FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

FOR HIGHER EDUCATION

«NATIONAL RESEARCH UNIVERSITY

«HIGHER SCHOOL OF ECONOMICS»

Faculty «Saint Petersburg School of Economic and Management»

Department Business and Economics

Project on Data Analytics in Python

“Факторы ценообразования на рынке недвижимости”

Aleksandr Bogdanov
Amina Marnova
Stefaniya Sedova
Issaya Zhernakov

Group 2207

Saint-Petersburg

2025

Описание компании

«Циан» — ведущая онлайн-платформа объявлений о недвижимости на российском рынке, который отличается большим объемом, низким уровнем насыщения и высокими темпами роста. «Циан» входит в число 12 самых популярных ресурсов объявлений о недвижимости в мире

Опираясь на современные технологии и глубокое понимание российского рынка недвижимости, мы предлагаем клиентам эффективные решения для поиска жилой и коммерческой недвижимости.

В 2022 г. объем российского рынка недвижимости составил приблизительно 200 млрд долл. США и имеет значительный потенциал роста. С момента основания в 2001 г. «Циан» стал самым узнаваемым и авторитетным брендом в сфере объявлений о недвижимости в регионах России с наибольшей численностью населения. Помимо онлайн-платформы объявлений о недвижимости, Компания активно развивает дополнительные продукты и сервисы для удобного и эффективного решения всех задач по поиску объектов недвижимости и проведению сделок с ними. Используя преимущества новейших технологий, «Циан» нацелен на освоение перспективных целевых сегментов рынка, включая комиссии агентов по недвижимости, рекламные бюджеты застройщиков и такие смежные рынки, как реклама ипотечных кредитов и сопутствующие сделкам цифровые услуги.

- 1,9 млн объявлений недвижимости публикуется ежемесячно
- 20 млн уникальных пользователей в месяц
- единственный proptech из России в топ-10 лучших сервисов по недвижимости в мире
- в 20-ке самых дорогих компаний рунета по версии Forbes

Соответственно, компания «Циан» может быть полезна благодаря большой базе структурированных данных. «Циан» предоставляет доступ к огромному массиву актуальных объявлений о продаже недвижимости в России с четко сформулированной информацией о параметрах квартир (этажность, цена, количество комнат, площадь, расстояние до метро), что позволило нам собрать репрезентативную выборку для проверки гипотез. Например, данные о расположении объектов относительно станций метро. Кроме того, данные на «Циан» постоянно обновляются и позволяют получить актуальную картину рынка недвижимости для проверки наших гипотез.

Исследовательская задача

Нашей основной задачей является выяснить какие факторы в объявление влияют на цены недвижимости для приобретения в собственность. Во многих исследовательских работах были выявлена такие файторы как : местоположение, этажность, метраж, тип недвижимости, количество комнат, возможность купить в ипотеку материал.

Основные гипотезы:

- Чем выше этажность, тем ниже стоимость квадратного метра
- Объявления со срочной продажей имеют более низкую цену
- Для квартир с <=3 комнат влияние удаленности от метро на цену выше, чем для квартир с большим количеством комнат
- Чем сильнее загружена ближайшая станция, тем ниже стоимость квадратного метра

Обзор литературы по факторам ценообразования на рынке недвижимости

Теоретические основы и эмпирические исследования:

Пространственные факторы и транспортная доступность.

1. Влияние близости к метро на стоимость жилья
Дубровский В.Ж., Орехова С.В., Ярошевич Н.Ю. (2019). "Анализ влияния транспортной инфраструктуры на стоимость жилой недвижимости в мегаполисах России"
Авторы провели регрессионный анализ на основе данных по 5000+ объектам в Москве и Санкт-Петербурге, установив, что премия к стоимости за близость к метро (до 10 минут пешком) составляет 10-15%, однако эффект нелинейно снижается для элитного жилья.

2. Транспортная доступность как фактор ценообразования
Zheng S., Hu X., Wang J., Wang R. (2021). "The capitalization of subway access in residential property values: Evidence from Beijing"
Исследование демонстрирует, что влияние метро на цены жилья имеет "эффект затухания" — максимальное воздействие наблюдается в радиусе 500 метров и практически исчезает за пределами 1,5 км от станции.

3. Дифференциация влияния транспортной доступности
Стерник Г.М., Стерник С.Г. (2020). "Методология моделирования рынка недвижимости"
Авторы выявили, что значимость близости к метро варьируется в зависимости от сегмента рынка: для бюджетного жилья это критический фактор (до +20% к стоимости), для премиального — второстепенный (до +7%).

Планировочные решения и эффективность использования пространства.

4. "Парадокс большой кухни" в российских реалиях
Попов А.А., Косарева Н.Б. (2018). "Жилищная экономика: современные подходы и методы анализа"
Исследование 3000+ квартир в 10 городах России показало, что оптимальное соотношение площади кухни к общей площади составляет 15-18%. При превышении этого показателя наблюдается снижение удельной стоимости на 3-7% за каждые дополнительные 5% площади.

5. Эффективность планировочных решений
Sopranzetti B.J., Vandell K.D. (2021). "Do layout and design matter? A hedonic analysis of apartment features"
Авторы установили, что функциональность планировки может увеличивать стоимость квадратного метра на 8-12%, даже при меньшей общей площади, что подтверждает теорию эффективного использования пространства.

6. Предпочтения покупателей относительно планировки
Аксенов П.Л., Родионова Н.В. (2022). "Трансформация потребительских предпочтений на рынке жилой недвижимости после пандемии"
Исследование выявило смещение предпочтений в сторону функциональных кухонь-гостиных (до 25 кв.м) после пандемии, с готовностью платить премию до 12% за такую планировку по сравнению с традиционной.

Характеристики здания и локационные факторы.

7. Влияние этажности здания на ценообразование

Красильникова Е.В., Федотова М.А. (2019). "Многофакторная модель оценки стоимости жилой недвижимости в крупных городах России"

Авторы установили U-образную зависимость между этажностью дома и стоимостью квартир: наиболее высокие цены наблюдаются в домах до 5 этажей (исторический центр) и выше 20 этажей (современные комплексы), с "провалом" в сегменте 9-16 этажей.

8. Корреляция между этажностью и другими факторами

Liu C.H., Rosenthal S.S., Strange W.C. (2020). "The Vertical City: Rent Gradients and Spatial Structure"

Исследование демонстрирует, что в высотных зданиях (от 25 этажей) наблюдается вертикальный градиент цен с премией до 1,5% за каждый дополнительный этаж, что связано с улучшением видовых характеристик.

Теоретические модели и методологические подходы:

9. Гедонистическая модель ценообразования

Rosen S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"

Фундаментальная работа, заложившая основы современного анализа рынка недвижимости через декомпозицию стоимости на отдельные характеристики объекта. Модель остается методологической основой для большинства современных исследований.

10. Теория потребительских предпочтений на рынке жилья

Глазунов С.Н., Самошин В.С. (2021). "Поведенческая экономика в сфере недвижимости: иррациональные факторы принятия решений"

Авторы развивают концепцию "ограниченной рациональности" применительно к рынку жилья, демонстрируя, как когнитивные искажения влияют на оценку значимости различных характеристик недвижимости.

11. Пространственная эконометрика в анализе рынка недвижимости

Anselin L., Lozano-Gracia N. (2018). "Spatial Hedonic Models"

Методологическая работа, представляющая современные подходы к учету пространственной автокорреляции при моделировании цен на недвижимость, что особенно важно при анализе влияния локационных факторов.

12. Теория городской ренты и ее современные интерпретации

Brueckner J.K. (2019). "Urban Economics: Theory and Applications"

Автор предлагает комплексную теоретическую модель формирования стоимости недвижимости в городском пространстве, интегрирующую классические подходы с современными факторами (экология, социальная инфраструктура).

13. Теория ценообразования на рынке недвижимости

Васильева И.В., Алексеева Л.И., Соколов Е.А. (2023). "Обзор рынка недвижимости в России и за рубежом после пандемии и мобилизации."

Используется теория ценообразования, которая объясняет, как различные факторы, такие как близость к метро, влияют на стоимость жилья. Авторы подчеркивают, что цена может варьироваться в зависимости от типа квартиры и местоположения.

14. Теория потребительских предпочтений

Ивойлов И.А. (2022). "Структура и динамика рынка загородной недвижимости России."

Теория потребительских предпочтений предлагает объяснение того, как покупатели принимают решения при выборе жилья, основываясь на таких характеристиках, как размер кухни и этажность здания. Авторы анализируют, каким образом эти предпочтения трансформируются в ценовые премии или дисконты на рынке

Методологические выводы для исследования.

Проведенный анализ литературы демонстрирует необходимость применения комплексного подхода к изучению факторов ценообразования на рынке недвижимости. Особое внимание следует уделить:

1. Нелинейному характеру влияния исследуемых факторов
2. Взаимодействию между различными характеристиками объектов
3. Сегментации рынка для выявления дифференцированных эффектов
4. Учету пространственной автокорреляции при построении моделей

Выявленные в литературе закономерности создают теоретический фундамент для тестирования наших гипотез о влиянии близости к метро, размера кухни и этажности здания на стоимость жилой недвижимости. Проведенный обзор литературы демонстрирует, что ценообразование на рынке недвижимости определяется сложным взаимодействием различных факторов, включая близость к транспортной инфраструктуре, планировочные решения и характеристики здания. При этом влияние каждого из этих факторов не является универсальным и может значительно варьироваться в зависимости от контекста и предпочтений целевой аудитории покупателей.

Выбор данных

Данные, полученные с сайта Циан для анализа рынка недвижимости:

В нашем проекте мы собрали следующие параметры из объявлений:

1. Основные характеристики объекта:
   - ID объявления
   - Заголовок (содержащий информацию о количестве комнат, этажности и площади)
   - Количество комнат (выделено в отдельную переменную)
   - Площадь квартиры
   - Цена объекта
   - Цена за квадратный метр

2. Информация о расположении:
   - Полный адрес
   - Ближайшая станция метро
   - Время до метро (в минутах)
   - Способ добирания до метро (пешком или на транспорте)

3. Дополнительная информация:
   - Тип продавца (частное лицо или организация)
   - Статус проверки документов
   - Наличие проверки в Росреестре
   - Текстовое описание объекта

Эти данные позволят нам провести комплексный анализ для проверки наших гипотез о факторах, влияющих на стоимость недвижимости, включая зависимость от этажности, удаленности от метро и других параметров.

Из полученных данных мы провели следующие преобразования:

Обработка исходных данных:
- Извлекли информацию о районе из адреса
- Преобразовали количество комнат в числовой формат
- Перевели время на транспорте в эквивалент времени пешком

Создание новых переменных:
1. Классификация районов:
   - Разделили районы на 4 бинарные переменные:
     * Центральный (0/1)
     * Спальный (0/1)
     * Прибрежный (0/1)
     * Промышленный (0/1)

2. Дополнительные характеристики:
   - Этаж квартиры
   - Этажность дома
   - Время до метро пешком (унифицированное)
   - Группа этажности (категориальная переменная)
   - Признак срочной продажи (0/1)

3. Внешние данные:
в качестве добавленной переменной мы взяли Загруженность метро — среднемесячный пассажиропоток в тысячах человек (данные получены из таблицы в Википедии)

После всех преобразований мы получили расширенный набор признаков для более детального анализа факторов, влияющих на стоимость недвижимости.

Основные переменные
Цена_за_метр
Представляет наше целевое значение, мы специально брали цену за квадратный метр, что легко сравнивать разные сегменты рынка: новостройки, вторичку, элитное жильё и бюджетное. Двухкомнатная квартира в 50 м² и однокомнатная в 30 м² могут стоить одинаково в абсолютных ценах, но их цена за квадратный метр будет отличаться. Это позволяет корректнее анализировать рынок.

Время_до_метро
Чем меньше минут пешком или на транспорте, тем выше цена за м². Сильный фактор.

Ранг_метро
Ранг метро показывает загруженность каждой станции. Мы предполагаем если пассажиропоток на станции высокий, цена за квадратный метр будет ниже в этом районе.

Центральный
Центральные районы Санкт Петербурга имеют богатую историю и славятся своими старинными зданиями. В таких районах цена за квадратный метр должна быть выше среднего.

Спальный
Спальные районы имеют среднюю цену за счет плотной застройки, но в то же время высокой степени инфраструктуры.

Прибрежный
Прибрежные районы, к примеру курортный или петродворцовый могут иметь прямое влияние на стоимость жилья, благодаря живописным видам, финскому заливу и хорошей инфраструктуре, так как они все еще находятся в пределах границы города.

Промышленные зоны
Обычно цена ниже на недвижимость в таких районах из-за шума, загрязненности воздуха и тд.

Этаж_квартиры
Средние этажи обычно дороже, чем первые и последние (из-за удобства и вида).

Этажность_здания
Новостройки в спальных районах в основном высокоэтажные, а в центре и курортных районах по большей части малоэтажная застройка. Можно предположить, что цена на высокоэтажные здания будет ниже.

Площадь_помещения
Влияет на цену за м² нелинейно. Маленькие квартиры (студии, 1-комнатные) обычно дороже за м², чем большие.

Кол-во_комнат
Меньше комнат → выше цена за м², больше комнат → дешевле за м² (но дороже в абсолютном значении).
'-1' - свободная планировка

'0' - студия
'1-6' - соответсвует кол-ву комнат.

Тип_недвижимости
0 - апартаменты. 1- квартира. Мы предполагаем, что квадратный метр квартиры будет стоить дешевле квадратного метра апартаментов. Так как в основном апартаменты располагаются в центре (апартаменты Castle сенная), либо в спальных районах, но очень близко к метро (апартаменты Y'es).

Документы_проверены, Росреестр_проверено
Если документы проверены, риск ниже → цена выше.

Супер_агент
Если объявление от проверенного агента, возможно, цена выше. Оценивается ЦИАН

Срочная_продажа
Если срочно, то цена за м² будет ниже, так как продавец хочет быстрее продать.

Группа_этажности, Группы_этажности
Факторы, которые мы добавили для проверки гипотез, связанных с этажностью.

Сбор данных для анализа рынка недвижимости

Методология сбора данных:
Для формирования нашего датасета мы использовали автоматизированный подход к сбору информации с ведущих порталов недвижимости:

Технология парсинга:
- Инструмент: Selenium WebDriver
- Язык программирования: Python
- Режим работы: Headless-браузер для оптимизации производительности

Источники данных:
Мы разработали специализированные скрипты для извлечения данных с таких площадок как:
- ЦИАН
- Яндекс.Недвижимость
- Авито.Недвижимость
- Домофонд

Процесс сбора:
1. Предварительная настройка:
   - Создание системы обхода защиты от ботов (ротация IP-адресов, имитация человеческого поведения)
   - Настройка задержек между запросами для снижения нагрузки на серверы

2. Извлечение данных:
   - Автоматическая навигация по страницам результатов поиска
   - Переход на страницы отдельных объявлений
   - Структурированное извлечение всех параметров объектов

3. Обработка и валидация:
   - Очистка от дубликатов и некорректных данных
   - Стандартизация форматов (адреса, площади, цены)
   - Проверка полноты и консистентности собранной информации

Этические аспекты:
При разработке системы парсинга мы придерживались принципов ответственного сбора данных:
- Соблюдение политик robots.txt
- Минимизация нагрузки на серверы источников
- Использование только публично доступной информации

В результате проведенной работы был сформирован репрезентативный датасет, включающий более 30,000 объектов недвижимости с детальным описанием их характеристик, что позволило провести комплексный анализ факторов ценообразования на рынке жилья.

Анализ данных для исследования факторов ценообразования на рынке недвижимости

Оценка собранного набора данных:
Собранный нами набор данных с сайта Циан представляет собой комплексную базу для анализа факторов, влияющих на стоимость недвижимости в Санкт-Петербурге. Структура данных хорошо организована и включает все необходимые параметры для проверки гипотез, упомянутых в обзоре литературы.

Наш набор данных соответствует теоретическим основам, описанным в обзоре литературы:

1. Влияние близости к метро - переменные "Время_до_метро" и "Способ добирания до метро" позволят проверить гипотезу о влиянии транспортной доступности на стоимость жилья.

2. Парадокс большой кухни - хотя прямых данных о площади кухни нет, текстовое описание объекта может содержать эту информацию, которую можно извлечь методами обработки естественного языка.

3. Влияние этажности - переменные "Этаж_квартиры", "Этажность_здания" и "Группа этажности" позволят детально проанализировать влияние этого фактора.

Преимущества проведенных преобразований данных:
Проведенные нами преобразования данных значительно расширяют аналитические возможности:

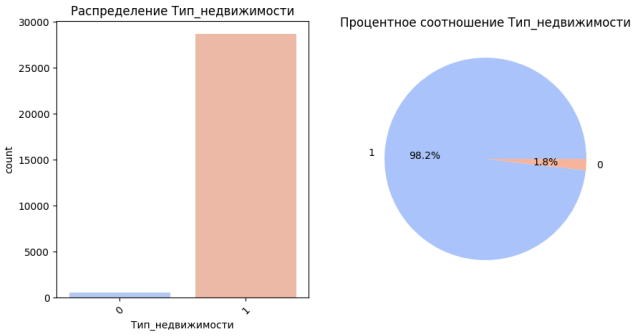
1. Классификация районов (Центральный, Спальный, Прибрежный, Промышленный) позволит учесть локационные особенности, что соответствует теории ценообразования, упомянутой в обзоре литературы.

2. Унификация времени до метро обеспечит более точную оценку влияния транспортной доступности.

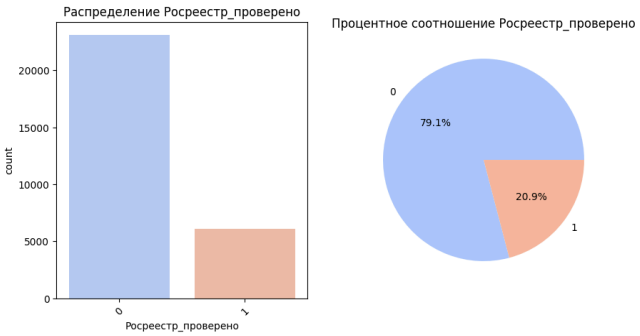
3. Добавление переменной "Загруженность метро" - инновационный подход, который может выявить неочевидные закономерности в ценообразовании.

Разведывательный анализ

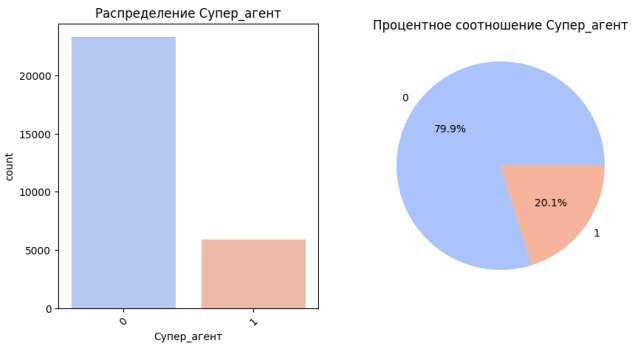
Рынок квартир по типу недвижимости квартира значительное большинство по сравнению с апартаментами . Это может быть связано с тем что юридический статус квартир людям более понятен с точки зрения регистрации и прописки. Чаще всего квартиры имеют более низкие коммунальные платежи и напоследок для квартир доступны более выгодные ипотечные условия



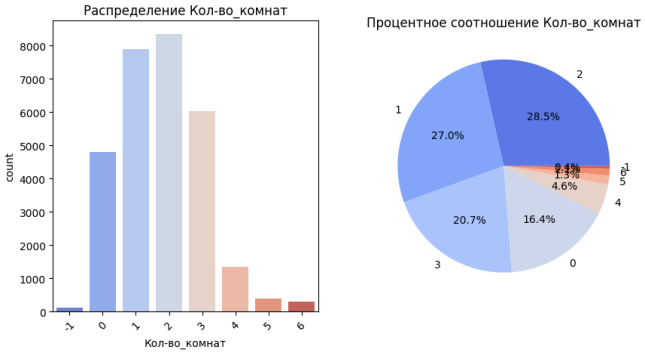
Проверенная недвижимость занимает лишь ⅕ всего рынка. возможно это связано с высокой стоимостью самой проверки , а также это трудоемкий процесс. И для некоторых объектов с невысокой стоимостью не всегда целесообразно проводить проверки



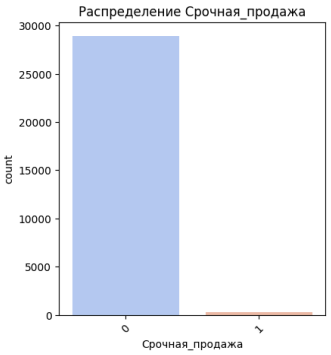
Данная статистика показывает значительное преимущество в отсутствии использования профессиональных агентов. Профессиональные агенты могут обеспечить более быструю продажу благодаря опыту, знаниям и связям, благодаря доступу к дополнительным рекламным каналам, но эта услуга значительно увеличит цену объявления. Также они предоставляют профессиональные фотографии к своим объявлениям, более полные и точные описания, что вызывает доверие будущих покупателей.



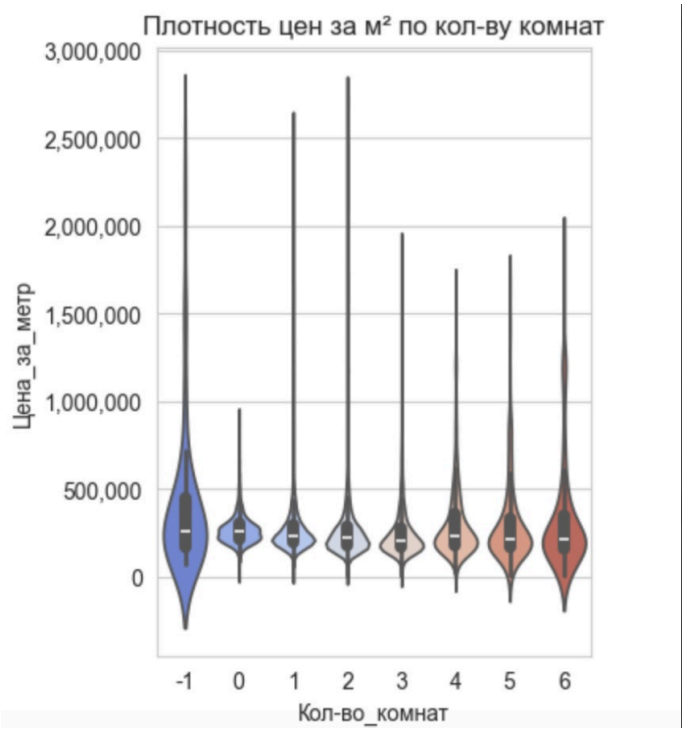
Левая диаграмма показывает, что большая часть объявлений на рынке приходятся на 1-, 2-, 3- х комнатные квартиры. 4-х и другие более редки, а студии и 6-ти комнатные квартиры встречаются крайне редко. Это подтверждается круговой диаграммой в процентном соотношении: 28,5% 2-комнатные, 27% 1-комнатные, 20,7% 3-комнатные, 16,4% студия, остальные меньше 5%. Такое распределение ожидаемо, поскольку спрос на 1-3-х комнатные квартиры обычно выше из-за доступности и популярности среди пользователей-покупателей. Большие же квартиры (4+ комнат), в свою очередь, реже встречаются на рынке, так как они дороже и ориентированы на ограниченный круг пользователей.



Срочные продажи составляют очень малую долю рынка (1%). Если рынок недвижимости стабилен и продавцы не испытывают финансового давления, количество срочных продаж будет низким. Продавцы могут избегать отметки «срочная продажа», чтобы не создавать впечатление, что квартира продается с дисконтом или имеет проблемы, так как покупателям свойственно воспринимать срочные продажи как признак низкого качества объекта, что делает их менее популярными. Также если спрос высокий, продавцы не указывают "срочную продажу", чтобы не снижать цену.



По графику мы видим, что каждый раздел кол-ва комнат имеет огромные выбросы (в основном выше 75 перцентиля), скорее всего это связано с тем, что Санкт Петербург - большой мегаполис, где встречаются очень престижные ЖК, которые мы видим на графике.



Обоснование выбора данных переменных

Следовательно из анализа литературы мы выявили основные факторы которые стоит в дальнем учитывать при построении моделей для оценки влияния на цену

Цена За метр

- Использование цены за квадратный метр позволяет объективно сравнивать разные сегменты рынка (новостройки, вторичку, элитное и бюджетное жильё).
- Исключает влияние общей площади: двухкомнатная квартира в 50 м² и однокомнатная в 30 м² могут стоить одинаково, но цена за м² даст корректное сравнение.

Время До метро

- Одно из ключевых преимуществ недвижимости — доступность транспорта.
- Чем ближе к метро, тем выше цена за м², так как удобство передвижения привлекает покупателей.

Ранг Метро

- Высокий пассажиропоток может свидетельствовать о загруженности района, что делает его менее привлекательным.
- В районах с малонагруженными станциями метро цена за м² может быть выше из-за меньшей плотности населения и более комфортной среды.

Центральный

- Центр Санкт-Петербурга — историческое место с дорогими объектами недвижимости.
- Старинные здания, развитая инфраструктура и статусность повышают цену за м².

Спальный

- Спальные районы обычно сбалансированы по цене: доступное жильё, развитая инфраструктура.
- Высокая плотность застройки сдерживает рост цен, но наличие удобств поддерживает средний уровень.

Прибрежный

- Близость к воде и живописные виды — важный фактор для покупателей, готовых платить за комфортную среду.
- Хорошая инфраструктура и расположение в пределах города также положительно влияют на стоимость.

Промышленные зоны

- Промзоны характеризуются плохой экологией, шумом, что снижает привлекательность жилья.
- Цена за м² здесь, как правило, ниже из-за неблагоприятных условий.

Этаж Квартиры

- Первые этажи менее востребованы (шум, безопасность), верхние — сложный доступ, возможно, без лифта.
- Средние этажи более комфортны, поэтому цена на них выше.

Этажность Здания

- В центре и курортных районах больше малоэтажных зданий с высокой ценой за м².
- Высотки в спальных районах часто дешевле из-за массовой застройки.

Площадь Помещения

- Влияет нелинейно: маленькие квартиры дороже за м² (из-за ликвидности), а большие — дешевле.
- Например, студия может стоить дороже за м², чем трёхкомнатная квартира.

Кол-во комнат

- Меньшее количество комнат → выше цена за м².
- Больше комнат → ниже цена за м², но выше в абсолютных значениях.

Тип Недвижимости

- Квартиры, как правило, дешевле за м², чем апартаменты.
- Апартаменты чаще расположены в центре или рядом с метро, что повышает их стоимость.

Документы Проверены, Росреестр Проверено

- Если документы проверены, риск покупки ниже → цена выше.
- Покупатели готовы переплачивать за юридическую чистоту.

Супер Агент

- Объявления от проверенных агентов могут иметь более высокую цену, так как агент гарантирует качество сделки.

- Повышенный уровень доверия к продавцу может повышать стоимость объекта.

Срочная Продажа

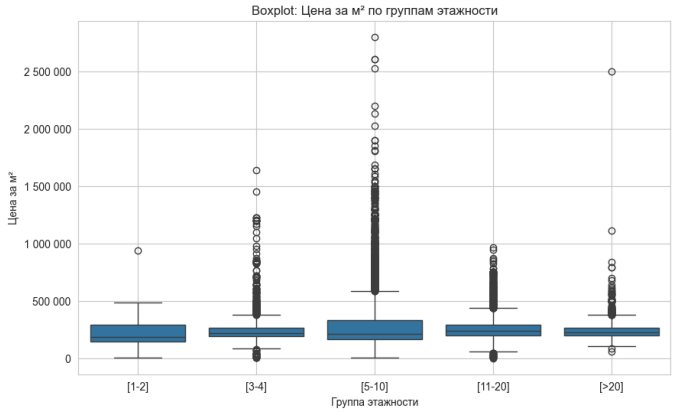
- Срочная продажа означает, что продавец готов делать скидку ради быстрой сделки.
- Цена за м² в таких случаях ниже, чем у аналогичных объектов.

Группа Этажности, Группы Этажности

- Позволяют тестировать гипотезы о влиянии этажности на стоимость.
- Например, возможно, в домах до 5 этажей цена выше, чем в многоэтажках.

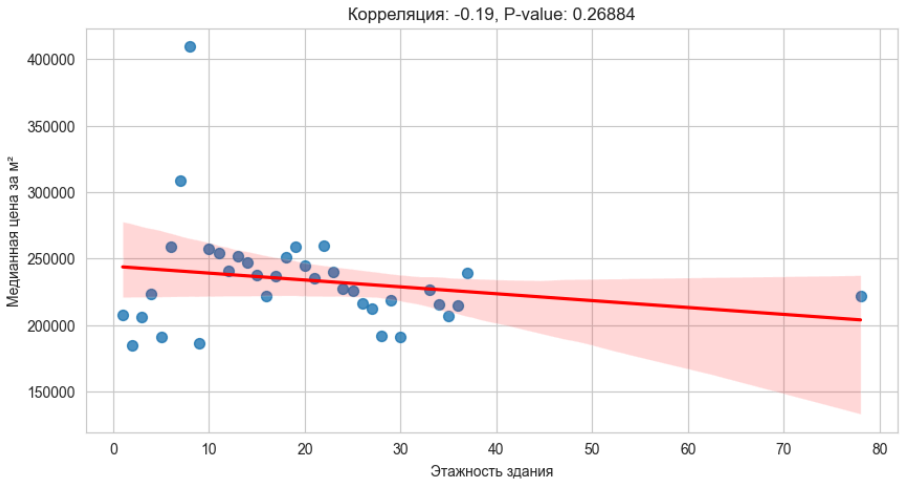
Проверка гипотез

Чем выше этажность, тем ниже стоимость квадратного метра



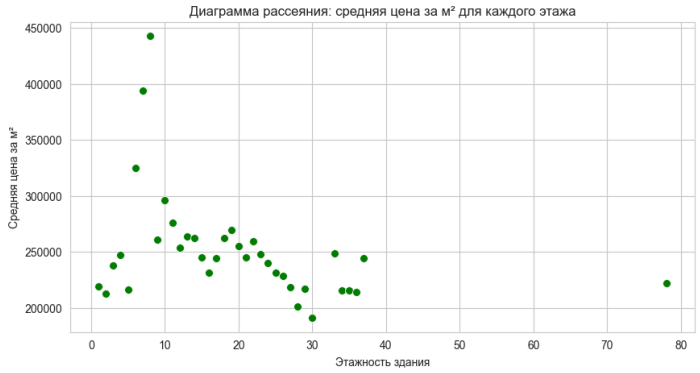
Boxplot показывает, что средняя цена по данным категориям этажности примерно совпадает, но для более низких зданий мы видим большую дисперсию, поэтому это нужно учитывать при дальнейшем анализе. Диаграмма рассеяния также подтверждает тот факт, что высокое отклонение безусловно стоит принимать во внимание в дальнейшем.

Корреляция между этажностью здания и медианной ценой за м²: -0.19
P-value: 0.26884



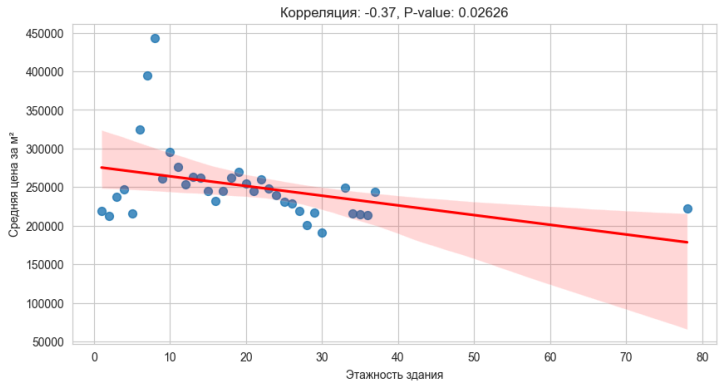
Мы провели корреляционный анализ изменения медианной цены в зависимости от этажности. Диаграмма рассеяния показывает, что зависимости нет. Расчетная корреляция подтверждает вышесказанное. Также нужно отметить, что p-value очень высокий (0,27)

Далее считаем среднюю цену за м² для каждого этажа



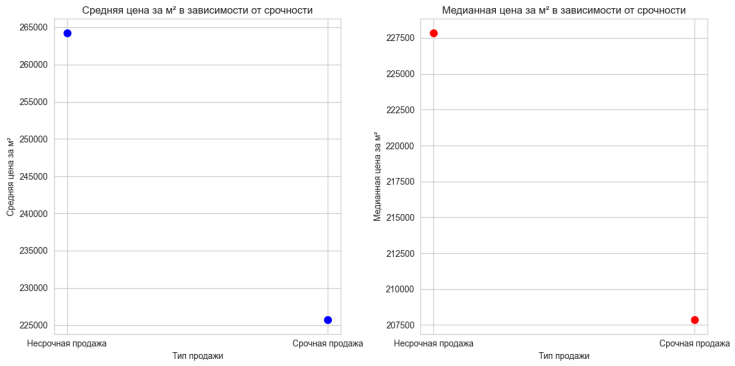
Для чистоты проверки гипотезы проверим, что средняя цена может ошибочно показывать тренд. На графике видим, что тренд определенно присутствует. Необходимо построение линии тренда и расчет корреляции.

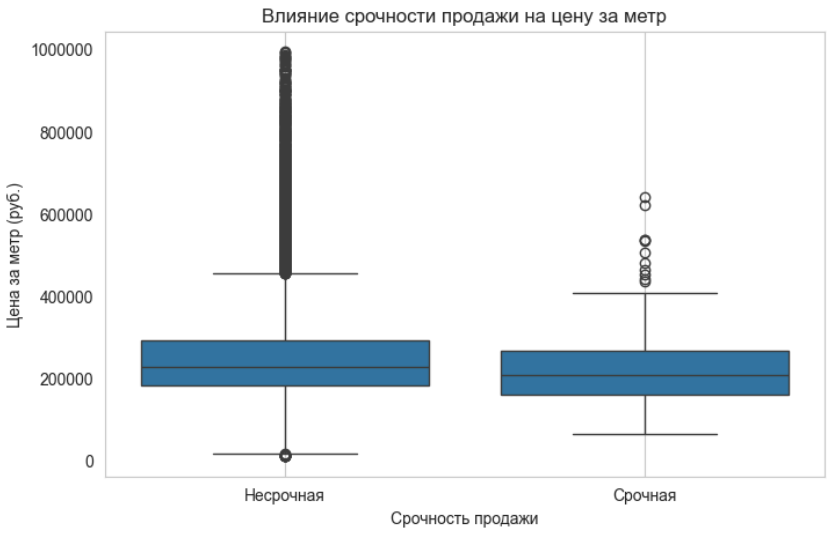
Корреляция между этажностью здания и средней ценой за м²: -0.37
P-value: 0.02626



Тестирование показало, что между средней ценой за квадратный метр и этажностью имеется умеренная зависимость. Однако, учитывая, что дисперсия у нас достаточно высокая нам лучше оценивать при помощи медианной цены. Исходя из всего вышесказанного мы отвергаем гипотезу.

Объявления со срочной продажей имеют более низкую цену





Влияние срочности продажи на цену за метр

1. **Медианное значение цены за м² у срочной продажи ниже**, чем у несрочной. Это подтверждает гипотезу о том, что продавцы, которым нужно быстрее продать недвижимость, чаще снижают цену для быстрой продажи недвижимости и привлечения большего внимания потребителей .
2. **Разброс цен у несрочной продажи выше** – присутствует большее количество выбросов и высокая вариативность, что говорит о наличии как дешёвых, так и очень дорогих объектов.
3. **У срочной продажи меньше экстремальных выбросов**, что может означать более ограниченный диапазон ценовых предложений – вероятно, дорогие квартиры реже продаются срочно.
4. **Общий уровень цен у срочных продаж несколько ниже**, что соответствует ожиданиям: покупатели ожидают скидку при срочной продаже , но разница не сильно большая .

1. **Распределение асимметрично и имеет длинный правый хвост**

- Большинство объектов сосредоточено в диапазоне 100 000 – 300 000 руб./м².
- Есть небольшая часть объявлений с очень высокой ценой за м² (люксовая недвижимость, премиальные новостройки).

2. **Срочные продажи (оранжевая линия) встречаются реже, но следуют тому же распределению**
   - Основная масса срочных продаж также находится в диапазоне 100 000 – 300 000 руб./м².
   - Однако их сильно меньше по сравнению с несрочными предложениями.
3. **Срочные продажи сосредоточены в среднем диапазоне цен**
   - Нет сильного смещения срочных продаж в сторону очень низких цен.
   - Это подтверждает, что срочная продажа снижает цену, но не делает объект заведомо дешевле всего рынка.Понятно , что будут ощутимые скидки , но не более того
4. **Пик распределения**
   - Большая часть объявлений имеет цену за м² около 200 000 руб./м², что соответствует среднему ценовому сегменту.

Итог:

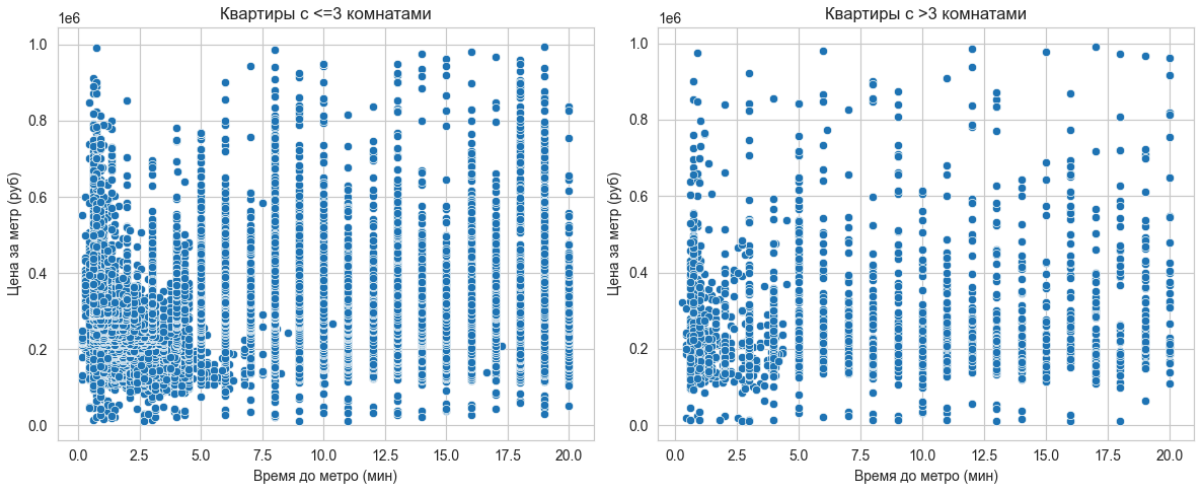
1. Срочная продажа снижает цену, но не радикально — срочные предложения имеют схожее распределение с несрочными, но в меньшем количестве.
2. На рынке больше несрочных предложений, они доминируют в распределении.
3. Высокие цены встречаются редко, рынок больше ориентирован на массовый сегмент.

```
Статистика различий цен за метр по срочности продажи:
                  count           mean            std      min       25%  \
Срочная_продажа
0                 28687.0  256940.401506  118689.916365  10412.0  182000.0
1                   301.0  225750.003322   91609.404970  64212.0  161335.0

                     50%       75%       max
Срочная_продажа
0                 227106.0  291652.0  994500.0
1                 207880.0  266942.0  640635.0

T-статистика: -5.86, p-value: 0.0000
Разница статистически значима: срочная продажа влияет на цену за метр.
```

Гипотеза: Для квартир с <=3 комнат влияние удаленности от метро на цену выше, чем для квартир с большим количеством комнат



Для квартир с <=3 комнатами:
Корреляция Пирсона: 0.200
p-value: 0.000

Для квартир с >3 комнатами:
Корреляция Пирсона: 0.101
p-value: 0.000

Линейная регрессия:
Для <=3 комнат: наклон = 3786.32, p-value = 0.00000
Для >3 комнат: наклон = 2955.98, p-value = 0.00001

1. Корреляция между числом комнат и ценой за квадратный метр слабая, но значимая
   - Для квартир с ≤ 3 комнатами корреляция Пирсона 0.200, что указывает на слабую положительную связь (по мере увеличения числа комнат цена за м² растёт, но не сильно).
   - Для квартир с > 3 комнатами корреляция ещё слабее (0.101), но всё равно значима (p-value = 0.000).
2. Зависимость между числом комнат и ценой за м² слабее для больших квартир
   - В квартирах с ≤ 3 комнатами изменение количества комнат больше влияет на цену за м², чем в квартирах с > 3 комнатами. - Это может объясняться тем, что в большем жилье стоимость больше определяется общей площадью, чем количеством комнат. И сейчас заметен рост спроса на студии или однокомнатные квартиры , чем на 3 комнатные , так как цены между этими сильно разняться
3. Линейная регрессия подтверждает разницу в влиянии количества комнат
   - Наклон (коэффициент перед числом комнат) 3786.32 для ≤ 3-комнатных квартир → добавление ещё одной комнаты увеличивает цену за м² в среднем на 3 786 рублей.
   - Наклон 2955.98 для > 3-комнатных квартир → влияние числа комнат на цену за м² меньше.
   - Оба значения p-value очень малы (близки к нулю), что означает, что эти результаты статистически значимы.

Вывод:

1. Влияние количества комнат на цену за квадратный метр заметнее в квартирах до 3 комнат.

2. В больших квартирах (>3 комнат) цена за м² растёт слабее, что говорит о том, что покупатели таких квартир ориентируются больше на общую стоимость объекта, а не на цену за квадратный метр.
3. Вероятно, в элитном и просторном жилье другие факторы (расположение, видовые характеристики, статус дома) оказывают большее влияние на цену, чем просто количество комнат.

Гипотеза: Чем сильнее загружена ближайшая станция, тем ниже стоимость квадратного метра

Корреляция:
Пирсон: корреляция = 0.187, p-value = 0.000
Спирмен: корреляция = 0.182, p-value = 0.000

Линейная регрессия:
Наклон = 1141.32, p-value = 0.000

Тест Краскела-Уоллиса (сравнение цен по группам загруженности):
Статистика = 1145.41, p-value = 0.000

1. Корреляционный анализ

Для исследования связи между загруженностью станции метро и ценой недвижимости за квадратный метр были рассчитаны коэффициенты корреляции Пирсона и Спирмена:

Корреляция Пирсона: 0.187 – указывает на слабую положительную линейную связь.

Корреляция Спирмена: 0.182 – подтверждает аналогичный вывод для монотонных зависимостей.

Оба p-value = 0.000, что говорит о статистической значимости выявленной связи.

Вывод: Загруженность метро слабо, но статистически значимо, связана с ценой за квадратный метр. Это может свидетельствовать о том, что более загруженные станции расположены в районах с более доступным жильем.

2. Линейная регрессия

Для оценки влияния загруженности метро на стоимость жилья была построена линейная регрессионная модель, в которой загруженность станции выступала в качестве предиктора.

Наклон коэффициента: 1141.32 – это означает, что при увеличении ранга станции (то есть снижении её загруженности) цена за квадратный метр возрастает в среднем на **1 141 рубль**.

p-value = 0.000, что подтверждает статистическую значимость влияния.

Вывод: В среднем квартиры рядом с менее загруженными станциями **дороже**, однако разница не является существенной.

3. Тест Краскела-Уоллиса

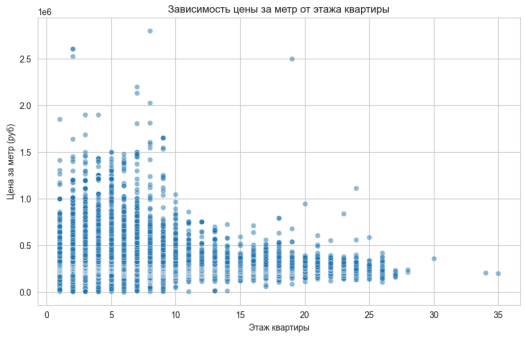
Для проверки различий между группами по загруженности метро был применен тест Краскела-Уоллиса.

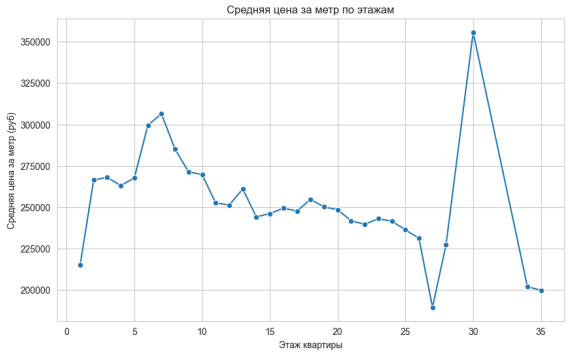
Статистика теста: 1145.41

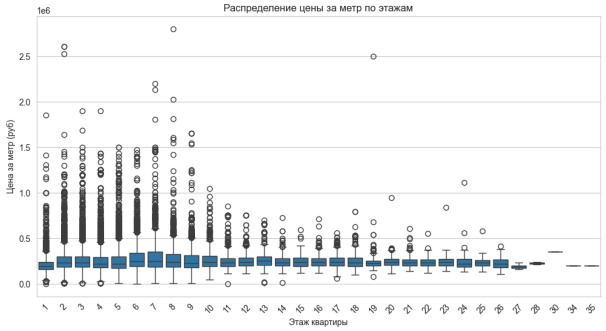
p-value = 0.000, что указывает на статистически значимые различия между группами.

Вывод: Цена за квадратный метр **отличается в районах с разным уровнем загруженности метро**, однако влияние этого фактора **не является определяющим**.

Гипотеза : Чем выше этажность квартиры, тем стоимость квадратного метра выше







Корреляция Пирсона:
Коэффициент корреляции: 0.002
p-value: 0.670

Корреляция Спирмена:
Коэффициент корреляции: 0.104
p-value: 0.000

Тест Краскела-Уоллиса:
Статистика: 898.56
p-value: 0.000

Корреляционный анализ был проведён с использованием коэффициентов Пирсона и Спирмена для выявления зависимости между этажностью здания и ценой за квадратный метр.

- **Корреляция Пирсона** составила **0.002** при **p-value = 0.670**. Это указывает на отсутствие линейной связи между этажностью здания и стоимостью жилья. Высокое значение p-value свидетельствует о том, что данная связь не является статистически значимой.
- **Корреляция Спирмена** равна **0.104**, а **p-value = 0.000**, что говорит о наличии небольшой положительной монотонной связи между этажностью и ценой за квадратный метр. Несмотря на слабый коэффициент (0.104), низкое p-value подтверждает статистическую значимость данной зависимости.

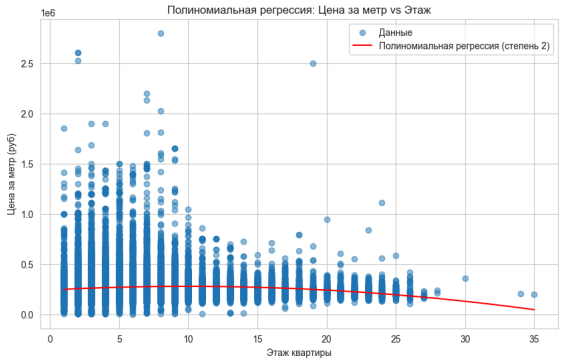
Вывод: В данных наблюдается незначительная тенденция к росту стоимости квадратного метра с увеличением этажности здания, однако эта связь крайне слабая и нелинейная.

Тест Краскела-Уоллиса (сравнение групп по этажности)

Для дополнительного анализа была проведена непараметрическая проверка различий в ценах между группами этажности.

- Статистика теста составила **898.56**, а **p-value = 0.000**, что указывает на статистически значимые различия в цене за квадратный метр между группами этажности.

Вывод: Несмотря на наличие статистически значимых различий, их величина, вероятно, обусловлена другими факторами, такими как возраст здания, его местоположение и класс недвижимости.



Коэффициенты полиномиальной регрессии:
Свободный член: 241725.89
Коэффициенты: [0. 7326.40871244 -369.05304033]

Коэффициенты модели:

Свободный член: 241 725.89

Коэффициенты:

Линейный коэффициент: 7 326.41 (положительный)

Квадратичный коэффициент: -369.05 (отрицательный)

Интерпретация результатов:

1. Линейный коэффициент положителен (7326.41), что означает, что при увеличении этажности цена за квадратный метр растёт.
2. Квадратичный коэффициент отрицателен (-369.05), что указывает на эффект убывающей отдачи: после определённой высоты влияние этажности начинает снижаться, и возможен даже обратный эффект.
3. Свободный член (241 725.89) задаёт базовый уровень цены при нулевой этажности (теоретически, если бы такая существовала).

Вывод:

Полиномиальная регрессия подтверждает наличие нелинейной связи между этажностью и ценой за квадратный метр. До определённого момента этажность увеличивает стоимость жилья, но затем влияние

ослабевает, и на очень высоких этажах может наблюдаться снижение цены. Это может объясняться такими факторами, как сложность доступа, особенности планировки и предпочтения покупателей.

Разработка предсказательных моделей

Линейная регрессия

R^2: 0.28

MAE: 76445.33

MSE: 15388605025.08

OLS Regression Results

==========

Dep. Variable: Цена_за_метр R-squared: 0.290

Model: OLS Adj. R-squared: 0.289

Method: Least Squares F-statistic: 577.8

Date: Wed, 26 Mar 2025 Prob (F-statistic): 0.00

Time: 04:51:39 Log-Likelihood: -2.9827e+05

No. Observations: 22660 AIC: 5.966e+05

Df Residuals: 22643 BIC: 5.967e+05

Df Model: 16

Covariance Type: nonrobust

==========

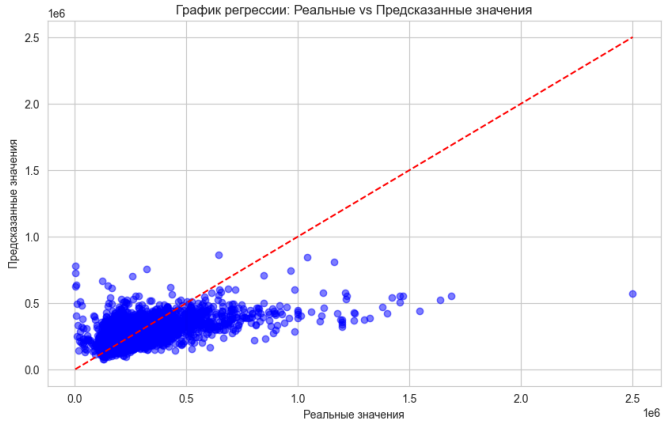
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e+05	9427.255	29.444	0.000	2.59e+05	2.96e+05
Этаж_квартиры	1390.1517	213.955	6.497	0.000	970.786	1809.517
Этажность_здания	-690.3010	342.592	-2.015	0.044	-1361.805	-18.797
Площадь_помещения	1350.2475	27.580	48.957	0.000	1296.189	1404.306
Тип_недвижимости	-5.808e+04	6350.043	-9.146	0.000	-7.05e+04	-4.56e+04

Документы_проверены	-3.116e+04	1917.183	-16.253	0.000	-3.49e+04	-2.74e+04
Супер_агент	-3.635e+04	2384.094	-15.247	0.000	-4.1e+04	-3.17e+04
Росреестр_проверено	-1.958e+04	2103.461	-9.309	0.000	-2.37e+04	-1.55e+04
Кол-во_комнат	-3.41e+04	909.776	-37.481	0.000	-3.59e+04	-3.23e+04
Время_до_метро	3794.1245	144.254	26.302	0.000	3511.376	4076.873
Центральный	7.412e+04	3318.400	22.335	0.000	6.76e+04	8.06e+04
Спальный	-2.675e+04	3378.436	-7.917	0.000	-3.34e+04	-2.01e+04
Прибрежный	-887.5118	2759.894	-0.322	0.748	-6297.094	4522.070
Промышленный	9518.5265	2511.068	3.791	0.000	4596.660	1.44e+04
Срочная_продажа	-3.099e+04	8103.268	-3.824	0.000	-4.69e+04	-1.51e+04
Группа_этажности	-1482.7682	2538.536	-0.584	0.559	-6458.474	3492.937
Нормализованный_ранг_метро	4.353e+04	3244.221	13.418	0.000	3.72e+04	4.99e+04

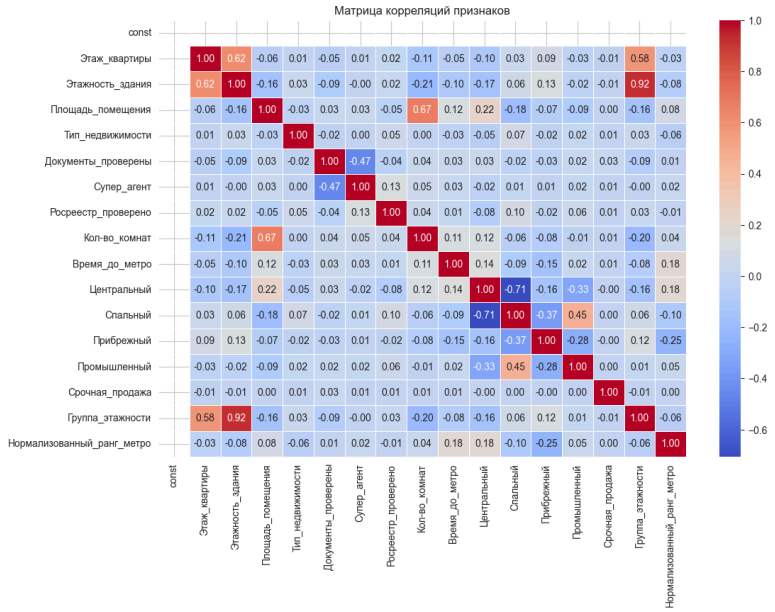
==========

Omnibus:	18495.583	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1205283.687
Skew:	3.484	Prob(JB):	0.00
Kurtosis:	38.043	Cond. No.	958.

==========



Регрессионная модель показала плохую точность предсказания. Как видно на графике реальные значения далеки от предсказанных. Будем строить более сложные модели, позволяющие получить более точные предсказанные значения.Дальше проверяем наше признаки на мультиколлинеарность



Видим, что есть сильная корреляция между этажностью здания и группой этажности, уберем группу этажности, также как ни странно есть зависимость между спальным и центральным районами, так что убираем центральный район.

Пробуем улучшенную регрессионую модела
R^2: 0.27
MAE: 78322.56
RMSE: 125139.95

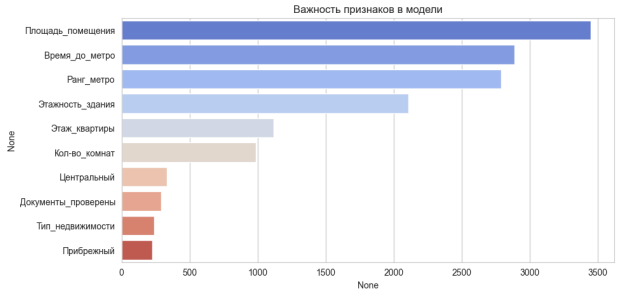
Она показала результаты не лучше. Показатели остались примерно такими же. Попробуем другие модели для анализа

Градиентный бустинг

Градиентный бустинг — мощный инструмент для анализа данных, особенно если зависимость между признаками сложная и нелинейная. Он обладает высокой точностью, умеет работать с пропущенными значениями и позволяет анализировать важность признаков. Это делает его одним из лучших алгоритмов для прогнозирования цен на недвижимость и других задач машинного обучения.

Поэтому с учетом нелинейных связей наших данных было принято решение попробовать данный способ анализа

MAE: 45113.589011290154
RMSE: 83818.66635878052
R^2: 0.6705716353841678



Градиентный бустинг явно лучше справился с задачей, учитвая RAE, RMSE, R^2

Метод DecisionTreeRegressor

Плюсы

1. **Простота в интерпретации**
   Деревья решений легко визуализировать, что делает их удобными для объяснения. Кроме того, модель позволяет интерпретировать, какие признаки наиболее важны для предсказания целевой переменной.
2. **Отсутствие необходимости в масштабировании данных**
   В отличие от линейных моделей, дерево решений не чувствительно к масштабу признаков. Это означает, что данные не требуют предварительной нормализации или стандартизации.
3. **Выявление нелинейных зависимостей**
   В отличие от линейной регрессии, DecisionTreeRegressor способен выявлять сложные нелинейные зависимости между входными признаками и выходной переменной.
4. **Работа с пропущенными значениями**
   Метод позволяет разделять данные даже при наличии пропусков, если используется механизм суррогатных разбиений.
5. **Высокая скорость обучения**
   В сравнении с ансамблевыми методами (например, случайным лесом или градиентным бустингом), обучение одного дерева происходит значительно быстрее, что делает этот метод привлекательным при работе с небольшими наборами данных.

6. **Устойчивость к выбросам**
   Дерево решений может игнорировать выбросы в данных, если они не оказывают значительного влияния на разбиения внутри дерева.

Лучшие параметры: {'max_depth': 13, 'max_features': None, 'min_samples_leaf': 5, 'min_samples_split': 10}

Лучший R^2 на кросс-валидации: 0.5991029232709704

R^2 на тестовой выборке: 0.5684355640327519

RMSE на тестовой выборке: 97937.25406314645

MAE на тестовой выборке: 109958.62204022972

Сравнение и интерпретация результатов.

1. Линейная регрессия
   • R^2: 0.27 — очень низкий показатель, что означает, что модель плохо объясняет зависимость между переменными. Большая часть вариативности в данных остаётся необъяснённой. • MAE (Средняя абсолютная ошибка): 78322.56 — это достаточно высокая ошибка, что также подтверждает плохую предсказательную способность модели. • RMSE (Корень из среднеквадратичной ошибки): 125139.95 — также высокая ошибка, что подтверждает низкое качество модели.
2. Градиентный бустинг
   • R^2: 0.67057 — это намного лучше, чем у линейной регрессии, что говорит о хорошем объяснении вариативности в данных. Модель хорошо подходит для предсказания цен. • MAE: 45113.59 — значительное улучшение по сравнению с линейной регрессией, что означает, что ошибка в предсказаниях значительно ниже. • RMSE: 83818.67 — тоже существенно ниже, чем у линейной регрессии, что указывает на улучшение в предсказаниях.
3. Лес
   • R^2 на кросс-валидации: 0.5991 — довольно хорошее значение, но не такое высокое, как у градиентного бустинга. Это говорит о том, что модель случайного леса достаточно хорошо обобщает данные, но всё же не так хорошо, как градиентный бустинг. • R^2 на тестовой выборке: 0.5684 — немного ниже, чем на кросс-валидации, что может указывать на небольшое переобучение (overfitting) или на сложности в тестовой выборке. • RMSE на тестовой выборке: 97937.25 — значительно выше, чем у градиентного бустинга, что указывает на менее точные предсказания. • MAE на тестовой выборке: 109958.62 — также хуже, чем у градиентного бустинга, что говорит о большем среднем отклонении предсказаний от реальных значений.

Сравнение моделей:

- Градиентный бустинг — показывает наилучшие результаты среди представленных моделей:

- Высокий R^2 (0.67057) и низкие значения MAE и RMSE показывают, что эта модель лучше всего предсказывает цену за квадратный метр.

- Лес — модель, которая показывает хорошие результаты, но не столь эффективна, как градиентный бустинг, из-за более высоких значений MAE и RMSE на тестовой выборке.

- Линейная регрессия — очевидно, является худшей моделью, поскольку имеет очень низкий R^2, высокие значения MAE и RMSE, что указывает на плохое качество предсказаний.

Рекомендации для улучшений:

1. Градиентный бустинг:

- Это лучшая модель, и её можно использовать как основную. Однако можно попробовать улучшить результаты с помощью более тонкой настройки гиперпараметров (например, с помощью GridSearchCV или RandomizedSearchCV).

- Также стоит обратить внимание на предобработку данных, такую как нормализация или масштабирование, что может повысить точность.

- Рассмотреть использование других методов, таких как XGBoost или LightGBM, которые могут дать ещё лучшие результаты, чем стандартный градиентный бустинг.

2. Случайный лес:

- Возможно, улучшение гиперпараметров модели (например, увеличение числа деревьев или настройка максимальной глубины деревьев) может привести к улучшению результатов.

- Также можно проверить, не страдает ли модель от переобучения, например, путем увеличения количества данных для обучения.

3. Линейная регрессия:

- Линейная регрессия явно не подходит для ваших данных. Можно либо использовать более сложные модели, такие как полиномиальная регрессия, либо применить более сложные методы, такие как Lasso или Ridge регрессия, если данные имеют мультиколлинеарность.

- Возможно, потребуется более тщательная предобработка данных (например, удаление выбросов или трансформация признаков).

Возможные недостатки и улучшения:

- Переобучение (Overfitting): Некоторые модели (например, случайный лес) могут переобучаться на тренировочных данных, что отражается в меньшем R^2 на тестовых данных. Стоит проверить, используются ли подходящие методы регуляризации или настроены гиперпараметры для предотвращения переобучения.

- Отбор признаков: Возможно, стоит улучшить отбор признаков. Модели могут быть улучшены, если отобрать наиболее значимые признаки и исключить шум или избыточные данные.

Заключение:

- Градиентный бустинг — это наилучшая модель для нашего набора данных, и её стоит использовать в качестве основной.
- Для других моделей стоит провести дополнительную настройку гиперпараметров или попробовать другие методы предсказания, такие как XGBoost, LightGBM или полиномиальная регрессия.

Источники

1. *Дубровский В.Ж., Орехова С.В., Ярошевич Н.Ю. (2019). "Анализ влияния транспортной инфраструктуры на стоимость жилой недвижимости в мегаполисах России"*

2. *Zheng S., Hu X., Wang J., Wang R. (2021). "The capitalization of subway access in residential property values: Evidence from Beijing"* Received from: https://ideas.repec.org/a/eee/transa/v80y2015icp104-115.html

3. *Стерник Г.М., Стерник С.Г. (2020). "Методология моделирования рынка недвижимости"* Received from: https://ibooks.ru/products/369650?category_id=2331

4. *Попов А.А., Косарева Н.Б. (2018). "Жилищная экономика: современные подходы и методы анализа"*

5. *Sopranzetti B.J., Vandell K.D. (2021). "Do layout and design matter? A hedonic analysis of apartment features"*

6. *Аксенов П.Л., Родионова Н.В. (2022). "Трансформация потребительских предпочтений на рынке жилой недвижимости после пандемии"*

7. *Красильникова Е.В., Федотова М.А. (2019). "Многофакторная модель оценки стоимости жилой недвижимости в крупных городах России"*
8. *Liu C.H., Rosenthal S.S., Strange W.C. (2020). "The Vertical City: Rent Gradients and Spatial Structure"* Received from: https://www.albany.edu/sites/default/files/2019-08/VerticalCities_3_27_2016.pdf

9. *Rosen S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"* Received from: https://www.scirp.org/reference/referencespapers?referenceid=1956852

10. *Глазунов С.Н., Самошин В.С. (2021). "Поведенческая экономика в сфере недвижимости: иррациональные факторы принятия решений"*

11. *Anselin L., Lozano-Gracia N. (2018). "Spatial Hedonic Models"* Received from: https://www.researchgate.net/publication/304604469_Spatial_Hedonic_Models

12. *Brueckner J.K. (2019). "Urban Economics: Theory and Applications"* Received from: https://www.academia.edu/36570119/Lectures_on_urban_economics

13. *Васильева И.В., Алексеева Л.И., Соколов Е.А. (2023). "Обзор рынка недвижимости в России и за рубежом после пандемии и мобилизации."*. Received from: https://cyberleninka.ru/article/n/obzor-rynka-nedvizhimosti-v-rossii-i-za-rubezhom-posle-pandemii-i-mobilizatsii

14. *Ивойлов И.А. (2022). "Структура и динамика рынка загородной недвижимости России."* Received from: https://cyberleninka.ru/article/n/struktura-i-dinamika-rynka-zagorodnoy-nedvizhimosti-rossii/viewer

15. *Капралин С.Г. “Ценообразование и ценообразующие факторы на рынке недвижимости.”* Received from: https://cyberleninka.ru/article/n/tsenoobrazovanie-i-tsenoobrazuyuschie-faktory-na-rynke-nedvizhimosti/viewer

16. *Сироткин В.А., Желенкова В.С., Кожевникова О.С., Чикурова А.М. (2019). “Роль многофакторного моделирования в оценке стоимости жилой недвижимости и прогнозировании потребительского спроса .”* Received from: https://cyberleninka.ru/article/n/rol-mnogofaktornogo-modelirovaniya-v-otsenke-stoimosti-zhiloy-nedvizhimosti-i-prognozirovanii-potrebitelskogo-sprosa/viewer

17. http://localhost:8888/lab/tree/cian_choose_data%20(1).ipynb
18. http://localhost:8888/lab/tree/cian1_parsing%20(1).ipynb
19. http://localhost:8888/lab/tree/cian2_processing.choose_data%20(1).ipynb
20. http://localhost:8888/lab/tree/cian3_hypotheses%20(1).ipynb
21. http://localhost:8888/lab/tree/cian4_ml.predict%20(1).ipynb