. Запускаем его и х-загружаем необходимые библиотеки Pandas, Matplotlib и Seaborn.

Полученный файл данных имеет формат csv. Экспортируем его через Pandas.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv('uln_08-23.csv', delimiter=';')
df.head()
```

	Название	Цена	Дата	Телефон	Оператор	Контактное лицо (автор объявления)	Тип автора	Регион	Город	Метро/Район	...	Источник	lat	
0	2-к. квартира, 48.5 м², 4/5 эт.	3000000	2023-08-16 07:24:33	89063935217	Вымпел-Коммуникации	недоступно	Частное лицо (фильтр)	Ульяновская область	Ульяновск	Железнодорожный	...	avito.ru	54.248197	48
1	2-к. квартира, 51 м², 3/10 эт.	4850000	2023-08-16 06:50:15	89867364237	Мобильные ТелеСистемы	Барса Агентство недвижимости	Агентство	Ульяновская область	Ульяновск	Засвияжский	...	avito.ru	54.272045	48
2	2-к. квартира, 45.8 м², 7/24 эт.	6800000	2023-08-16 06:43:37	89829002636	Мобильные ТелеСистемы	ЭТАЖИ УЛЬЯНОВСК	Агентство	Ульяновская область	Ульяновск	Засвияжский	...	avito.ru	54.308603	48
3	2-к. квартира, 52.6 м², 12/25 эт.	3999600	2023-08-16 06:31:18	89170504179	Мобильные ТелеСистемы	недоступно	Частное лицо (фильтр)	Ульяновская область	Ульяновск	Заволжский	...	avito.ru	54.350770	48
4	2-к. квартира, 41.1 м², 1/4 эт.	2240000	2023-08-16 06:18:59	89829313310	Мобильные ТелеСистемы	ЭТАЖИ УЛЬЯНОВСК	Агентство	Ульяновская область	Ульяновск	Засвияжский	...	avito.ru	54.293114	48

5 rows × 26 columns

lng	Персона для контактов	Доп.параметры	URL	Ссылки на картинки	Регион мобильного телефона	Номер подменён	Расстояние до метро, км
.306779	NaN	Площадь кухни=18| Жилая площадь=30| Тип объявлен...	https://www.avito.ru/ulyanovsk/kvartiry/2-k._k...	https://10.img.avito.st/image/1/1.6ZE96ra5RXgL...	Ульяновская область	да	неизвестно
.277710	NaN	Площадь кухни=10.3| Жилая площадь=29| Тип объявл...	https://www.avito.ru/ulyanovsk/kvartiry/2-k._k...	https://80.img.avito.st/image/1/1.sRQjjLa5Hf0V...	Ульяновская область	да	неизвестно
.358315	NaN	Площадь кухни=8| Тип объявления=Продам| Количест...	https://www.avito.ru/ulyanovsk/kvartiry/2-k._k...	https://20.img.avito.st/image/1/1.MQf52ba5ne7P...	Тюменская область	да	неизвестно
.530060	NaN	Тип объявления=Продам| Количество комнат=2|Вид ...	https://www.avito.ru/ulyanovsk/kvartiry/2-k._k...	https://40.img.avito.st/image/1/1.r8Lvpra5AyvZ...	Ульяновская область	да	неизвестно
.303338	NaN	Площадь кухни=5.5| Тип объявления=Продам| Количе...	https://www.avito.ru/ulyanovsk/kvartiry/2-k._k...	https://40.img.avito.st/image/1/1.r66zt7a5A0eF...	Тюменская область	да	неизвестно

Посмотрим информацию о нашем загруженном датасете, проверим есть ли повторяющиеся и отсутствующие данные.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 26 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Название                            4898 non-null   object
 1   Цена                                4898 non-null   int64
 2   Дата                                4898 non-null   object
 3   Телефон                             4898 non-null   int64
 4   Оператор                            4898 non-null   object
 5   Контактное лицо (автор объявления)  4897 non-null   object
 6   Тип автора                          4898 non-null   object
 7   Регион                              4898 non-null   object
 8   Город                               4898 non-null   object
 9   Метро/Район                         4853 non-null   object
 10  Адрес                               4898 non-null   object
 11  Описание                            4898 non-null   object
 12  Тип объявления                      4898 non-null   object
 13  Категория1                          4898 non-null   object
 14  Категория2                          4898 non-null   object
 15  ID на сайте                         4898 non-null   int64
 16  Источник                            4898 non-null   object
 17  lat                                 4898 non-null   float64
 18  lng                                 4898 non-null   float64
 19  Персона для контактов               0 non-null      float64
 20  Доп.параметры                       4898 non-null   object
 21  URL                                 4898 non-null   object
 22  Ссылки на картинки                  4827 non-null   object
 23  Регион мобильного телефона          4898 non-null   object
 24  Номер подменён                      4898 non-null   object
 25  Расстояние до метро, км             4898 non-null   object
dtypes: float64(3), int64(3), object(20)
memory usage: 995.0+ KB
```

```
duplicateRows = df[df.duplicated ()]
print(duplicateRows)
```

```
Empty DataFrame
Columns: [Название, Цена, Дата, Телефон, Оператор, Контактное лицо (автор объявления), Тип автора, Регион, Город, Метро/Райо
Адрес, Описание, Тип объявления, Категория1, Категория2, ID на сайте, Источник, lat, lng, Персона для контактов, Доп.парамет
URL, Ссылки на картинки, Регион мобильного телефона, Номер подменён, Расстояние до метро, км]
Index: []

[0 rows x 26 columns]
```

Как видно, набор данных, который был получен достаточно полный, пропущенные, отсутствующие значения имеются в таких колонках, как 'Персона для контактов', 'Ссылки на картинки', 'Контактное лицо (автор объявления)', но значения в них для дальнейшего анализа будут не нужны. Дубликатов данных так же не имеется. По отсутствующим данным в колонке 'Метро/Район' работу проведем позднее.

Удалим все персональные данные, данные, относящиеся к сайту Avito и ненужные данные типа 'Расстояние до метро, км' (все равно в Ульяновске нет метро).

Колонку 'Метро/Район' для большего удобства п переименуем в 'Район':

```
df=df.drop(['Телефон', 'Оператор', 'Контактное лицо (автор объявления)', 'Тип автора',
            'Тип объявления', 'Категория1', 'Категория2', 'ID на сайте', 'Источник',
            'Персона для контактов', 'URL', 'Ссылки на картинки', 'Регион мобильного телефона',
            'Номер подменён', 'Расстояние до метро, км'], axis=1)
df.rename(columns = {'Метро/Район':'Район'}, inplace = True )
df.head(0)
```

Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Доп.параметры

Посмотрим, какие типы квартир имеются в нашем наборе данных:

```
df['first'] = df['Название'].str.split(' ').str[0]
df.groupby('first').count().iloc [:, 0]
```

```
first
1-к.                    1596
2-к.                    1669
3-к.                    1125
4-к.                     199
5-к.                      16
8-к.                       1
Апартаменты-студия,        1
Аукцион:                   4
Доля                      17
Квартира-студия,         207
Своб.                     63
Name: Название, dtype: int64
```

```
df.loc[~df['Название'].astype(str).str.contains(r'^(?:[1-9])')]
```

	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Доп.параметры	first
13	Квартира-студия, 22.1 м², 16/24 эт.	2650000	2023-08-16 05:30:48	Ульяновская область	Ульяновск	Заволжский	Ульяновск	Дом сдан. Квартира с ремонтом не от застройщик...	54.350370	48.531060	Жилая площадь=13|Тип объявления=Продам|Количес...	Квартира-студия,
51	Квартира-студия, 26 м², 1/3 эт.	2250000	2023-08-16 00:43:02	Ульяновская область	Ульяновск	Заволжский	Ульяновск, улица Генерала Кашубы, 3, подъезд 1	Предлагаю Вашему вниманию квартиру - студию, р...	54.362175	48.567192	Жилая площадь=14.1|Тип объявления=Продам|Колич...	Квартира-студия,
65	Своб. планировка, 58 м², 6/9 эт.	6750000	2023-08-15 22:54:42	Ульяновская область	Ульяновск	Ленинский	Ульяновск	Видовая квартира с панорамным видом, на шестом...	54.318150	48.380470	Жилая площадь=43|Тип объявления=Продам|Количес...	Своб.
73	Квартира-студия, 30 м², 8/13 эт.	2961090	2023-08-15 21:30:37	Ульяновская область	Ульяновск	Засвияжский	Ульяновск	ДОМ СДАН Жилой комплекс ULTRAGRAD - полноценны...	54.279830	48.260370	Тип объявления=Продам|Количество комнат=Студия...	Квартира-студия,
93	Своб. планировка, 110.2 м², 1/4 эт.	8000000	2023-08-15 19:04:23	Ульяновская область	Ульяновск	Ленинский	пр-т Нариманова, 132к2	Продается новая квартира в кирпичном доме клуб...	54.359000	48.356393	Жилая площадь=90|Тип объявления=Продам|Количес...	Своб.
...	...	...	...	...	...	...	...	...	...	...	...	...
4793	Доля в 3-к. квартире, 64.4 м², 7/9 эт.	1850000	2023-07-17 09:18:35	Ульяновская область	Ульяновск	Засвияжский	ул. Корунковой, 12	Продам 1/2 долю трехкомнатной квартиры. Кварти...	54.274860	48.304847	Площадь кухни=7.2|Жилая площадь=39.7|Тип объяв...	Доля
4818	Квартира-студия, 28.6 м², 23/24 эт.	2600000	2023-07-16 22:32:18	Ульяновская область	Ульяновск	Засвияжский	ул. Александра Невского, 2И	Продам квартиру с прекрасным видом и удачным р...	54.285016	48.317100	Тип объявления=Продам|Количество комнат=Студия...	Квартира-студия,
4838	Квартира-студия, 22 м², 5/24 эт.	3320000	2023-07-16 19:27:14	Ульяновская область	Ульяновск	Засвияжский	ул. Аблукова, 18	Продам новую студию , сделан хороший, качестве...	54.310524	48.359077	Тип объявления=Продам|Количество комнат=Студия...	Квартира-студия,
4864	Своб. планировка, 51 м², 4/21 эт.	4600000	2023-07-16 13:50:29	Ульяновская область	Ульяновск	Ленинский	ул. Федерации, 130А	Продаётся квартира свободной планировки, перед...	54.337977	48.396296	Жилая площадь=47|Тип объявления=Продам|Количес...	Своб.
4880	Квартира-студия, 27.3 м², 19/19 эт.	2300000	2023-07-16 10:06:45	Ульяновская область	Ульяновск	Засвияжский	ул. Александра Невского, 2Жк1	Продам новую студию в микрорайоне "Новая жизнь...	54.286130	48.318014	Тип объявления=Продам|Количество комнат=Студия...	Квартира-студия,

292 rows × 12 columns

Удалим объявления о продаже долей в квартире и аукционы:

```
df=df[df['Название'].str.contains("Доля|Аукцион")== False]
```

Теперь займемся пропущенными данным по колонке 'Район'. У нас имеется 45 таких объявлений. Посмотрим на них:

У нас имеется 45 таких объявлений. Посмотрим на них:

69	1-к. квартира, 32 м², 2/2 эт.	550000	2023-08-15 22:43:45	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам квартиру. Срочно. Собственник. Частично...	54.197311	48.100554	Площадь кухни=8.5| Жилая площадь=16| Тип объявле...	1-к.
317	1-к. квартира, 32 м², 1/2 эт.	1500000	2023-08-14 16:07:48	Ульяновская область	Ульяновск	NaN	Ульяновская область, Чердаклинский р-н, Красно...	Квартира в отличном состоянии, раздельный сану...	54.322673	48.565017	Площадь кухни=7| Жилая площадь=16| Тип объявлени...	1-к.
425	1-к. квартира, 35.5 м², 1/2 эт.	700000	2023-08-14 00:35:22	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продам 1-комн.квартиру , кухня гарнитур, свое ...	54.398603	48.105619	Площадь кухни=9|Тип объявления=Продам| Количест...	1-к.
642	1-к. квартира, 30 м², 1/2 эт.	575000	2023-08-12 19:53:26	Ульяновская область	Ульяновск	NaN	Ульяновская область, г.о. Новоульяновск, пос. ...	Все объекты инфраструктуры рядом, 15 км от гор...	54.117524	48.327803	Площадь кухни=6| Жилая площадь=16| Тип объявлени...	1-к.
718	1-к. квартира, 32.4 м², 1/2 эт.	550000	2023-08-12 11:26:01	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Квартира находится в центральной части, тёплая...	54.160465	48.021294	Площадь кухни=8.8| Жилая площадь=15.8| Тип объяв...	1-к.
727	1-к. квартира, 39.9 м², 3/3 эт.	800000	2023-08-12 10:13:52	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Продам квартиру в посёлке Зелёная Роща, от гор...	54.155156	48.016892	Площадь кухни=7.5| Жилая площадь=19| Тип объявле...	1-к.
921	1-к. квартира, 30 м², 2/2 эт.	1270000	2023-08-11 07:55:06	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Посёлок находится не далеко от города (18км). ...	54.426273	48.170810	Площадь кухни=5.8| Тип объявления=Продам| Количе...	1-к.
1009	3-к. квартира, 60 м², 1/2 эт.	1599999	2023-08-10 14:48:58	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продается квартира в Тихом, экологически чисто...	54.407255	48.207075	Площадь кухни=7| Жилая площадь=45| Тип объявлени...	3-к.
1036	3-к. квартира, 70 м², 1/1 эт.	900000	2023-08-10 09:44:14	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продам квартиру в доме. Есть гараж, сарай (кто...	54.397125	48.106418	Площадь кухни=16| Жилая площадь=50| Тип объявлен...	3-к.
1102	3-к. квартира, 60.9 м², 1/2 эт.	1000000	2023-08-09 23:42:14	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н	продается 3-х комнатная квартира в п Зеленая Р...	54.164732	48.003711	Площадь кухни=7.8| Жилая площадь=42.4| Тип объяв...	3-к.
1136	2-к. квартира, 55.9 м², 1/2 эт.	1099000	2023-08-09 19:44:31	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Доступное жилье - старт для семейной жизни.Ваш...	54.197194	48.100606	Площадь кухни=7.4| Жилая площадь=27.8| Тип объяв...	2-к.
1195	2-к. квартира, 47 м², 2/2 эт.	850000	2023-08-09 13:36:10	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, пос. Зел...	Удобное расположение, рядом (в шаговой доступн...	54.160465	48.021294	Площадь кухни=7|Тип объявления=Продам| Количест...	2-к.
1657	1-к. квартира, 27 м², 1/2 эт.	800000	2023-08-06 18:44:34	Ульяновская область	Ульяновск	NaN	Ульяновск	Однокомнатная квартира + сарай с погребом + 1,...	54.424660	48.171187	Площадь кухни=5|Тип объявления=Продам| Количест...	1-к.
1770	2-к. квартира, 51 м², 1/2 эт.	700000	2023-08-05 15:55:38	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, пос. Зел...	Разумный торг. Для реального покупателя.	54.156342	48.017395	Площадь кухни=10| Тип объявления=Продам| Количес...	2-к.
1969	3-к. квартира, 66 м², 3/3 эт.	1280000	2023-08-04 04:31:49	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, пос. Зел...	Продам уютную, теплую и светлую квартиру, не д...	54.154488	48.017144	Площадь кухни=11|Тип объявления=Продам| Количес...	3-к.
2055	3-к. квартира, 60 м², 1/2 эт.	1350000	2023-08-03 18:18:12	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Очень удобная квартира для того кто понимает ,...	54.423599	48.142384	Площадь кухни=10| Тип объявления=Продам| Количес...	3-к.
2150	2-к. квартира, 50.4 м², 2/2 эт.	1090000	2023-08-03 04:33:16	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Данный актив продается посредством онлайн аукц...	54.406522	48.209195	Площадь кухни=12.5| Жилая площадь=28| Тип объявл...	2-к.
2230	2-к. квартира, 43 м², 1/2 эт.	1600000	2023-08-02 11:18:44	Ульяновская область	Ульяновск	NaN	Ульяновская область, Чердаклинский р-н, Красно...	Продается уютная двухкомнатная квартира, идеал...	54.322256	48.572471	Площадь кухни=8|Тип объявления=Продам| Количест...	2-к.
2333	3-к. квартира, 65.4 м², 1/9 эт.	4299000	2023-08-01 11:59:07	Ульяновская область	Ульяновск	NaN	Ульяновский пр-т, 18	ПОСМОТРИТЕ КВАРТИРУ!!! Продается солнечная, тё...	54.378088	48.579597	Площадь кухни=8.4| Тип объявления=Продам| Количе...	3-к.

2496	2-к. квартира, 45.1 м², 1/5 эт.	1000000	2023-07-31 14:45:02	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам квартиру в селе Тетюшском 45 квадратов ...	54.312114	48.014390	Площадь кухни=4|Тип объявления=Продам| Количест...	2-к.
2566	2-к. квартира, 38.9 м², 2/2 эт.	950000	2023-07-31 12:35:57	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Идеальное сочетание городского комфорта и заго...	54.162563	48.020764	Площадь кухни=5.2| Жилая площадь=20.9| Тип объяв...	2-к.
2725	2-к. квартира, 40 м², 2/2 эт.	900000	2023-07-30 19:04:56	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Продам уютную 2-комнатную квартиру в п.Красноа...	54.226069	48.092727	Площадь кухни=6|Тип объявления=Продам| Количест...	2-к.
2815	3-к. квартира, 77 м², 1/1 эт.	480000	2023-07-29 14:56:00	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам половину кирпичного дома на центральной...	54.399944	47.960732	Площадь кухни=11|Тип объявления=Продам| Количес...	3-к.
2903	2-к. квартира, 48 м², 1/2 эт.	1300000	2023-07-28 21:27:39	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам 2х-комнатную квартиру, имеется водонагр...	54.312476	48.041856	Площадь кухни=8|Тип объявления=Продам| Количест...	2-к.
2913	2-к. квартира, 44.6 м², 2/2 эт.	1400000	2023-07-28 19:59:48	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продам 2-комнатную квартиру на 2 этаже 2-х эт...	54.425687	48.170998	Площадь кухни=5.8| Жилая площадь=38.8| Тип объяв...	2-к.
3116	1-к. квартира, 29.8 м², 2/2 эт.	800000	2023-07-27 10:57:07	Ульяновская область	Ульяновск	NaN	Ульяновский район, Тимирязевское сельское посе...	Однокомнатная квартира,в санузле сделан ремонт...	54.407381	48.208853	Площадь кухни=2.7| Тип объявления=Продам| Количе...	1-к.
3211	2-к. квартира, 46.1 м², 2/2 эт.	1050000	2023-07-26 23:23:00	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Вашему вниманию предлагается двухкомнатная ква...	54.225800	48.093111	Площадь кухни=6| Жилая площадь=30.1| Тип объявле...	2-к.
3238	3-к. квартира, 63.3 м², 10/10 эт.	3799000	2023-07-26 17:23:14	Ульяновская область	Ульяновск	NaN	ул. Рябикова, 116	Продам уютную светлую 3х комнатную квартиру на...	54.267336	48.281096	Площадь кухни=8| Жилая площадь=40| Тип объявлени...	3-к.
3240	2-к. квартира, 37.9 м², 2/2 эт.	800000	2023-07-26 16:23:39	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам уютную квартиру в посёлке Станция Лаише...	54.407292	48.208224	Площадь кухни=7|Тип объявления=Продам| Количест...	2-к.
3610	2-к. квартира, 44.4 м², 2/2 эт.	1300000	2023-07-24 16:21:06	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам 2х комнатную квартиру, в хорошем состоя...	54.312217	48.019039	Площадь кухни=7.6| Жилая площадь=27| Тип объявле...	2-к.
3898	3-к. квартира, 68 м², 2/2 эт.	3800000	2023-07-22 20:04:41	Ульяновская область	Ульяновск	NaN	Ульяновская область, Чердаклинский р-н, Мирнов...	В шаговой доступности школа,детский сад,ФАП,ма...	54.428897	48.656080	Площадь кухни=12| Жилая площадь=53| Тип объявлен...	3-к.
3927	2-к. квартира, 62 м², 2/2 эт.	2850000	2023-07-22 15:45:10	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продаётся уютная и светлая 2-х комн. квартира ...	54.428887	48.175319	Площадь кухни=7.9| Жилая площадь=36.4| Тип объяв...	2-к.
3941	3-к. квартира, 47.8 м², 5/5 эт.	870000	2023-07-22 14:22:32	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам 3х комнатную квартиру. Площадь 54 кв.м....	54.312248	48.016111	Площадь кухни=6|Тип объявления=Продам| Количест...	3-к.
3969	1-к. квартира, 30.1 м², 1/2 эт.	400000	2023-07-22 10:22:43	Ульяновская область	Ульяновск	NaN	Ульяновская область, г.о. Новоульяновск, пос. ...	ПРОДАЖА ОТ СОБСТВЕННИКА! Продается 1-комнатная...	54.118046	48.327576	Площадь кухни=5| Жилая площадь=16| Тип объявлени...	1-к.
4215	2-к. квартира, 52 м², 9/10 эт.	4500000	2023-07-20 19:09:33	Ульяновская область	Ульяновск	NaN	Ульяновск, улица Генерала Мельникова, 14	Продам 2х комнатную квартиру. Автономное отопл...	54.277641	48.268158	Площадь кухни=11|Тип объявления=Продам| Количес...	2-к.
4222	2-к. квартира, 58 м², 1/1 эт.	1550000	2023-07-20 16:19:44	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продается двухкомнатная квартира, 58 кв.м., со...	54.424613	48.176137	Площадь кухни=12| Жилая площадь=36| Тип объявлен...	2-к.
4377	3-к. квартира, 58.2 м², 2/2 эт.	1850000	2023-07-19 18:55:43	Ульяновская область	Ульяновск	NaN	Ульяновск	Продаётся квартира. В экологически чистом мест...	54.363790	48.129388	Площадь кухни=5.7| Тип объявления=Продам| Количе...	3-к.
4395	1-к. квартира, 28 м², 2/2 эт.	520000	2023-07-19 15:52:27	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Зеленоро...	Дом расположен в детском доме Имени А.Матросов...	54.196683	48.100948	Площадь кухни=4.7| Жилая площадь=23| Тип объявле...	1-к.
4433	1-к. квартира, 32 м², 2/2 эт.	1350000	2023-07-19 11:27:28	Ульяновская область	Ульяновск	NaN	Ульяновская область, Майнский р-н, Тагайское с...	Продам квартиру ,цена договорная	54.290707	47.845045	Площадь кухни=13| Жилая площадь=19| Тип объявлен...	1-к.
4466	1-к. квартира, 33 м², 2/2 эт.	400000	2023-07-19 08:26:31	Ульяновская область	Ульяновск	NaN	Ульяновская область, Майнский р-н, Тагайское с...	Однокомнатная квартира.под ремонт.хоз постройк...	54.291485	47.837750	Площадь кухни=9| Жилая площадь=16| Тип объявлени...	1-к.

4516	1-к. квартира, 34.8 м², 3/3 эт.	760000	2023-07-18 22:20:18	Ульяновская область	Ульяновск	NaN	Ульяновская область, г.о. Новоульяновск, пос. ...	Квартира в идеальном состоянии, один собственн...	54.115887	48.326073	Площадь кухни=8.8| Жилая площадь=16| Тип объявле...	1-к.
4517	2-к. квартира, 59 м², 1/1 эт.	2000000	2023-07-18 21:46:22	Ульяновская область	Ульяновск	NaN	Ульяновск	Продается квартира в двухквартирном доме , вхо...	54.423639	48.123882	Площадь кухни=19| Жилая площадь=40| Тип объявлен...	2-к.
4699	2-к. квартира, 37.9 м², 2/2 эт.	800000	2023-07-17 20:36:23	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тимирязе...	Продам уютную квартиру в посёлке Станция Лаише...	54.407266	48.208284	Площадь кухни=7| Жилая площадь=30.9| Тип объявле...	2-к.
4729	1-к. квартира, 31.8 м², 4/5 эт.	730000	2023-07-17 16:35:45	Ульяновская область	Ульяновск	NaN	Ульяновская область, Ульяновский р-н, Тетюшско...	Продам однокомнатную квартиру в с. Тетюшское. ...	54.312117	48.015132	Площадь кухни=5.7| Жилая площадь=17.9| Тип объяв...	1-к.
4862	3-к. квартира, 53.3 м², 1/2 эт.	670000	2023-07-16 14:30:32	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам 3-х комнатную квартиру в Ульяновском ра...	54.399861	48.104954	Площадь кухни=7|Тип объявления=Продам| Количест...	3-к.

Сразу же видно, что в набор данных попали объявления не только по городу Ульяновску, но и по различным населенным пунктам Ульяновской области. Удалим все объявления, где адрес содержит 'Ульяновская область':

```
df=df[df['Адрес'].str.contains('Ульяновская область,')== False]
```

Выведем оставшиеся объявления с пустым значением 'Район':

```
df[df['Район'].isnull()]
```

	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Д
69	1-к. квартира, 32 м², 2/2 эт.	550000	2023-08-15 22:43:45	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам квартиру. Срочно. Собственник. Частично...	54.197311	48.100554	Плоі Жил
1657	1-к. квартира, 27 м², 1/2 эт.	800000	2023-08-06 18:44:34	Ульяновская область	Ульяновск	NaN	Ульяновск	Однокомнатная квартира + сарай с погребом + 1,...	54.424660	48.171187	Площ объяв
2333	3-к. квартира, 65.4 м², 1/9 эт.	4299000	2023-08-01 11:59:07	Ульяновская область	Ульяновск	NaN	Ульяновский пр-т, 18	ПОСМОТРИТЕ КВАРТИРУ!!! Продается солнечная, тё...	54.378088	48.579597	Плоі объяв
3116	1-к. квартира, 29.8 м², 2/2 эт.	800000	2023-07-27 10:57:07	Ульяновская область	Ульяновск	NaN	Ульяновский район, Тимирязевское сельское посе...	Однокомнатная квартира,в санузле сделан ремонт...	54.407381	48.208853	Плоі объяв
3238	3-к. квартира, 63.3 м², 10/10 эт.	3799000	2023-07-26 17:23:14	Ульяновская область	Ульяновск	NaN	ул. Рябикова, 116	Продам уютную светлую 3х комнатную квартиру на...	54.267336	48.281096	Пл Жил Т
3240	2-к. квартира, 37.9 м², 2/2 эт.	800000	2023-07-26 16:23:39	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам уютную квартиру в посёлке Станция Лаише...	54.407292	48.208224	Площ объяв
4215	2-к. квартира, 52 м², 9/10 эт.	4500000	2023-07-20 19:09:33	Ульяновская область	Ульяновск	NaN	Ульяновск, улица Генерала Мельникова, 14	Продам 2х комнатную квартиру. Автономное отопл...	54.277641	48.268158	Площа объяв
4377	3-к. квартира, 58.2 м², 2/2 эт.	1850000	2023-07-19 18:55:43	Ульяновская область	Ульяновск	NaN	Ульяновск	Продаётся квартира. В экологически чистом мест...	54.363790	48.129388	Плоі объяв
4517	2-к. квартира, 59 м², 1/1 эт.	2000000	2023-07-18 21:46:22	Ульяновская область	Ульяновск	NaN	Ульяновск	Продается квартира в двухквартирном доме , вхо...	54.423639	48.123882	Плс Жил
4862	3-к. квартира, 53.3 м², 1/2 эт.	670000	2023-07-16 14:30:32	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам 3-х комнатную квартиру в Ульяновском ра...	54.399861	48.104954	Площ объяв

Теперь осталось 10 таких объявлений, их можно просмотреть и поправить "вручную". У трех объявлений есть адрес и по названию улицы можно определить и внести название района.

Оставшиеся семь объявлений изучим дополнительно:

```
df.at [2333, 'Район'] = 'Заволжский'
df.at [4215, 'Район'] = 'Засвияжский'
df.at [3238, 'Район'] = 'Ленинский'

df[df['Район'].isnull()]
```

	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat
69	1-к. квартира, 32 м², 2/2 эт.	550000	2023-08-15 22:43:45	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам квартиру. Срочно. Собственник. Частично...	54.197311
1657	1-к. квартира, 27 м², 1/2 эт.	800000	2023-08-06 18:44:34	Ульяновская область	Ульяновск	NaN	Ульяновск	Однокомнатная квартира + сарай с погребом + 1,...	54.424660
3116	1-к. квартира, 29.8 м², 2/2 эт.	800000	2023-07-27 10:57:07	Ульяновская область	Ульяновск	NaN	Ульяновский район, Тимирязевское сельское посе...	Однокомнатная квартира,в санузле сделан ремонт...	54.407381
3240	2-к. квартира, 37.9 м², 2/2 эт.	800000	2023-07-26 16:23:39	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам уютную квартиру в посёлке Станция Лаише...	54.407292
4377	3-к. квартира, 58.2 м², 2/2 эт.	1850000	2023-07-19 18:55:43	Ульяновская область	Ульяновск	NaN	Ульяновск	Продаётся квартира. В экологически чистом мест...	54.363790
4517	2-к. квартира, 59 м², 1/1 эт.	2000000	2023-07-18 21:46:22	Ульяновская область	Ульяновск	NaN	Ульяновск	Продается квартира в двухквартирном доме , вхо...	54.423639
4862	3-к. квартира, 53.3 м², 1/2 эт.	670000	2023-07-16 14:30:32	Ульяновская область	Ульяновск	NaN	Ульяновск	Продам 3-х комнатную квартиру в Ульяновском ра...	54.399861

Видно, что у всех есть географические координаты - проверим координаты из оставшихся объявлений через какой-нибудь геосервис. Я воспользовался https://geotree.ru.

Все эти семь объявлений касались квартир, находящихся в области и не относящихся к городу Ульяновск, поэтому удаляем их из нашего датасета.

```
df=df.dropna()
df[df['Район'].isnull()]
```

Преобразование данных.

Этап очистки окончен и теперь переходим к преобразованию данных. В объявлениях все данные квартиры записаны в одном текстовом поле. С помощью

сплит-оператора выделим из этой строки данные о площади, этаже и количеству комнат в отдельные колонки. Для этого создаем временный датасет – new_df.

```
new_df = df['Название'].str.split(',',expand=True)
new_df.columns=['квартира','площадь','этаж/высота дома']
new_df['этаж'] = new_df['этаж/высота дома'].str.split('/').str[0].astype(int)
new_df['кол-во комнат'] = new_df['квартира'].str[0]
new_df
```

	квартира	площадь	этаж/высота дома	этаж	кол-во комнат
0	2-к. квартира	48.5 м²	4/5 эт.	4	2
1	2-к. квартира	51 м²	3/10 эт.	3	2
2	2-к. квартира	45.8 м²	7/24 эт.	7	2
3	2-к. квартира	52.6 м²	12/25 эт.	12	2
4	2-к. квартира	41.1 м²	1/4 эт.	1	2
...	...	...	...	...	...
4893	1-к. квартира	28.9 м²	2/2 эт.	2	1
4894	3-к. квартира	77.9 м²	5/12 эт.	5	3
4895	1-к. квартира	38 м²	4/9 эт.	4	1
4896	1-к. квартира	40 м²	1/24 эт.	1	1
4897	2-к. квартира	56.5 м²	6/19 эт.	6	2

4775 rows × 5 columns

Так как в наборе данных находились объявления "Апартаменты-студия", "Квартира-студия", "Своб." - о квартирах свободной планировки (без определенного количества комнат), в колонке количества комнат появились значения "А", "К", "С". Заменим эти буквы на 0.

```
new_df['кол-во комнат'] = new_df['кол-во комнат'].replace('С','0').replace('А','0').replace('К','0').astype(int
new_df.sort_values(by=['кол-во комнат'])
```

	квартира	площадь	этаж/высота дома	этаж	кол-во комнат
340	Своб. планировка	44 м²	6/7 эт.	6	0
371	Своб. планировка	44 м²	5/7 эт.	5	0
370	Своб. планировка	44 м²	3/7 эт.	3	0
369	Своб. планировка	44 м²	6/7 эт.	6	0
368	Своб. планировка	106.9 м²	3/4 эт.	3	0
...	...	...	...	...	...
2106	5-к. квартира	102.7 м²	4/9 эт.	4	5
4220	5-к. квартира	115 м²	1/3 эт.	1	5
4027	5-к. квартира	83 м²	2/9 эт.	2	5
1829	5-к. квартира	100 м²	1/2 эт.	1	5
529	8-к. квартира	195.8 м²	8/9 эт.	8	8

4775 rows × 5 columns

Колонку с указанием площади из строкового формата переведем в числовой:

```
new_df['площадь'] = new_df['площадь'].str.replace('м²','')
```

```
new_df.sort_values(by=['площадь'])
```

	квартира	площадь	этаж/высота дома	этаж	кол-во комнат
2375	3-к. квартира	100	8/9 эт.	8	3
2263	3-к. квартира	100	3/3 эт.	3	3
3196	4-к. квартира	100	1/9 эт.	1	4
1088	3-к. квартира	100	2/3 эт.	2	3
3822	Квартира-студия	100	17/17 эт.	17	0
...	...	...	...	...	...
2752	4-к. квартира	99	1/16 эт.	1	4
2435	4-к. квартира	99	7/8 эт.	7	4
647	3-к. квартира	99.2	5/10 эт.	5	3
1140	3-к. квартира	99.6	2/10 эт.	2	3
2954	4-к. квартира	99.8	3/8 эт.	3	4

4775 rows × 5 columns

```
new_df['площадь'] = new_df['площадь'].astype(float)
```

Проверим результат преобразований:

```
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4775 entries, 0 to 4897
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   квартира          4775 non-null   object
 1   площадь           4775 non-null   float64
 2   этаж/высота дома  4775 non-null   object
 3   этаж              4775 non-null   int32
 4   кол-во комнат     4775 non-null   int32
dtypes: float64(1), int32(2), object(2)
memory usage: 186.5+ KB
```

Теперь имеются колонки “площадь”, “этаж” и ”кол-во комнат” с числовыми типами значений, которые можно использовать в анализе.

Результаты наших преобразований из датасета new_df и первоначальный датасет df сводим в новый датасет - df_realty, который и будет использоваться в дальнейшей работе по проведению анализа объявлений.

```
df_realty = pd.concat([df,new_df],axis=1)
df_realty
```

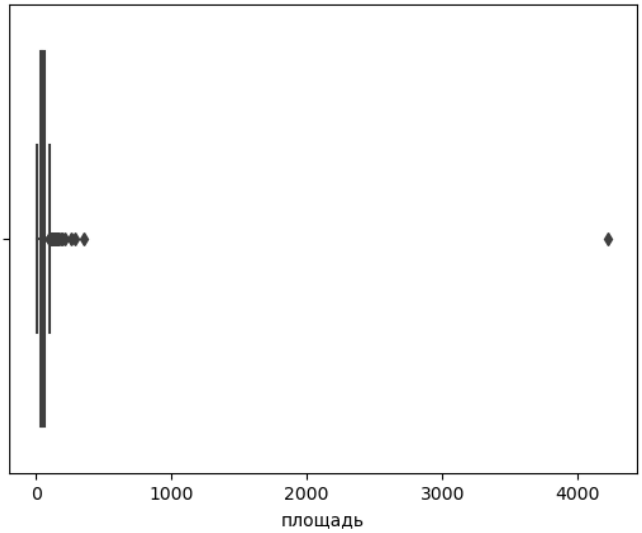
	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat
0	2-к. квартира, 48.5 м², 4/5 эт.	3000000	2023-08-16 07:24:33	Ульяновская область	Ульяновск	Железнодорожный	Профсоюзная ул., 38	Опытное поле.Отличный вариант расположения.Сту...	54.248197
1	2-к. квартира, 51 м², 3/10 эт.	4850000	2023-08-16 06:50:15	Ульяновская область	Ульяновск	Засвияжский	Отрадная ул., 83	Предлагаю Вашему вниманию двухкомнатную кварти...	54.272045
2	2-к. квартира, 45.8 м², 7/24 эт.	6800000	2023-08-16 06:43:37	Ульяновская область	Ульяновск	Засвияжский	ул. Аблукова, 4	СПЕШИТЕ КУПИТЬ КВАРТИРУ МЕЧТЫ! Продается прекр...	54.308603
3	2-к. квартира, 52.6 м², 12/25 эт.	3999600	2023-08-16 06:31:18	Ульяновская область	Ульяновск	Заволжский	Ульяновск	В ЖК «Сиреневый» сочетаются современная комфор...	54.350770
4	2-к. квартира, 41.1 м², 1/4 эт.	2240000	2023-08-16 06:18:59	Ульяновская область	Ульяновск	Засвияжский	ул. Ефремова, 15	УСПЕЙ КУПИТЬ! Продается 2х комнатная квартира ...	54.293114
...	...	...	...	...	...	...	...	...	...
4893	1-к. квартира, 28.9 м², 2/2 эт.	1550000	2023-07-16 07:21:36	Ульяновская область	Ульяновск	Заволжский	ул. Жуковского, 89	Продам полногабаритную уютную квартиру в добро...	54.349949
4894	3-к. квартира, 77.9 м², 5/12 эт.	5608080	2023-07-16 06:48:46	Ульяновская область	Ульяновск	Засвияжский	Ульяновск	ЖК ЯСНОВО По-новому комфортный парк-квартал на...	54.283350
4895	1-к. квартира, 38 м², 4/9 эт.	2890000	2023-07-16 06:44:38	Ульяновская область	Ульяновск	Заволжский	Новосондецкий б-р, 18	Продается светлая, теплая однокомнатная кварти...	54.381696
4896	1-к. квартира, 40 м², 1/24 эт.	3150000	2023-07-16 06:44:35	Ульяновская область	Ульяновск	Железнодорожный	ул. Кирова, 6/2	Продаётся однокомнатная квартира в Железнодоро...	54.302330
4897	2-к. квартира, 56.5 м², 6/19 эт.	4977110	2023-07-16 00:12:01	Ульяновская область	Ульяновск	Засвияжский	Ульяновск	Монолитный дом повышенной комфортности в знако...	54.278770

4775 rows × 17 columns

Проверим данные на выбросы в значениях.

Площадь:

```
sns.boxplot(x=df_realty['площадь']);
```



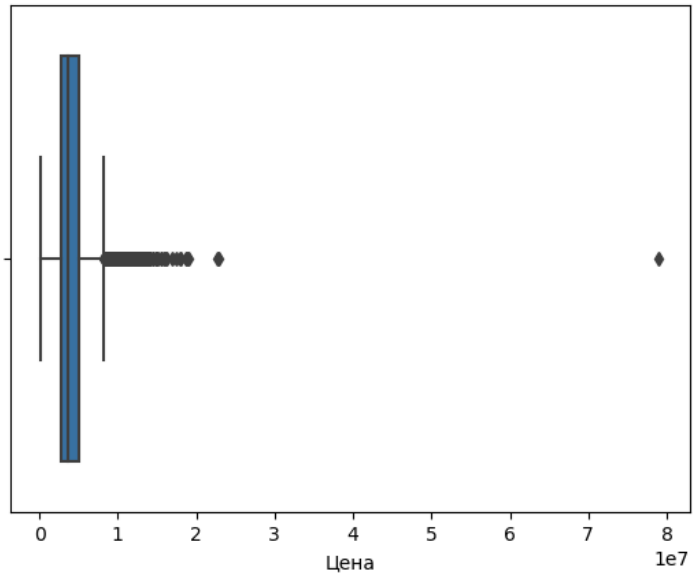

```
df_realty[df_realty['площадь']>1000]
```

Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Доп.параметры	first	квартира	площадь	вы
2-к. квартира, 4224 м², 1/5 эт.	3100000	2023-08-12 21:21:03	Ульяновская область	Ульяновск	Ленинский	ул. Докучаева, 16	Магазины рядом Гулливер., Магнит, Пятёрочка, к...	54.347504	48.386567	Площадь кухни=6.1| Жилая площадь=42| Тип объявле...	2-к.	2-к. квартира	4224.0	1

Имеется одна квартира с невозможной площадью в 4224 кв.метра, но так как в текстово описании другая площадь, перед нами ошибка ввода, должно быть 42,24 кв.метра.

Цена:

```
sns.boxplot(x=df_realty['Цена']);
```



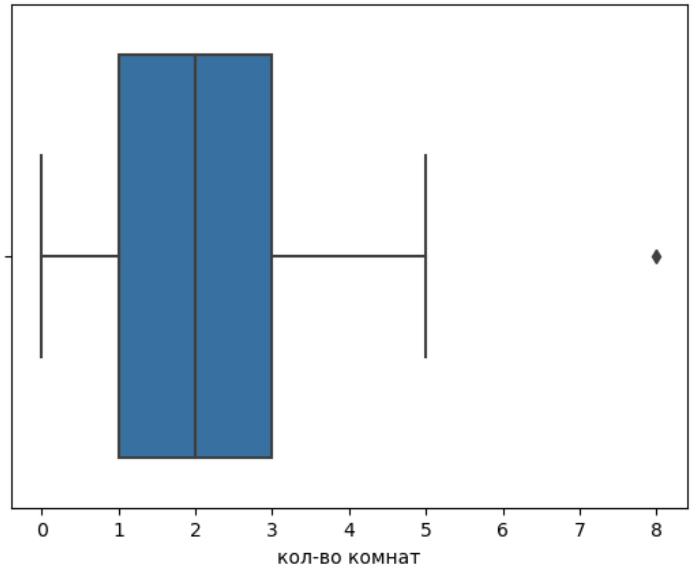
```
df_realty[df_realty['Цена']>50000000]
```

	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Доп.параметры	first	квартира
2777	4-к. квартира, 75 м², 5/9 эт.	79000000	2023-07-29 21:56:14	Ульяновская область	Ульяновск	Заволжский	Фестивальный б-р, 3	Продаётся 4 комнатная квартира в новом городе	54.369323	48.579485	Площадь кухни=8.1| Тип объявления=Продам| Количе...	4-к.	4-к квартира

Имеется одна квартира с достаточно стандартными показателями, но ценой в 79 млн. рублей, что для Ульяновска невозможно – возможно ошибка в данных – нормальная цена этой квартиры 7,9 млн. рублей

Количество комнат:

```
sns.boxplot(x=df_realty['кол-во комнат']);
```




```
df_realty[df_realty['кол-во комнат']>5]
```

	Название	Цена	Дата	Регион	Город	Район	Адрес	Описание	lat	lng	Доп.параметры	first	квартира	п
529	8-к. квартира, 195.8 м², 8/9 эт.	229840	2023-08-13 08:07:15	Ульяновская область	Ульяновск	Засвияжский	ул. Аблукова, 45	В процедуре исполнительного производства с тор...	54.314712	48.361532	Площадь кухни=20|Жилая площадь=60| Тип объявлен...	8-к.	8-к. квартира	

Есть одна квартира с 8-ю комнатами, но учитывая, что общая площадь 195,8 кв.метра – данные корректны. Такая квартира может быть в наличии.

Данные двух квартир с ошибками в площади и цене исправим:

```
df_realty['площадь']=df_realty['площадь'].replace(4224, 42.24)
df_realty['Цена']=df_realty['Цена'].replace(79000000, 7900000)
```

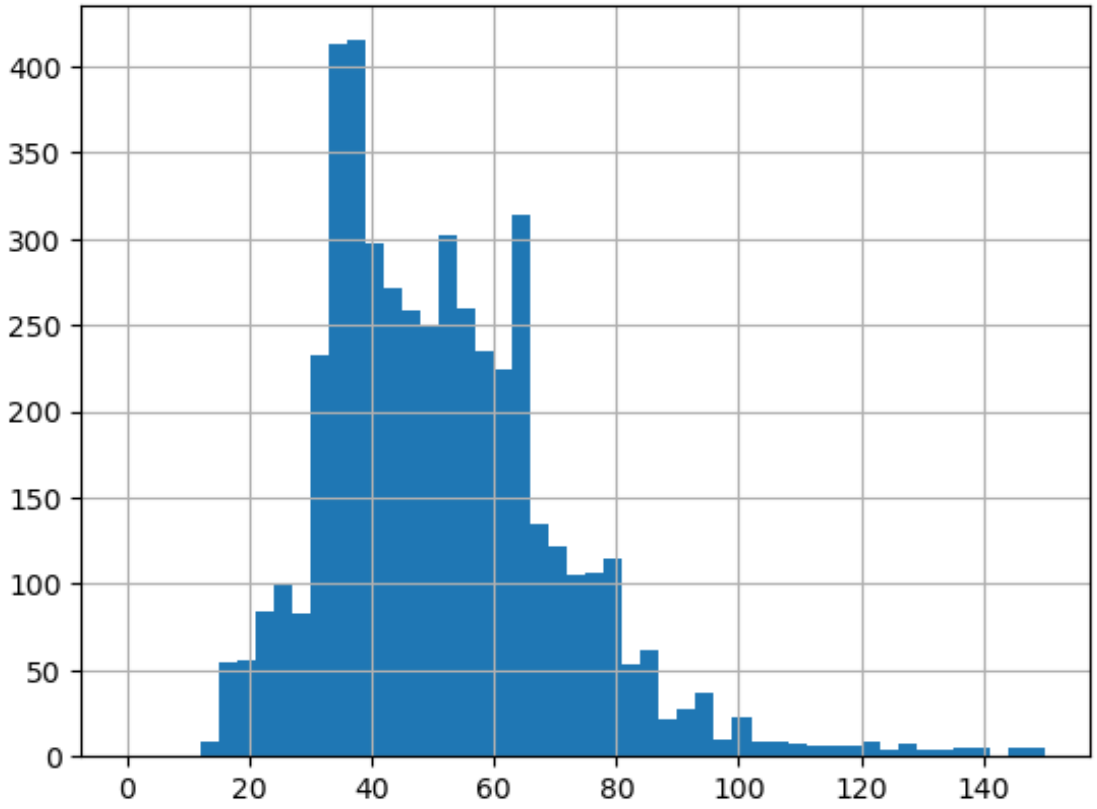
2.4 Исследование и визуализация данных.

Посмотрим на распределения количества объявлений по разным показателям. По площади квартир:

```
def draw_hist(col, xmin, xmax):
    df_realty[col].hist(bins=50, range=(xmin, xmax))
```

```
def draw_boxplot(col, ymin=-50, ymax=200):
    plt.ylim(ymin, ymax)
    df_realty.boxplot(col)
```

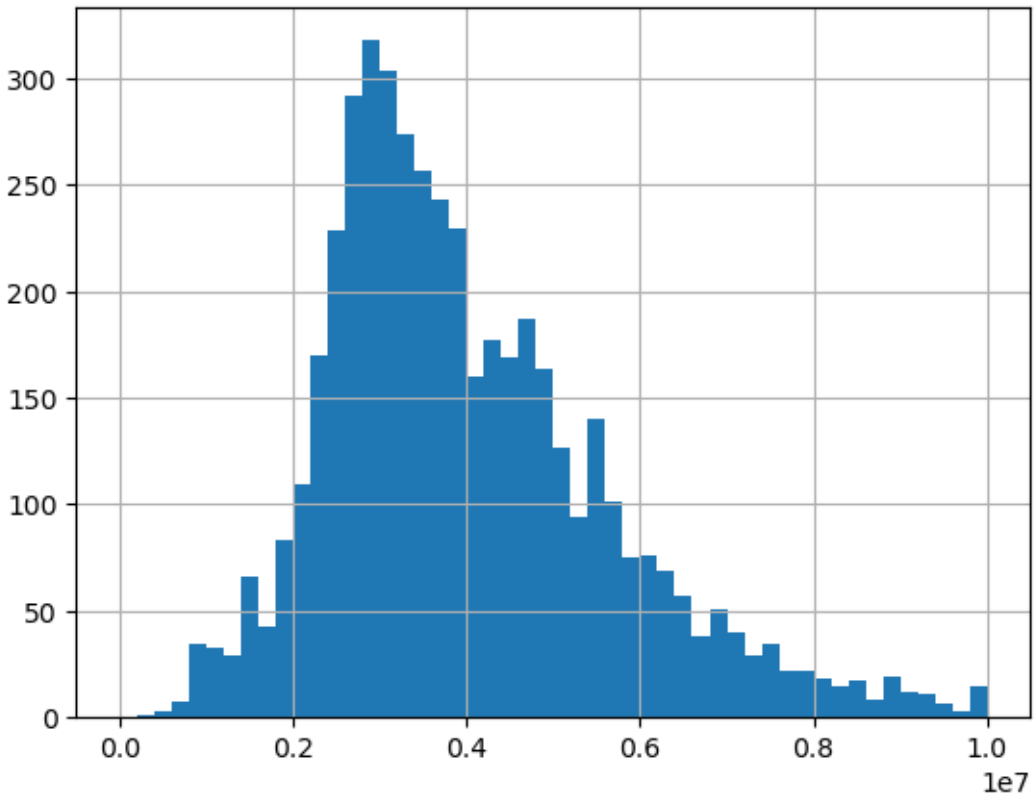
```
draw_hist('площадь', 0, 150)
```



Преобладают квартиры площадью чуть ниже 40 кв.м.. Есть всплеск на примерно 55 и 65 кв.м.

По цене:

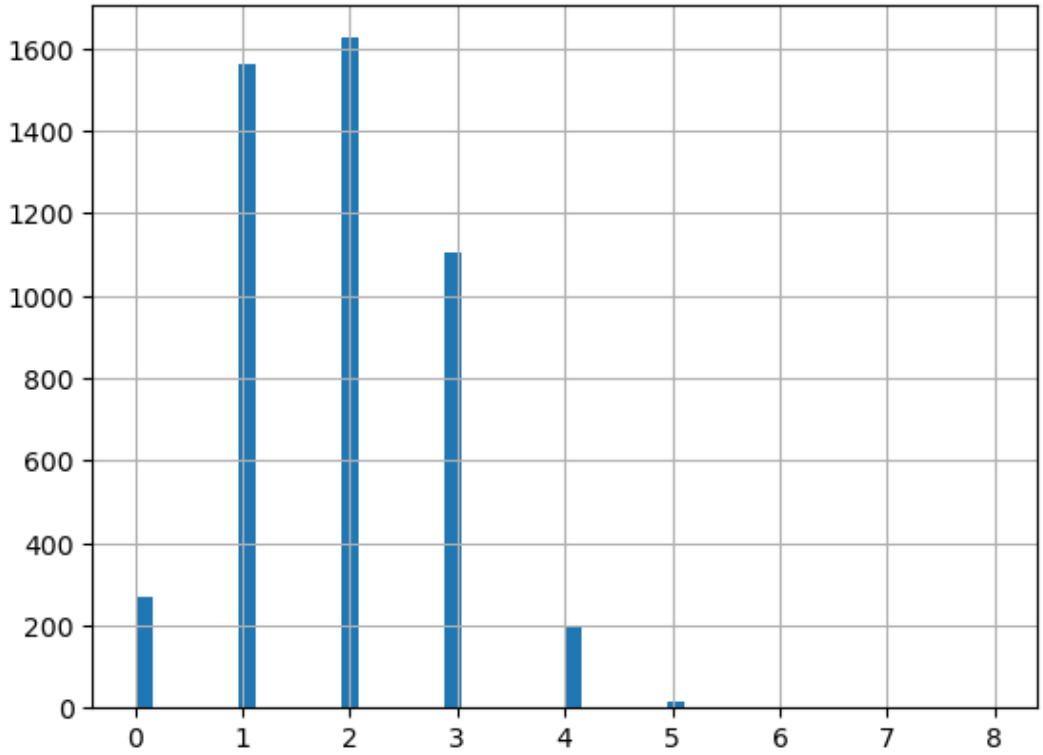
```
draw_hist('Цена', 0, 10000000)
```



Преобладают квартиры в ценовом диапазоне от 2,5 до 4 млн. руб.

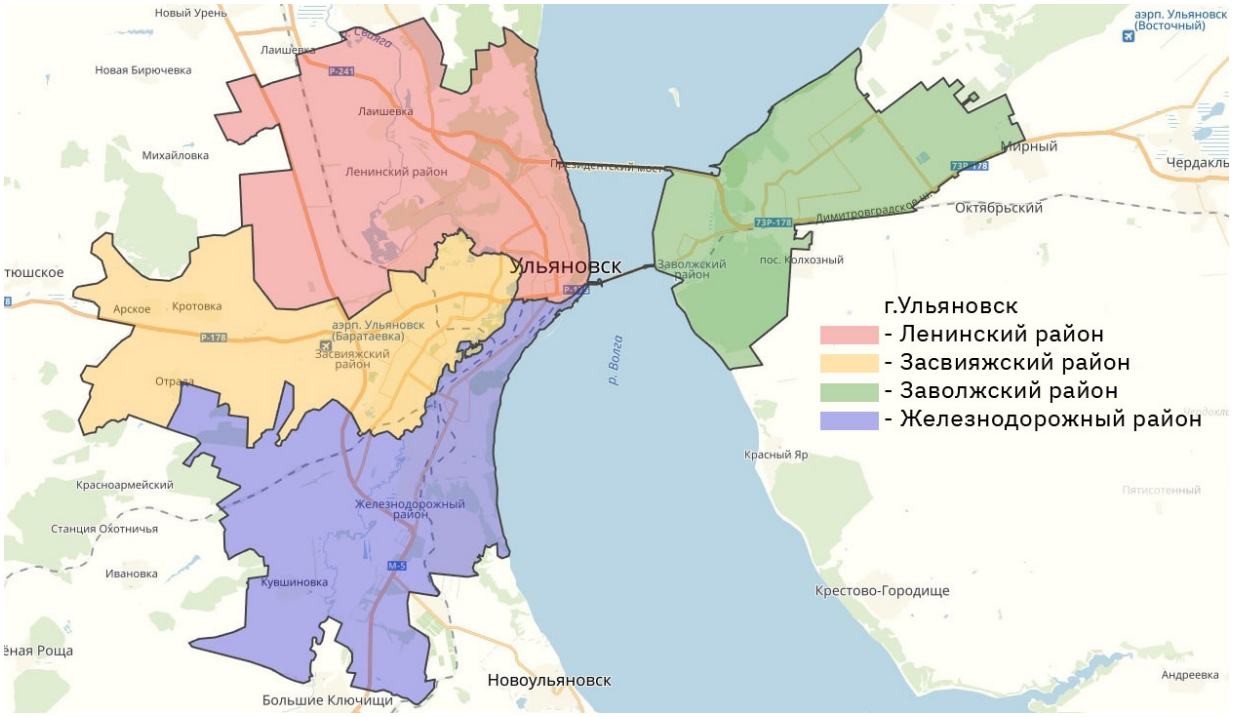
По количеству комнат:

```
: draw_hist('кол-во комнат', 0, 8)
```



Здесь явно видно, что преобладают объявления о продаже одно- и двух комнатных квартир.

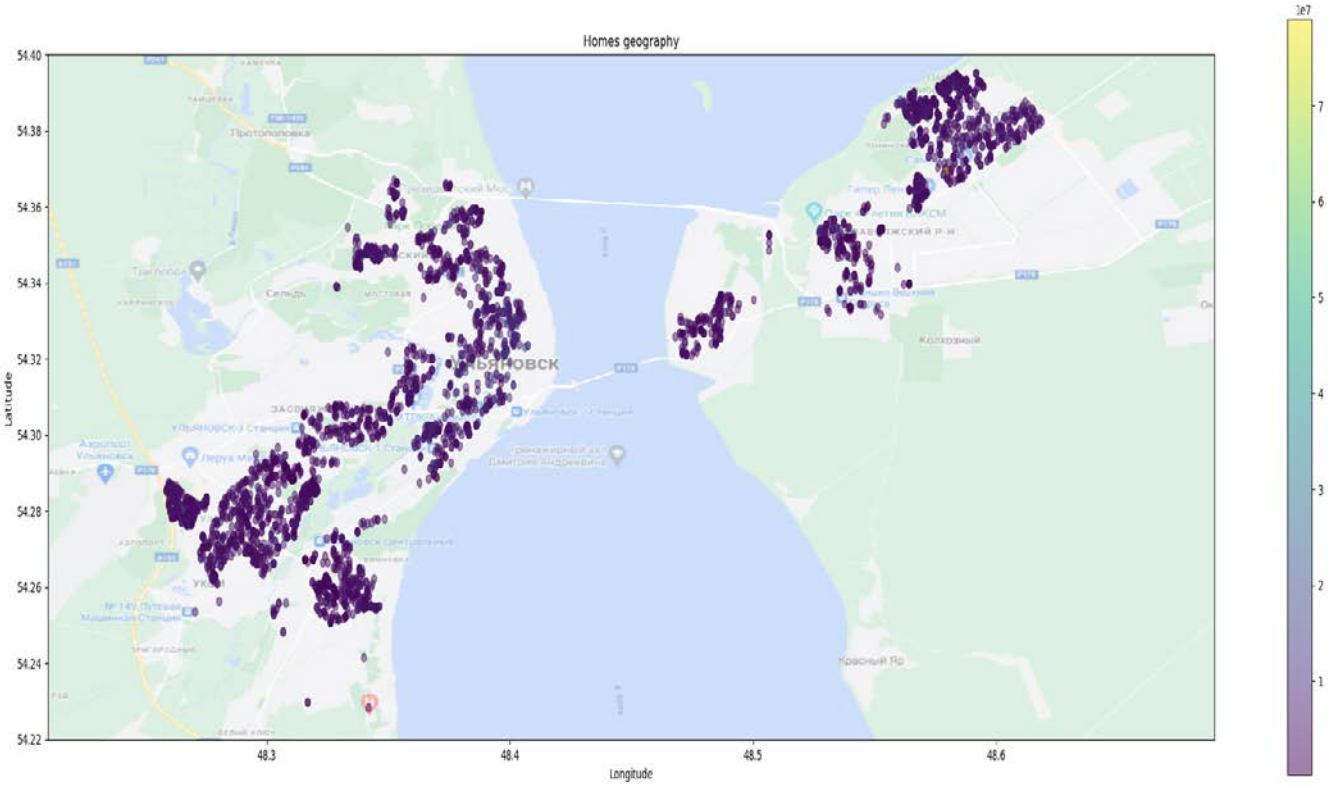
Посмотрим на географическое распределение объявлений по районам города. Город Ульяновск с районами представлен на изображении ниже.



С помощью графического редактора и скриншота с Яндекс.Карт было подготовлено изображение карты города в нужных границах по географическим координатам.

Нанесем на него данные по количеству объявлений:

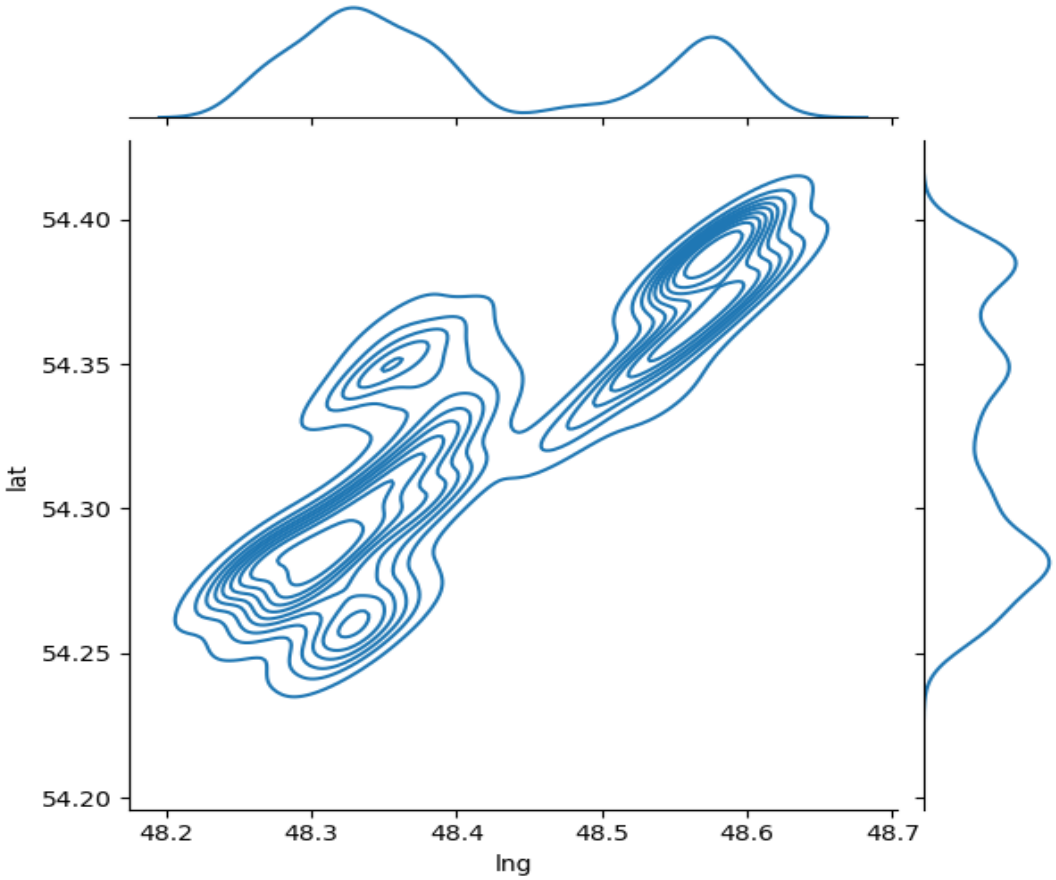
```
min_long = 48.21
max_long = 48.69
min_lat = 54.22
max_lat = 54.40

import matplotlib.image as img

ulv_map = img.imread('ulv_map.png')

plt.figure(figsize=(30, 10))
plt.imshow(ulv_map, extent=[min_long, max_long, min_lat, max_lat], alpha=0.5)

sc = plt.scatter(df_realty['lng'], df_realty['lat'], alpha=0.5, c=df_realty['Цена'])

plt.colorbar(sc)

plt.xlabel("Longitude")

plt.ylabel("Latitude")

plt.title("Homes geography");
```



Те же данные построим с помощью сводной диаграмм модуле seaborn:

```
sns.jointplot(x=df_realty['lng'], y=df_realty['lat'], kind='kde')
```

```
<seaborn.axisgrid.JointGrid at 0x248836352b0>
```



Обе визуализации показывают увеличение количества объявлений на юго-западной окраине города в Засвияжском районе и северо-восточной окраине города в Заволжском районе.

Обе эти зоны соответствуют зонам новостроек города Ульяновск - микрорайонам «Юго-Западный» в Засвияжском районе и «Новый город – Центральный» в Заволжском. И там и там преобладают дома с одно и двухкомнатными квартирами со стоимостью от 3 до 4 млн. рублей.

Наши визуализации подтвердили эти данные.

Теперь проверим зависимости цен на квартиры от других характеристик квартиры.

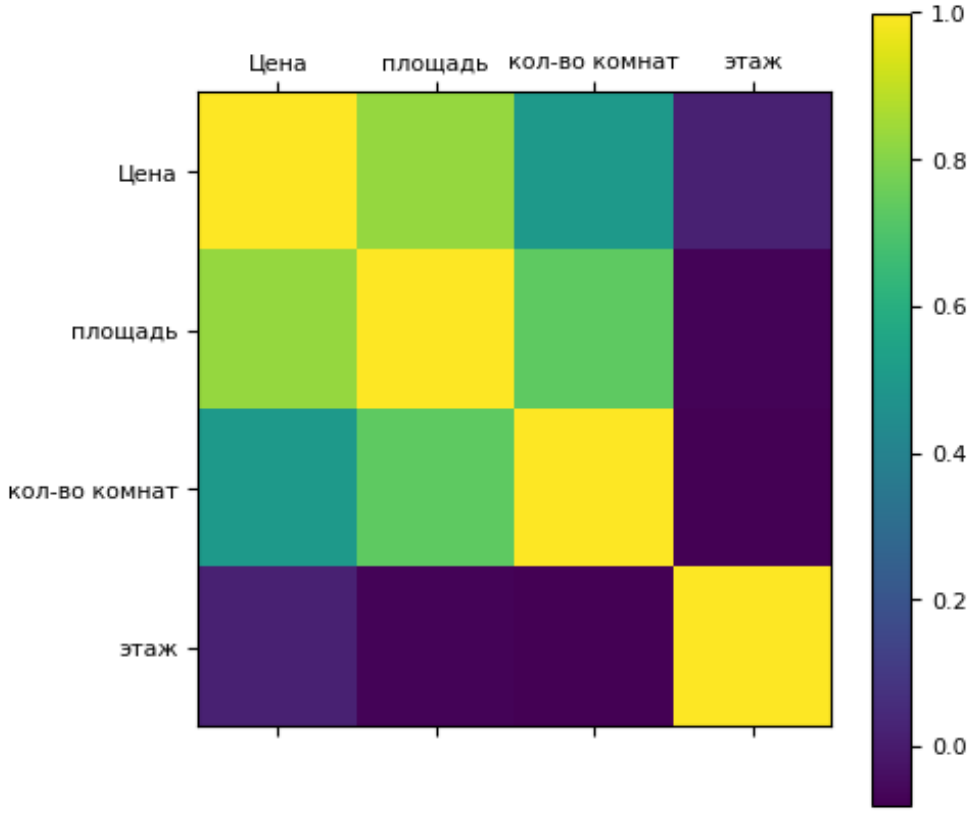
Наш набор данных позволяет проверить только взаимосвязь таких характеристик как 'Цена', 'площадь', 'кол-во комнат', 'этаж'.

С помощью библиотеки seaborn построим тепловую карту взаимозависимостей по этим характеристикам:

```
df_tmp = df_realty[['Цена','площадь','кол-во комнат','этаж']]
df_tmp.head()
```

	Цена	площадь	кол-во комнат	этаж
0	3000000	48.5	2	4
1	4850000	51.0	2	3
2	6800000	45.8	2	7
3	3999600	52.6	2	12
4	2240000	41.1	2	1

```
f = plt.figure(figsize=(5, 5))
plt.matshow(df_tmp.corr(), fignum=f.number)
plt.xticks(range(df_tmp.shape[1]), df_tmp.columns, fontsize=8)
plt.yticks(range(df_tmp.shape[1]), df_tmp.columns, fontsize=8)
cb = plt.colorbar()
cb.ax.tick_params(labelsize=8)
```

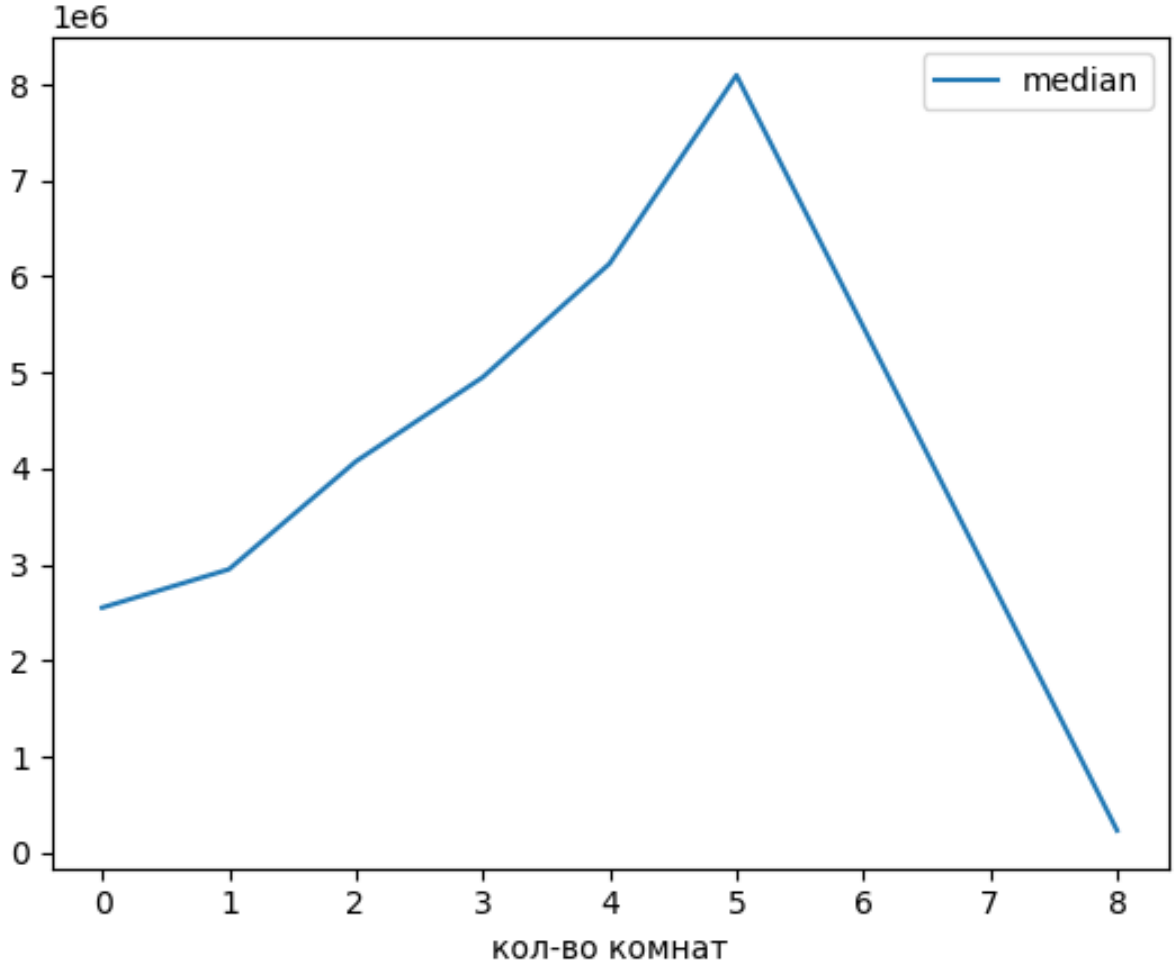


Тепловая карта показывает, что какая-то значимая зависимость есть между значениями ‘Цена’, ‘площадь’ и ‘кол-во комнат’, что в принципе логично - чем больше площадь или больше комнат комнат, тем большое цена.

Построим графики зависимости:

```
df_realty_rmexposition = df_realty.pivot_table(index = 'кол-во комнат', values = 'Цена',
                                               aggfunc = ['mean', 'count', 'median'])
df_realty_rmexposition.columns = ['mean', 'count', 'median']
df_realty_rmexposition.plot(y = 'median')
df_realty_rmexposition.sort_values('median', ascending = False)
```

кол-во комнат	mean	count	median
5	9.372000e+06	16	8095500
4	7.233262e+06	196	6133595
3	5.407990e+06	1103	4950000
2	4.203097e+06	1626	4073340
1	3.038315e+06	1564	2950000
0	3.233069e+06	269	2550400
8	2.298400e+05	1	229840

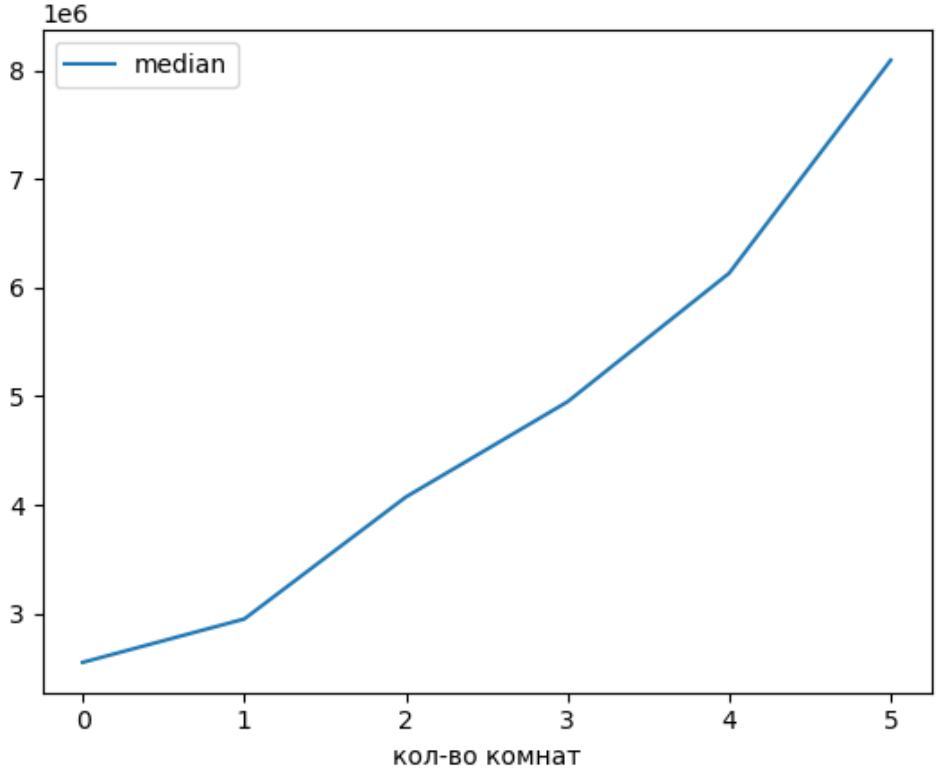


По выводу операции и графику видно, что существует выброс в данных: одна восьмикомнатная квартира с низкой ценой.

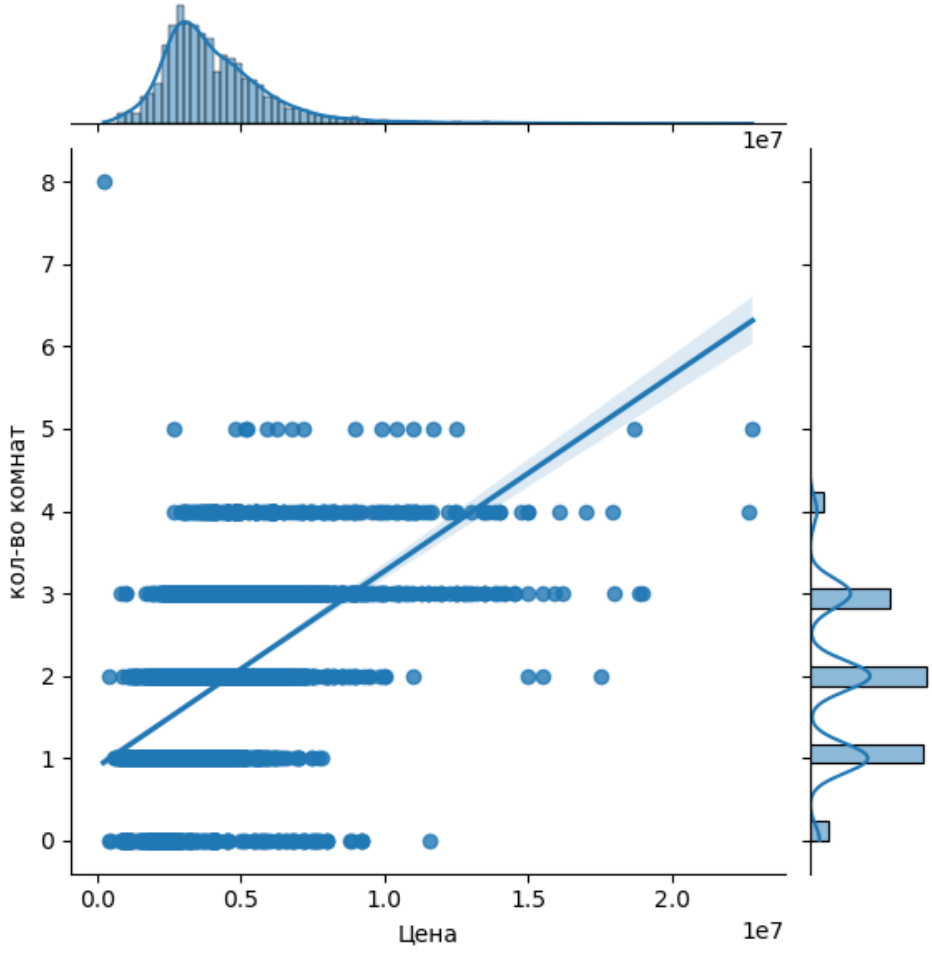
Уберем запись о этом объявлении и перестроим график:

```
df1 = df_realty[df_realty['кол-во комнат'] != 8 ]
df_realty_rmexposition = df1.pivot_table(index = 'кол-во комнат', values = 'Цена',
                                         aggfunc = ['mean', 'count', 'median'])
df_realty_rmexposition.columns = ['mean', 'count', 'median']
df_realty_rmexposition.plot(y = 'median')
df_realty_rmexposition.sort_values('median', ascending = False)
```

кол-во комнат	mean	count	median
5	9.372000e+06	16	8095500
4	7.233262e+06	196	6133595
3	5.407990e+06	1103	4950000
2	4.203097e+06	1626	4073340
1	3.038315e+06	1564	2950000
0	3.233069e+06	269	2550400

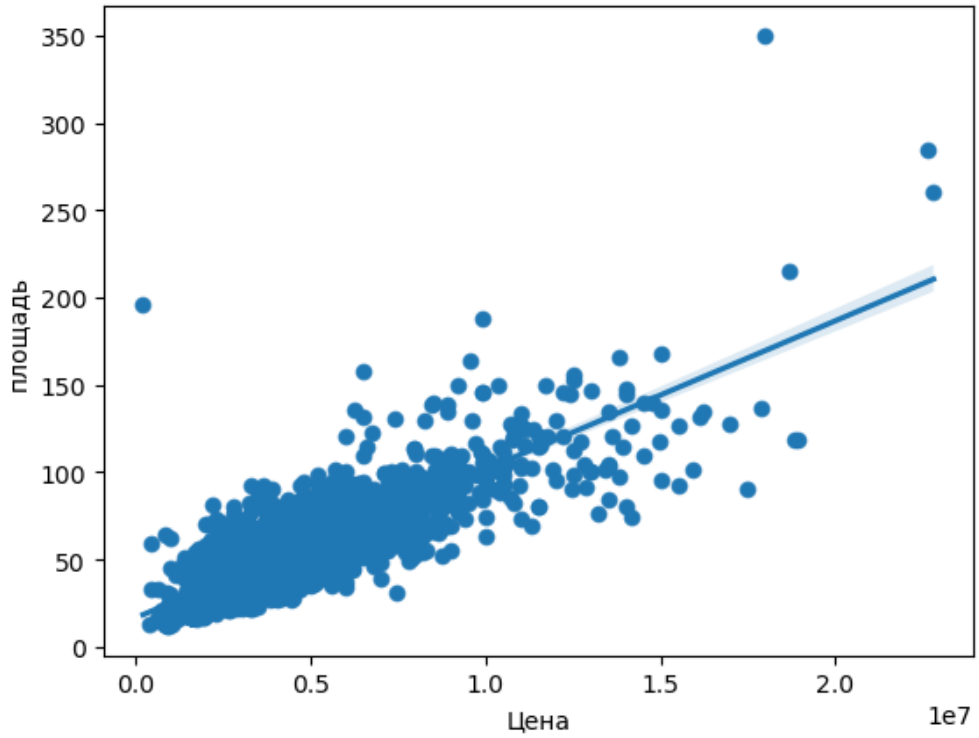


Теперь график зависимости стал более линейным. Также можно построить график зависимости через функция jointplot:



Посторим диаграммц распределения цен в зави цен от площади:

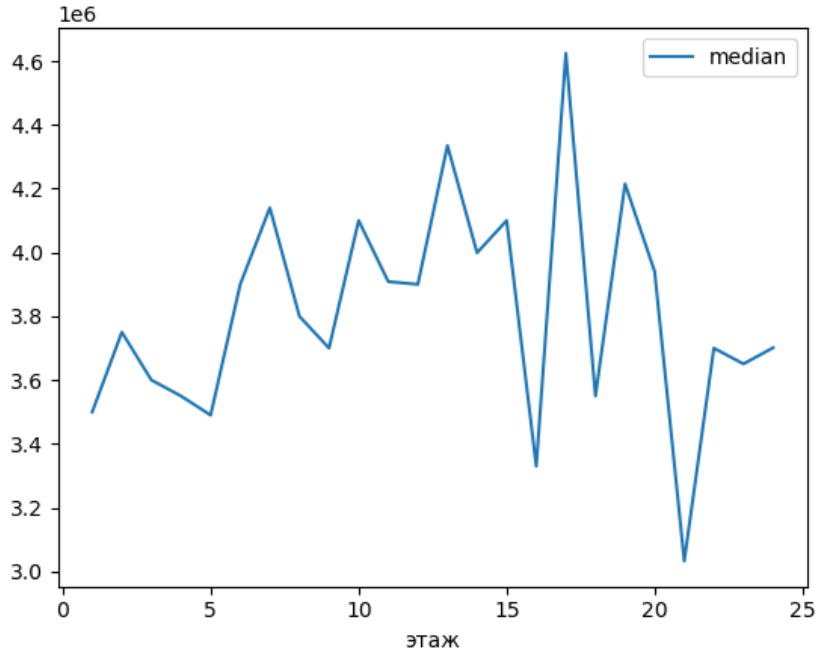
```
sns.scatterplot(data=df_realty, x='Цена', y='площадь')
sns.regplot(data=df_realty, x='Цена', y='площадь')
```



Несмотря на большой разброс в значениях – наличии зависимости имеется.

Посторим график зависимоти цен от этажности:

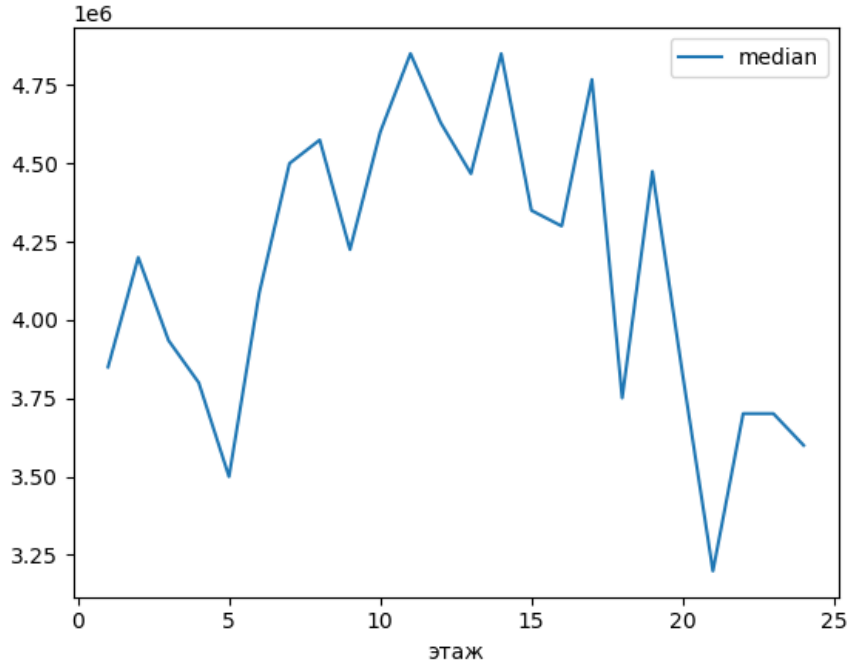
```
df_realty_floor_category = df_realty.pivot_table(index = 'этаж',
                                                 values = 'Цена', aggfunc = ['mean', 'count', 'median'])
df_realty_floor_category.columns = ['mean', 'count', 'median']
df_realty_floor_category.plot(y = 'median')
df_realty_floor_category
```



Данный график не совсем точен – так как сравниваются квартиры с разным количеством комнат.

Выберем только один тип – например только двухкомнатные квартиры:

```
df_realty_floor_category =  df_realty[df_realty['кол-во комнат'] == 2 ].pivot_table(index = 'эта
                                        values = 'Цена', aggfunc = ['mean', 'count', 'median'])
df_realty_floor_category.columns = ['mean', 'count', 'median']
df_realty_floor_category.plot(y = 'median')

df_realty_floor_category.query('этаж > 20')
```



По графику видны падения цен на отметках в 1, 5, 9, 12, 18, 21. Так как пяти- девяти- двенадцати- восемнадцатиэтажные дома являются стандартными городскими домами график показывает, что на первых и последних этажах цены на на квартиры дешевле. Особенно заметны падения на отметках 5 и 21. С пятиэтажными домами понятно – это типичная высота «хрущевок» без лифта и квартиры на 5-м этаже там всегда дешевые. Остались квартиры выше 20 этажа:

```
df_realty_floor_category.query('этаж > 20')
```

этаж	mean	count	median
21	3.687927e+06	7	3198550.0
22	4.172110e+06	9	3701240.0
23	4.109529e+06	11	3701240.0
24	3.771455e+06	11	3600000.0

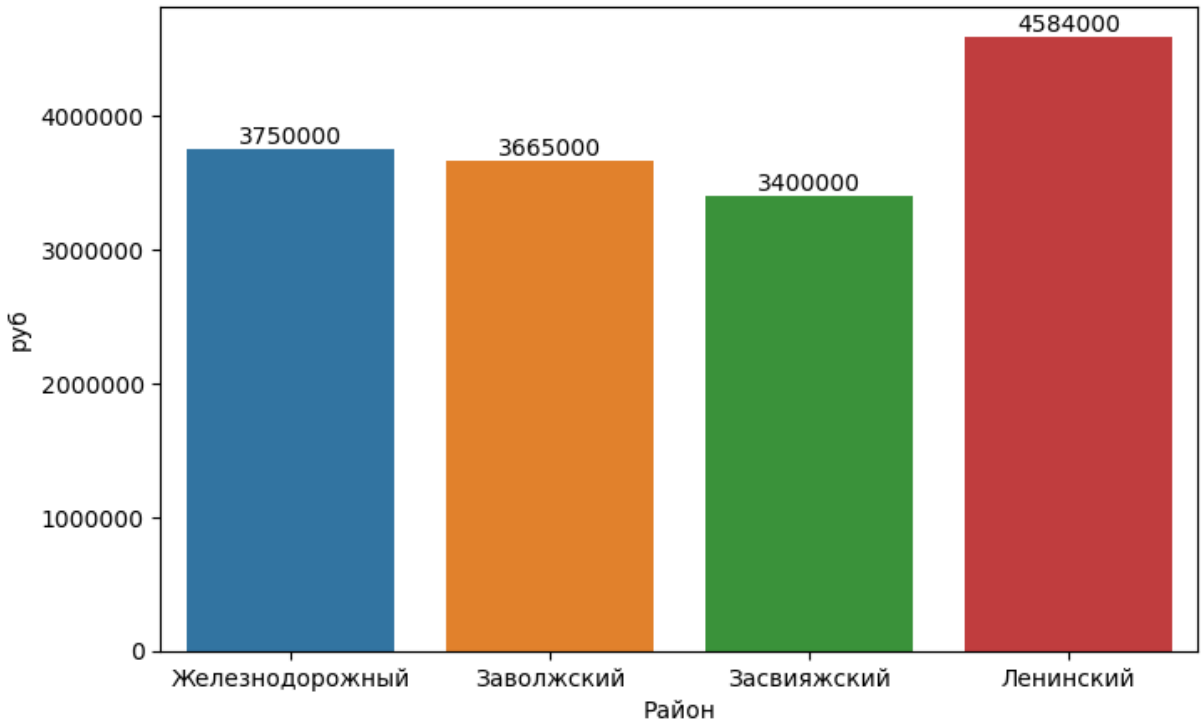
Двадцать первый этаж скорее всего тоже последний, а общее снижение цен на этажах выше двадцати требует дополнительных данных о типах этих домов, но

в нашем датасете таких данных недостаточно. Возможно, квартиры на высоких этажах просто не популярны.

Теперь посмотрим, как цены различаются по районам города. Будем использовать не среднюю величину а медиана, так как она более точно показывает уровень цен по районы (сглаживает выбросы).

Построим столбчатую диаграмму:

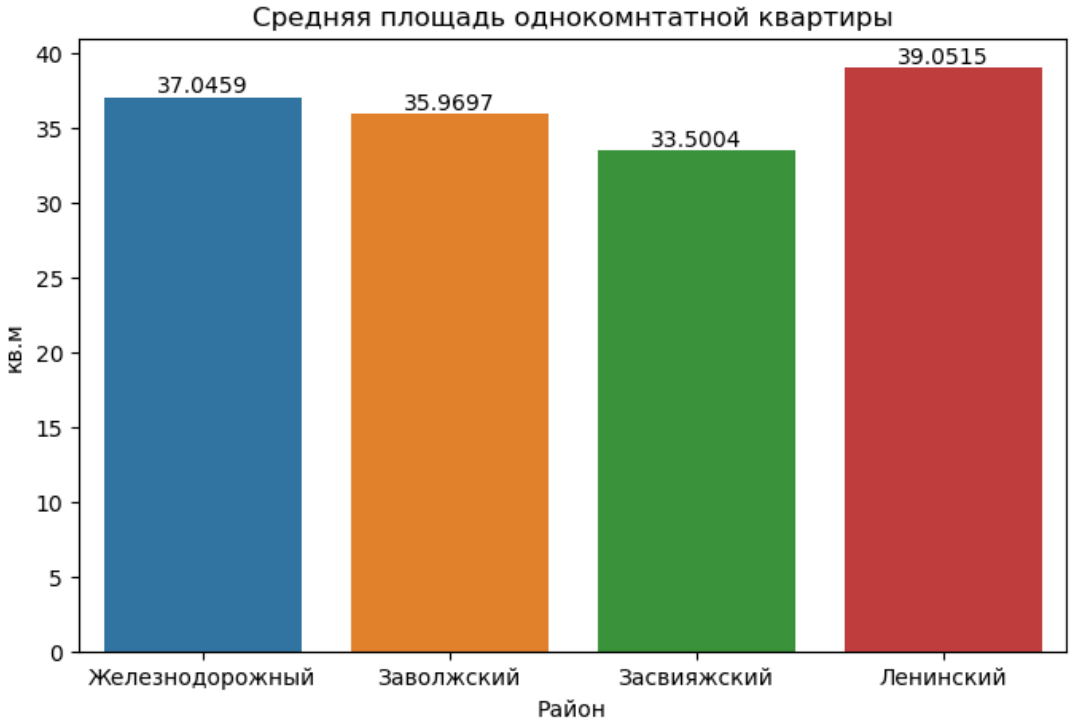
```
import matplotlib.ticker as tkr
df_realty_blk = df_realty.pivot_table(index =
                                      'Район', values = 'Цена', aggfunc = ['mean', 'count', 'median'
df_realty_blk.columns = ['mean', 'count', 'median']

plt.figure(figsize=(8, 5))
ax=sns.barplot(x = df_realty_blk.index, y = df_realty_blk['median'])
ax.bar_label(ax.containers[0], fontsize=10, fmt='%.0f');
ax.yaxis.get_major_formatter().set_scientific(False)
ax.yaxis.get_major_formatter().set_useOffset(False)
plt.title('Цена квартир (медианная)')
plt.xlabel('Район')
plt.ylabel('руб');
```



По ценам на квартиры некоторое снижение от общего уровня цен есть в Засвияжском районе. Возможно, это связано с тем, что в районе много квартир, построенных в 50-х годах. Но более всего по ценам выделяется Ленинский район

города. Это легко объяснимо, так как в Ульяновске центр города находится в этом районе и большинство «элитных» квартир так же находятся там.

Дополнительно выведем диаграмму средней площади однокомнатных комнатных квартир по районам города:



Как видно по диаграмме средняя площадь однокомнатный квартиры в центре немного выше чем в других района города.

Выведем еще один показатель, часто используемый риэлторами – среднюю цену квадратного метро по городу и по районам города.

```
df_realty['Цена'].sum()/df_realty['площадь'].sum()
```

```
80295.79139181695
```

```
block=df_realty['Район'].unique()
for i in range(4):
    print(f"\
    {block[i]} -\
    {df_realty[df_realty['Район'] \
     == block[i]]\['Цена'].sum() / df_realty[df_realty['Район'] \
     == block[i]]['площадь'].sum():.2f}")
```

```
Железнодорожный -     77162.92
Засвияжский -     78969.25
Заволжский -     76388.95
Ленинский -     90673.16
```

Для сравнения, данные по стоимости квадратного метра можно сравнить с данными другого сайта по продаже недвижимости domclick.ru за этот же период (август 2023 г.):

https:// opendata.domclick.ru/offers/table/ulyanovskaya-oblast/month/2023-08-01

Введите регион | Август 2023

Ульяновская Область

21 036 -302	4 873 -304	2 592 -644	73 823 ₽ +436 ₽	85 000 ₽ -1 067 ₽
Размещённых объявлений	Активных объявлений, вторичка	Активных объявлений, новостройки	Средняя стоимость м², вторичка	Средняя стоимость м², новос

№	Город	Активных объявлений о продаже	Активных объявлений, вторичка	Активных объявлений, новостройки	Средняя стоимость м², вторичка	Средняя стоимость м², новостройки
1	Ульяновск	7 635	3 856	2 592	78 025 ₽	85 000 ₽

Данные на этом сайте не сильно отличаются от полученных нами в результате анализа объявлений на Avito.

В качестве заключения выведем портрет средней, продаваемой в городе Ульяновск квартиры:

```
import warnings
warnings.filterwarnings('ignore')

meanDf = round(df_realty.mean(),2)
meanDf
```

```
Цена              4171240.08
lat                    54.32
lng                    48.41
площадь                51.95
этаж                    5.87
кол-во комнат           1.88
dtype: float64
```

Это двухкомнатная квартира площадью немногим больше 54 кв.метра на 5 или 6-м этаже и ценой в 4 млн.171 тыс. рублей.

2.5 Интерпретация результатов.

По результату импорта полученных данных для анализа, было выявлено, что в них дубликатов, и каких-либо серьезных ошибок нет. Единичные ошибки были исправлены, наиболее трудоемких оказался процесс определения местонахождения квартир по геометкам в некоторых объявлениях с отсутствующими адресами.

В результате проведенного анализа было выявлено, что в г. Ульяновск преобладают объявления о продаже одно- и двухкомнатных квартир в ценовом диапазон в ценовом диапазоне от 2,5 до 4 млн. рублей и площадью 40-50 кв. метров.

Средняя стоимость квадратного метра в городе составляет 80 295 рублей.

Цены в объявлениях зависят от:

- Количества комнат;
- Площади;
- Этажности (квартиры на первом этаж дешевле);
- Района города (квартиры в Ленинском районе – центре города дороже).

В целом объявления по продаже квартир на сайте Avito хорошо представляют рынок недвижимости города.

Заключение

Анализ данных — это процесс изучения и интерпретации данных с целью выявления закономерностей, трендов и важных характеристик.

Язык программирования Python с его богатой библиотекой дополнений позволяет быстро провести работу по аналитике данных и получить результаты данного анализа в виде понятных визуализаций. Все возможности языка Python для анализа данных реализуются при использовании такой интерактивной среды разработки как Jupyter Notebook. Он предоставляет удобную среду для написания кода, его выполнения и визуализации результатов. Notebook позволяет работать с данными в интерактивном режиме, что делает процесс анализа более наглядным и понятным. Это очень удобно для анализа данных, так как позволяет быстро проверять различные гипотезы и видеть, как они работают на реальных данных. Jupyter Notebook уже несколько лет считается одним из популярных инструментов для анализа данных. Jupyter Notebook является инструментом аналитика, которым пользуются практически каждый день – от загрузки данных до создания и развертывания моделей с его помощью.

Знание возможностей языка Python для анализа данных и умение работать в Jupyter Notebook позволило быстро загрузить, подготовить и провести исследование объявлений о продаже квартир в г.Ульяновск на основе объявлений с сайта Avito.

Список используемой литературы

1. М. А. Поручиков Анализ данных: учеб. пособие / М.А. Поручиков. – Самара: Изд-во Самарского университета, 2016. – 88 с. ISBN 978-5-7883-1085-5
2. Федоров, Д. Ю. Программирование на языке высокого уровня Python: учеб. пособие для прикладного бакалавриата / Д. Ю. Федоров. – 2-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2019. – 161 с. ISBN 978-5-534-10971-9
3. Как получить Jupyter Python Notebook на AWS https://questu.ru/articles/723449/
4. Программирование и научные вычисления на языке Python https://ru.wikiversity.org/wiki/Программирование_и_научные_вычисления_на_языке_Python
5. Официальная документация Project Jupyter: https://jupyter.readthedocs.io/en/latest/projects/architecture/content-architecture.html
6. Открытые данные «Домклик» https://opendata.domclick.ru/offers/table/ulyanovskaya-oblast/

Приложения

Файлы с исходными данными, и расчеты в формате Jupyter Notebook выложены на GitHub: https://github.com/anburd/DIPL