МИНОБРНАУКИ РОССИИ
ФГБОУ ВО «ИжГТУ имени М.Т. Калашникова»
Факультет «Информационные технологии»
Кафедра «Программное обеспечение»

Направление 09.03.04 «Программная инженерия»
Профиль «Разработка программно-информационных систем»

ОТЧЕТ ПО ПРАКТИКЕ

Учебная практика
«Научно-исследовательская работа
(получение первичных навыков научно-исследовательской работы)»

Место прохождения практики и структурного подразделения:
кафедра «Программное обеспечение»

Выполнил обучающийся

ФИО: Афанасьев Павел Юрьевич

Курс: 3

Номер группы: Б22-191-1з

Дата сдачи отчета: «_____» _____ 202 __ г.

Дата аттестации: «_____» _____ 202 __ г.

Оценка _____

Руководитель практики от ИжГТУ имени М.Т. Калашникова
_____ Леонов М.В., д.э.н, доцент, заведующий кафедрой ПО
подпись

Заведующий кафедрой
_____ Леонов М.В., д.э.н, доцент, заведующий кафедрой ПО
подпись

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Учебная практика "Научно-исследовательская работа (получение первичных навыков научно-исследовательской работы)" является важной частью образовательного процесса и направлена на формирование у студентов компетенций в области научных исследований.

Основная цель практики заключается в освоении методов постановки исследовательских задач, обработки данных, анализа научной литературы и представления результатов в виде отчета.

В рамках практики рассматривается вопрос ценообразования на вторичном рынке жилой недвижимости в **г. Барнауле**. Данное исследование является актуальным, так как стоимость жилья оказывает значительное влияние на экономику региона и благосостояние населения. Анализ факторов, определяющих цену недвижимости, может быть полезен для риэлторов, застройщиков, органов государственной власти и потенциальных покупателей.

Для выполнения исследования используется язык программирования Python. В качестве среды разработки применяется PyCharm (разработчик: JetBrains), что обеспечивает удобство написания, тестирования и отладки кода.

ОСНОВНАЯ ЧАСТЬ

1. Обоснование актуальности темы научного исследования

Актуальность темы обусловлена высокой динамичностью рынка недвижимости, влиянием макроэкономических факторов на стоимость жилья, а также необходимостью разработки методов оценки цен с использованием современных инструментов анализа данных.

В последние годы российский рынок недвижимости активно использует современные методы анализа данных, включая машинное обучение и искусственный интеллект, для оценки и прогнозирования цен на жилье. В 2024 году автоматизированная система оценки недвижимости SRG, применяющая ИИ и алгоритмы машинного обучения, проанализировала динамику цен на вторичное жилье в 50 крупных городах России. Результаты показали, что в большинстве городов рост цен составил от 0,02% до 2,06%, при этом в некоторых городах наблюдалось незначительное снижение цен. (https://realty.rbc.ru/news/6780f8669a79472506e0b14d)

Применение методов машинного обучения позволяет учитывать широкий спектр факторов, влияющих на стоимость недвижимости, таких как местоположение, инфраструктура, площадь и состояние объекта. Например, компания Циан разработала модели, способные с высокой точностью оценивать объекты недвижимости и предсказывать сроки их продажи, что способствует более эффективному функционированию рынка. (https://www.cian.ru/analiz-rynka-nedvizhimosti-b2b/)

Кроме того, исследование, опубликованное в 2024 году, продемонстрировало успешное применение кластеризации и метода градиентного бустинга для прогнозирования стоимости жилья в России. Особое внимание уделялось предварительной обработке данных и выбору оптимальных параметров моделей, что позволило повысить точность прогнозов. (https://cyberleninka.ru/article/n/issledovanie-rynka-zhilya-rossiyskoy-federatsii-s-ispolzovaniem-metodov-dobychi-znaniy)

Использование современных методов анализа данных и машинного обучения становится неотъемлемой частью исследований рынка недвижимости в России, позволяя более точно оценивать и прогнозировать динамику цен на жилье.

2. Обзор имеющейся научной литературы по теме исследования

Обзор научной литературы по теме исследования позволяет всесторонне изучить текущее состояние научных знаний о факторах ценообразования на рынке недвижимости и применяемых методах прогнозирования. Анализ показывает, что современные исследования основываются на математическом моделировании, методах машинного обучения и регрессионном анализе.

Бедрина С. А., Пименова М. И. "Применение метода сравнения продаж при определении рыночной стоимости квартиры" (2021, с. 201-202) рассматривают метод сравнения продаж при оценке стоимости недвижимости, что является классическим подходом, широко используемым в рыночной оценке.

Богданова Т. К. "Проблемы моделирования оценки стоимости жилой недвижимости" (2020, с. 7-23) акцентирует внимание на проблемах математического моделирования стоимости жилья и указывает на необходимость учета региональных факторов.

Лебедева Е. В., Ломакин Д. Е. "Применение алгоритмов интеллектуального анализа данных при моделировании стоимости квартир на вторичном рынке жилья" (2021, с. 354-359) демонстрируют применение алгоритмов интеллектуального анализа данных для построения прогнозных моделей, что позволяет повысить точность оценки стоимости квартир.

Васильченко Д. С., Руденко К. С., Чуйко А. В. "Оценка стоимости купли-продажи недвижимости (квартир) в городе Ставрополе в зависимости от расположения на этаже" (2022, с. 5720-5726) анализируют влияние этажности на цену недвижимости, используя методы статистического анализа.

Испольнов Е. В., Горюнов Ю. Ю. "Оценка стоимости жилья в Пензе с помощью нейронных сетей" (2020, с. 155-159) исследуют применение нейронных сетей для оценки стоимости жилья в Пензе, доказывая эффективность машинного обучения в данной сфере.

Лейфер и Черная (2020) проводят сравнительный анализ различных методов машинного обучения для оценки стоимости недвижимости, показывая, что градиентный бустинг демонстрирует наибольшую точность прогнозирования.

Лейфер Л. А., Чёрная Е. В. "Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир" (2020, с. 32-42) проводят сравнительный анализ различных методов машинного обучения для оценки стоимости недвижимости, показывая, что градиентный бустинг демонстрирует наибольшую точность прогнозирования.

Кропотухина Н. А., Хиревич С. А. "Анализ факторов, влияющих на стоимость объектов жилой недвижимости" (2021, с. 134-140) рассматривают

влияние таких факторов, как транспортная доступность, экология, качество инфраструктуры на стоимость жилья, применяя методы регрессионного анализа и кластеризации.

Мамедли М. О., Умнов А. В. "Оценка стоимости недвижимости на основе больших данных" (2022, с. 118-136) анализируют применение методов анализа больших данных для оценки стоимости жилья, что позволяет учитывать широкий спектр факторов, включая макроэкономические и социальные.

Обзор научной литературы подтверждает важность и актуальность применения современных методов анализа данных для прогнозирования стоимости недвижимости. Однако остаются вопросы, связанные с влиянием макроэкономических факторов, социальных характеристик районов и индивидуальных особенностей недвижимости, требующие дальнейшего изучения.

3. Формирование исследовательской гипотезы

3.1 Местоположение и цена недвижимости

Для формирования гипотезы о влиянии месторасположения на цену недвижимости необходимо учитывать степень доступности района, его близость к центру и транспортную инфраструктуру. Эта гипотеза основывается на предположении, что объекты, располагающиеся в более выгодных местах, имеют большую привлекательность, что обуславливает более высокую цену. Это предположение важно протестировать с использованием доступных данных о ценах на недвижимость и географическом положении объектов.

Таблица №1 «Гипотеза: месторасположение и цена недвижимости»

Гипотеза	"Чем ближе объект недвижимости расположен к центру города, тем выше его стоимость на вторичном рынке жилья."
Обоснование	Местоположение недвижимости считается одним из самых важных факторов, влияющих на ее стоимость. Центр города часто является более престижным местом с более развитой инфраструктурой, лучшими возможностями для работы и досуга. Люди готовы платить больше за удобство и близость к ключевым объектам города.
Методы проверки	Для проверки гипотезы можно проанализировать данные о стоимости квартир в разных районах города, сопоставив цены с расстоянием от центра. Важно учитывать площадь, тип недвижимости, этажность и другие параметры, чтобы сделать анализ более точным.
Предсказания	Если гипотеза верна, то можно ожидать, что квартиры, расположенные ближе к центру, будут стоить дороже, а при удалении от центра цена будет снижаться.

3.2 Престижность района и цена на жилье

Престижность района может значительно повлиять на стоимость жилья. В данном случае гипотеза заключается в том, что более престижные районы, характеризующиеся хорошей репутацией, развитой инфраструктурой и высоким уровнем безопасности, способствуют увеличению цен на недвижимость. Анализируя данные о ценах в различных районах и их социально-экономических характеристиках, мы сможем проверить данное предположение.

Таблица №2 «Гипотеза: престижность района и цена на жилье»

Гипотеза	"Повышение уровня престижности района приводит к увеличению цен на вторичное жилье в этом районе."
Обоснование	Престижность района, как правило, ассоциируется с высокой социальной статусностью, развитой инфраструктурой и улучшенными условиями жизни. Люди, стремящиеся улучшить свои условия жизни, чаще выбирают более престижные районы, что увеличивает спрос и соответственно цену.
Методы проверки	Для проверки гипотезы можно провести анализ цен на недвижимость в различных районах, которые могут быть различимы по статусу и престижности. Важно учитывать социально-экономические особенности районов (наличие элитных учебных заведений, престижных магазинов, уровень безопасности).
Предсказания	В престижных районах цена на вторичное жилье будет выше, чем в районах с менее высоко оцененной статусностью, даже если по прочим характеристикам они аналогичны.

3.3 Развитость инфраструктуры и ценовые колебания

Гипотеза о том, что развитость инфраструктуры прямо влияет на ценовые колебания в регионе, требует внимания к таким факторам, как транспортная доступность, наличие образовательных учреждений и медицинских услуг. Ожидается, что районы с более развитой инфраструктурой будут характеризоваться стабильностью и возможно более высокими ценами. Эмпирическая проверка этого предположения позволит выяснить степень влияния данных факторов на рыночную стоимость жилья.

Таблица №3 «Гипотеза: развитость инфраструктуры и ценовые колебания»

Гипотеза	"Наличие развитой транспортной инфраструктуры и социальных объектов в районе ведет к повышению цен на жилую недвижимость."
Обоснование	Развитая транспортная инфраструктура и наличие социальных объектов, таких как школы, магазины, медицинские учреждения, создают дополнительные удобства для жителей района. Эти факторы могут быть важными при выборе места для покупки жилья, поскольку обеспечивают удобный доступ к необходимым услугам.

Методы проверки	Для проверки гипотезы можно провести анализ цен на жилье в районах с различной степенью развитости инфраструктуры. Также можно использовать данные о наличии метро, автобусных маршрутов, школ, больниц и других объектов, которые могут повлиять на удобство жизни в этом районе.
Предсказания	В районах с высокой доступностью общественного транспорта, близостью к школам и другим социальным объектам цены на вторичное жилье будут выше по сравнению с районами, где такие объекты отсутствуют или имеют низкую доступность.

3.4 Экологическая обстановка и цены на жилье

Экологическая обстановка может играть ключевую роль в формировании цен на недвижимость. Гипотеза состоит в том, что более благоприятная экологическая среда способствует повышению стоимости жилья. Для проверки данного предположения потребуется анализ данных о состоянии окружающей среды и ценах недвижимости в различных районах, чтобы установить наличие причинно-следственной связи.

Таблица №4 «Гипотеза: экологическая обстановка и цены на жилье»

Гипотеза	"Негативная экологическая обстановка в районе снижает стоимость недвижимости на вторичном рынке."
Обоснование	Экологическая ситуация в районе может существенно повлиять на здоровье и благополучие жителей. Плохая экология может отпугивать покупателей, особенно тех, кто заинтересован в долгосрочном проживании. В результате спрос на жилье в таких районах снижается, что приводит к падению цен.
Методы проверки	Можно собрать данные о ценах на жилье в районах с различным уровнем загрязнения воздуха, наличием заводов и предприятий, а также с учетом показателей качества окружающей среды (уровень шума, наличие зеленых зон и т. п.). Затем нужно провести анализ корреляции между экологическими параметрами и стоимостью недвижимости.
Предсказания	В районах с плохой экологией цены на недвижимость будут ниже, чем в районах с хорошей экологической обстановкой, даже если прочие характеристики (размер, возраст зданий, престижность района) схожи.

3.5 Состояние жилья и его цена на вторичном рынке

Состояние жилья является одним из основных факторов, определяющих его цену на вторичном рынке. Гипотеза предполагает, что дом с хорошим ремонтом и современными удобствами будет стоить дороже, чем требующий ремонта. Проверка данной гипотезы включает сбор информации о состоянии объектов недвижимости и их рыночных ценах, что поможет выявить закономерности и подтвердить или опровергнуть выдвинутые предположения.

Таблица №5 «Гипотеза: состояние жилья и его цена на вторичном рынке»

Гипотеза	"Хорошее состояние жилого помещения (отремонтированное жилье) ведет к повышению его цены на вторичном рынке недвижимости."
Обоснование	Состояние жилья, включая наличие свежего ремонта, высококачественную отделку и современные коммуникации, значительно влияет на восприятие стоимости. Покупатели, выбирая жилье, часто оценивают не только площадь и расположение, но и готовность объекта к заселению, что увеличивает его стоимость при хорошем состоянии.
Методы проверки	Для проверки гипотезы можно провести анализ цен на квартиры с разным состоянием ремонта. Можно разделить жилье на несколько категорий: новое, с капитальным ремонтом, требующее косметического ремонта и требующее капитальной реставрации.
Предсказания	Если гипотеза верна, то квартиры с новым или свежим ремонтом будут иметь более высокую цену по сравнению с аналогичными по площади и расположению объектами, которые требуют ремонта.

4. Сбор данных для проведения научного исследования

На этом этапе для каждого объекта недвижимости определяются параметры недвижимости.

4.1 Выбор характеристик для сбора данных

Таблица №6 «Характеристики данных»

Наименование	Атрибут в СУБД	Тип данных
Цена	price	Цена объекта недвижимости.
Общая площадь	total_area	Площадь квартиры или дома в квадратных метрах.
Жилая площадь	living_area	Площадь жилых комнат.
Количество комнат	rooms	Число комнат в объекте недвижимости.
Этаж	floor	Этаж, на котором находится квартира.
Общая этажность	total_floors	Общее количество этажей в здании.
Год постройки	year	Год, в котором был построен дом или квартира.
Расстояние до общественного транспорта	distance	Расстояние до ближайшей станции метро, автобусной остановки или другого общественного транспорта.
Наличие балкона или лоджии	balcony	Указание, имеется ли балкон или лоджия.
Состояние ремонта	condition	Уровень ремонта (например, новый ремонт, под ремонт).
Тип недвижимости	type	Тип недвижимости (квартира, дом, студия и т.д.).
Район	district	Местоположение объекта недвижимости (например, центр города, окраина).
Объявление	url	URL-адрес объявления)

4.2 Определение источников данных

Для сбора данных используются открытые и доступные источники: Avito.ru.

Эти сайты содержат объявления о продаже недвижимости и предоставляют подробную информацию о характеристиках объектов, их стоимости и местоположении.

4.3 Сбор данных

В рамках исследования был сформирован набор данных (DataSet) на основе базы данных SQLite, содержащий информацию о 102 объектах жилой недвижимости. Для обеспечения репрезентативности выборки объекты отбирались случайным образом с различных страниц поисковых результатов на агрегаторе объявлений.

Сбор данных осуществлялся с использованием специально разработанного Python-скрипта, который выполнял автоматизированный парсинг информации о недвижимости. Алгоритм извлечения данных включал в себя:

- Анализ структуры веб-страниц и идентификацию ключевых элементов, содержащих целевые параметры объектов.
- Рандомизированный отбор объявлений для формирования выборки, что снижает риск систематической ошибки.
- Логирование выполняемых операций, позволяющее отслеживать процесс сбора данных и выявлять возможные ошибки.

Для повышения надежности полученной информации использовался метод визуализации работы скрипта в браузере, что позволило проводить верификацию корректности извлеченных данных. Отладка и настройка алгоритма осуществлялись на основе анализа лог-файлов, фиксирующих этапы взаимодействия программы с веб-ресурсом.

Полученные данные были сохранены в базе данных SQLite в структурированном виде, что обеспечивает их удобное последующее использование для анализа и обработки. База данных содержит ключевые характеристики объектов, включая стоимость, площадь, этажность, год постройки и расстояние до транспортной инфраструктуры, а также ссылки на оригинальные объявления, что позволяет проверить актуальность собранной информации.

5. Расчет описательной статистики и ее визуализация

5.1 Интерпретация статистики

Для количественного анализа собранных данных о недвижимости используются методы описательной статистики. Они позволяют охарактеризовать основные свойства распределения значений, таких как цена и площадь квартир, выявить закономерности и сделать предварительные выводы перед проведением более сложных статистических исследований.

Ключевыми показателями описательной статистики являются среднее значение, медиана и мода, которые дают представление о типичных значениях исследуемых параметров. Стандартное отклонение, размах и квартили помогают оценить степень разброса данных, а также выявить возможные аномалии или выбросы. В совокупности эти показатели позволяют провести комплексную интерпретацию собранных данных и подготовить их для дальнейшего анализа и визуализации.

Таблица №7 «Описательная интерпретация статистики»

Цена	**Среднее** (Mean)	Средняя цена квартиры
	Медиана (Median)	Цена, которая делит весь набор данных пополам
	Мода (Mode)	Наиболее частая цена квартиры
	Стандартное отклонение (Standard Deviation)	Насколько разбросаны цены вокруг среднего
	Максимум и минимум	Наибольшая и наименьшая цена
	Размах (Range)	Разница между максимальной и минимальной ценой
	Квартили (Quantiles)	Разделяют на 4 группы
Общая площадь	**Среднее** (Mean)	Средняя площадь квартиры
	Медиана (Median)	Центральная площадь квартиры
	Мода (Mode)	Наиболее часто встречающаяся площадь
	Стандартное отклонение (Standard Deviation)	Разброс площадей
	Максимум и минимум	Наибольшая и наименьшая площадь
	Размах (Range)	Разница между максимальной и

		минимальной площадью
	Квартили (Quantiles)	Разбивка на 4 части

5.2 Интерпретации

В ходе анализа данных была произведена интерпретация статистической информации для оценки характеристик недвижимости. Среднее значение дало представление о общей тенденции в ценах и площадях, в то время как медиана указала на центральное значение, минимизируя влияние выбросов. Стандартное отклонение позволило оценить вариативность цен, что важно для понимания рыночной стабильности. Размах и квартили помогли выделить экстремальные значения и понять распределение данных. Данный анализ способствует более глубокому пониманию рыночных условий и формированию обоснованных прогнозов.

Таблица №8 «Интерпретация статистики»

Цена квартиры	**Среднее** (Mean)	Средняя цена квартиры в данном наборе данных составляет 5 млн рублей, что дает представление о том, какова средняя стоимость на рынке.
	Медиана (Median)	Медиана может быть ниже среднего, если в наборе данных есть несколько очень дорогих квартир, что показывает, что распределение цен имеет скошенность.
	Мода (Mode)	Мода укажет, какая цена встречается чаще всего, например, если на рынке часто встречаются квартиры в пределах 4-5 млн рублей.
	Стандартное отклонение (Standard Deviation)	Если стандартное отклонение высокое, это означает, что цена квартир сильно варьируется.
	Максимум и минимум	Они покажут наименьшую и наибольшую цену, что важно для выявления выбросов.
	Размах (Range)	Для **цены квартиры** размах

		покажет, насколько сильно различаются самые дешевые и самые дорогие квартиры в выборке. Если размах очень велик, это может свидетельствовать о наличии высоких выбросов или различий на рынке, например, наличие как дешевых квартир, так и очень дорогих объектов.
	Квартили (Quantiles)	Для **цены квартиры**: если Q1 — 3 млн рублей, медиана — 5 млн рублей, а Q3 — 7 млн рублей, это значит, что 25% квартир стоят до 3 млн рублей, 50% — до 5 млн рублей, и 75% — до 7 млн рублей. Таким образом, 25% самых дорогих квартир стоят выше 7 млн рублей.
Общая площадь	**Среднее** (Mean)	Средняя площадь квартиры может составлять 50 м².
	Медиана (Median)	Медиана может показать, что большая часть квартир в выборке имеет площадь около 45 м², что может быть полезным для понимания рыночного сегмента.
	Мода (Mode)	Мода может показать, что наиболее популярными являются квартиры с площадью 40 м².
	Стандартное отклонение (Standard Deviation)	Это значение покажет, насколько сильно варьируется площадь от среднего.
	Максимум и минимум	Они могут показать диапазон от маленьких студий до более крупных квартир, что поможет увидеть разнообразие объектов

		на рынке.
	Размах (Range)	Для **общей площади квартиры** размах покажет, насколько сильно варьируются площади квартир в выборке. Это может быть полезным для оценки, есть ли в выборке как маленькие квартиры (например, студии), так и большие.
	Квартили (Quantiles)	Для **площади квартиры**: если Q1 — 35 м², медиана — 50 м², а Q3 — 65 м², то 25% квартир имеют площадь меньше 35 м², 50% — меньше 50 м², и 75% — меньше 65 м². Таким образом, 25% самых больших квартир будут иметь площадь больше 65 м².

5.3 Гистограммы (Визуализация данных)

5.3.1 Гистограмма распределения цен

На данной гистограмме представлено распределение цен на квартиры. Большинство объектов концентрируются в среднем ценовом диапазоне, что говорит о популярности среднемаркетных квартир на рынке. Мы видим значительное снижение частоты для более высоких цен, что указывает на редкость более дорогих объектов. Это распределение типично для жилой недвижимости и отражает предпочтения покупателей.

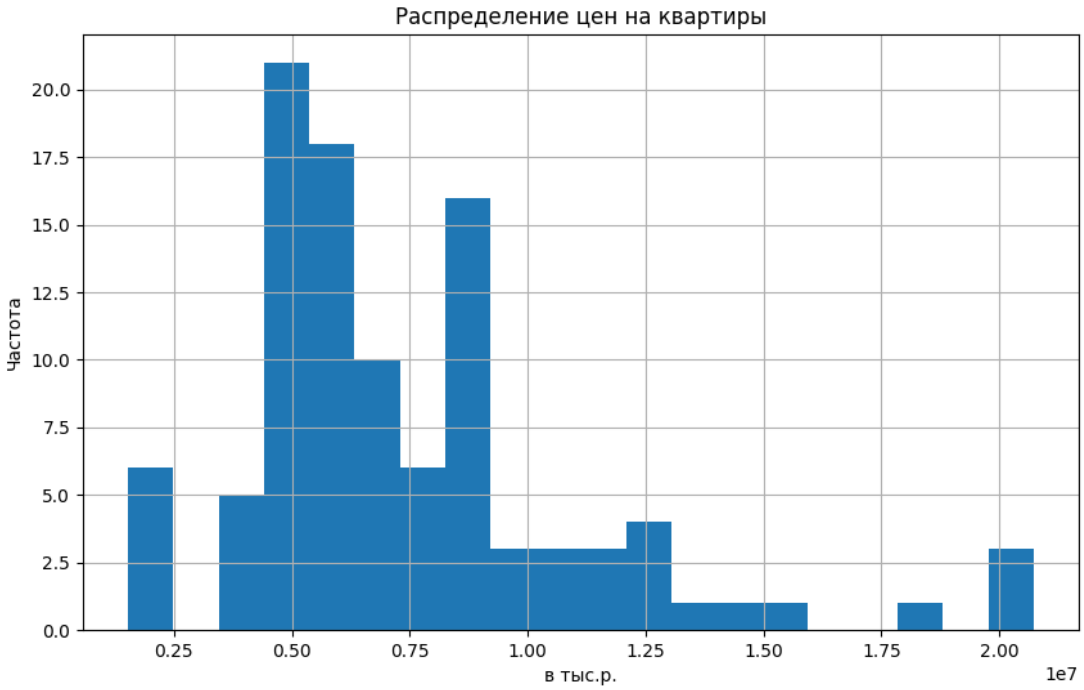

Рисунок 1 Гистограмма распределения цен

5.3.2 Коробчатая диаграмма (box plot)

Коробчатая диаграмма предоставляет визуализацию центральной тенденции и разброса цен на квартиры. Медиана цены расположена ниже среднего уровня, что указывает на наличие более дешевых объектов. Выбросы, представленные точками выше "усов", свидетельствуют о наличии нескольких элитных объектов с ценами, значительно превышающими основную часть данных.

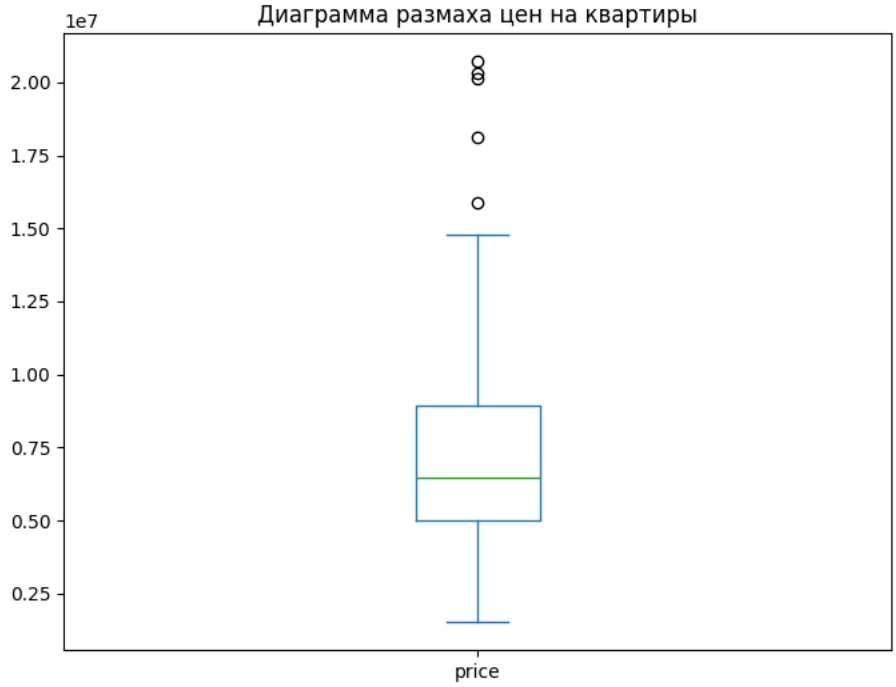

Рисунок 2 Коробчатая диаграмма (box plot)

6. Проведение корреляционного анализа и его визуализация

6.1 Диаграмма рассеяния цены и площади

Диаграмма рассеяния демонстрирует положительную корреляцию между площадью квартиры и её ценой. Как видно из графика, с увеличением площади цена также увеличивается, что соответствует ожиданиям. Эта информация полезна для понимания того, как размер объекта влияет на его рыночную стоимость и может быть использована при стратегическом планировании цен.

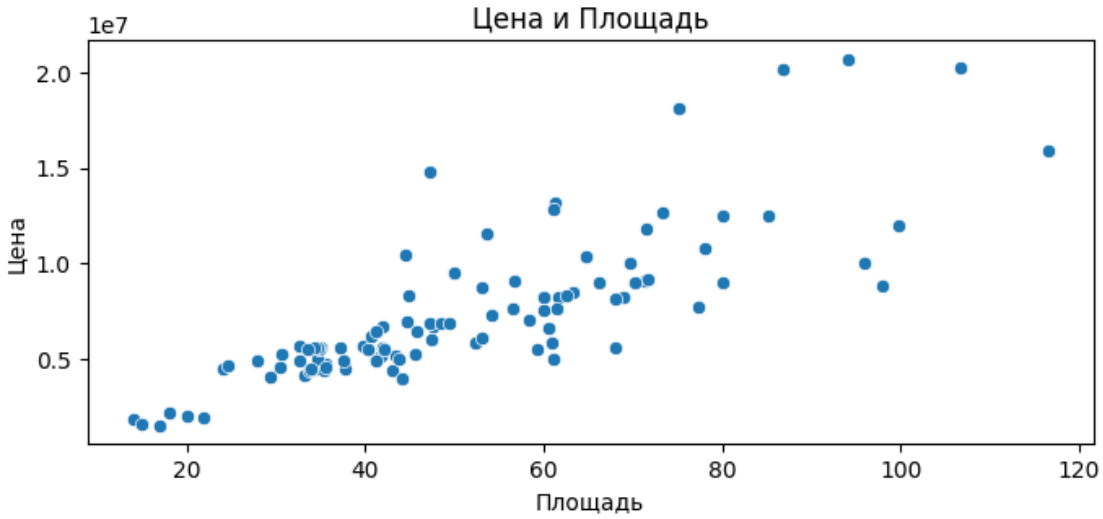

Рисунок 3 Диаграмма рассеяния цены и площади

6.2 Диаграмма рассеяния цены и этажа

Зависимость между ценой и этажом не является очевидной. Цена квартиры может варьироваться в зависимости от этажа из-за таких факторов, как удобство доступа, шумовые характеристики или вид из окна. Например, квартиры на средних этажах могут быть более популярны из-за баланса между удобством и видом, тогда как на верхних этажах цена может повышаться за счёт панорамного вида.

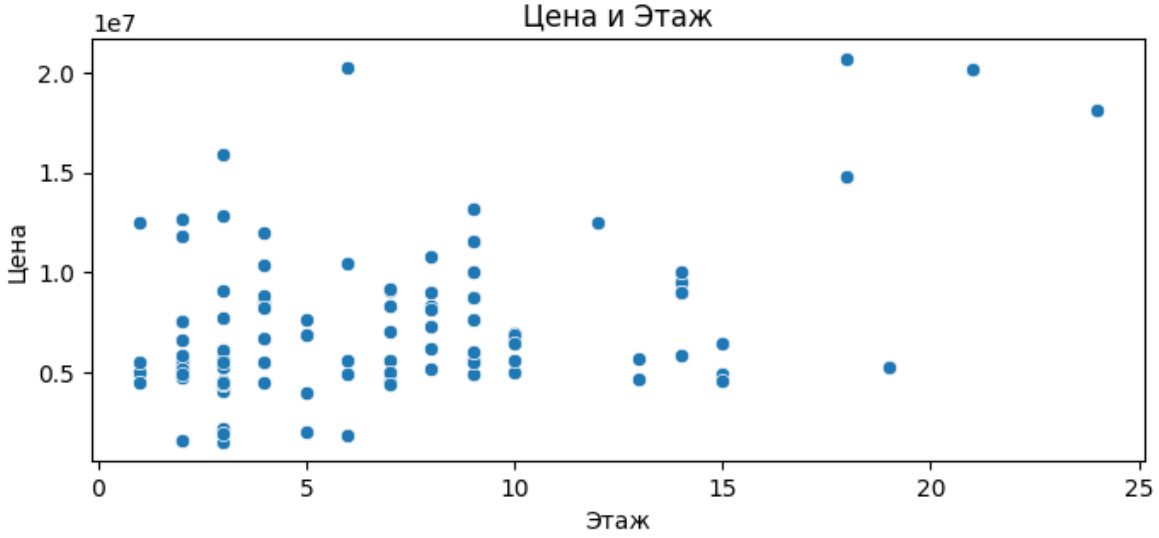

Рисунок 4 Диаграмма рассеяния цены и этажа

6.3 Диаграмма рассеяния цены и расстояния

Наблюдается слабая отрицательная зависимость: с увеличением логарифма расстояния цена, как правило, снижается. Это может быть связано с тем, что ближе к центру города или к важным объектам цена жилья обычно выше из-за большей доступности и развитой инфраструктуры.

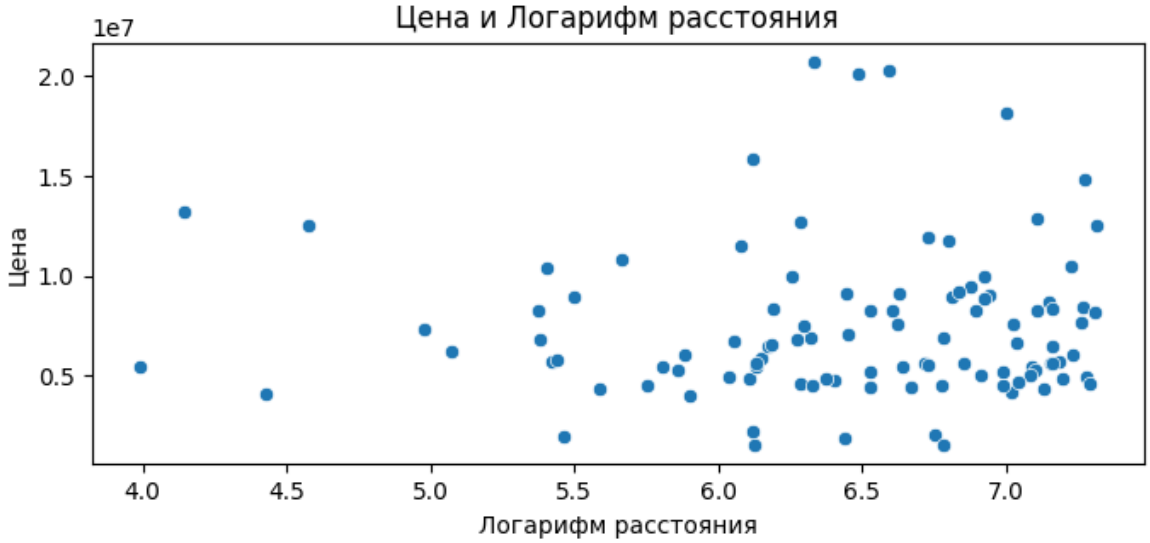

Рисунок 5 Диаграмма рассеяния цены и расстояния

6.4 Диаграмма рассеяния площади и этажа

Видимая зависимость отсутствует. Площадь жилых помещений не демонстрирует явной связи с этажностью. Это может быть связано с разнообразием планировок и архитектурных решений, которые не зависят от конкретного этажа.

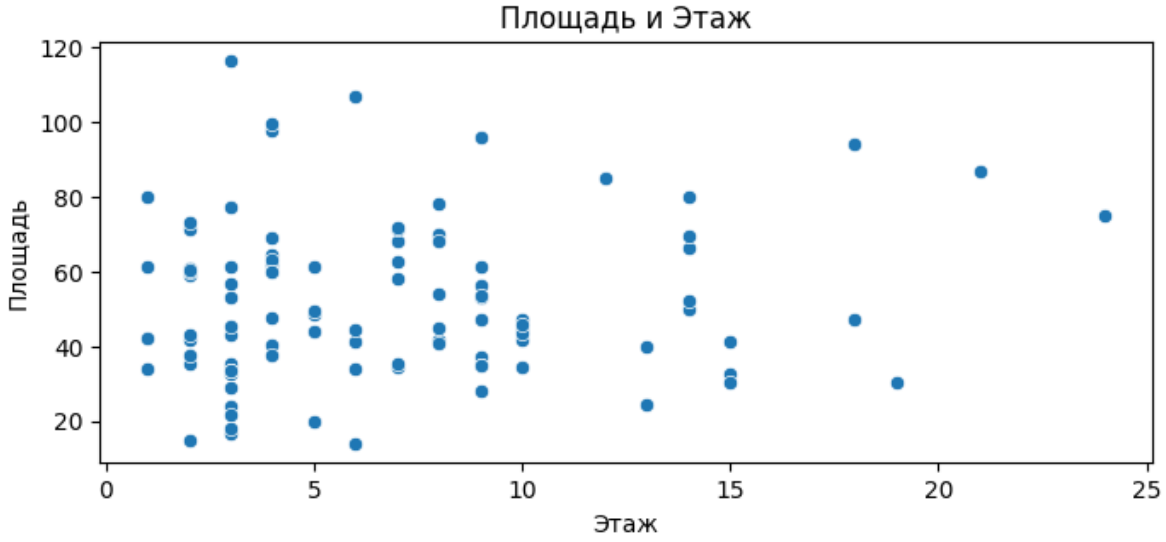

Рисунок 6 Диаграмма рассеяния площади и этажа

6.5 Диаграмма рассеяния площади и расстояния

Связь слабо выраженная или отсутствует. Это может свидетельствовать о том, что размер квартир предлагается равномерно по всему городу, и не зависит существенно от удалённости от центра или ключевых объектов.

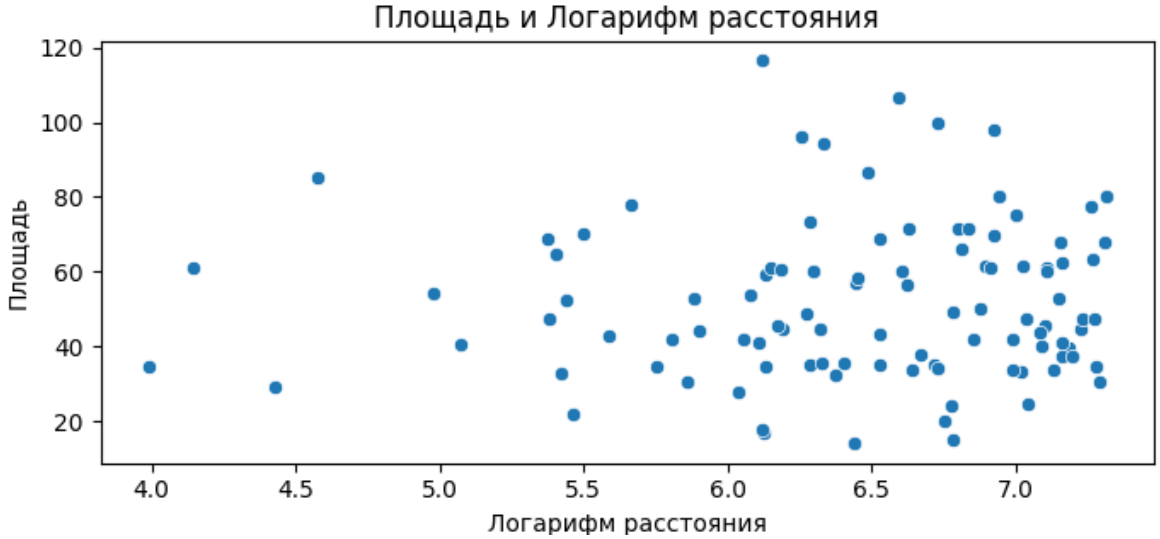

Рисунок 7 Диаграмма рассеяния площади и расстояния

6.6 Диаграмма рассеяния этажа и расстояния

Практически отсутствует зависимость. Это можно объяснить тем, что этажность чаще определяется архитектурными и строительными нормами, а не местоположением здания относительно центра города.

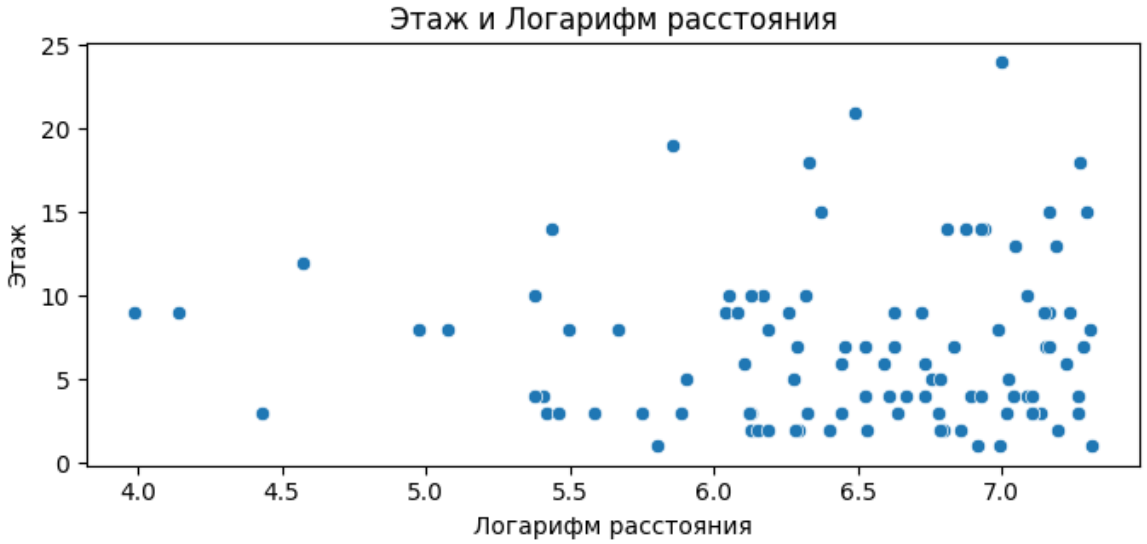

Рисунок 8 Диаграмма рассеяния этажа и расстояния

6.7 Матрица корреляций

Матрица корреляций отображает взаимосвязи между разными переменными, такими как цена, этаж и логарифмическое преобразование расстояния. Из всех переменных log_distance демонстрирует слабую корреляцию с ценой, что может указывать на его меньшую значимость в текущем наборе данных. Это подчеркивает необходимость пересмотра влияния этой переменной на цену или поиск дополнительных факторов, которые могут лучше объяснять вариабельность.

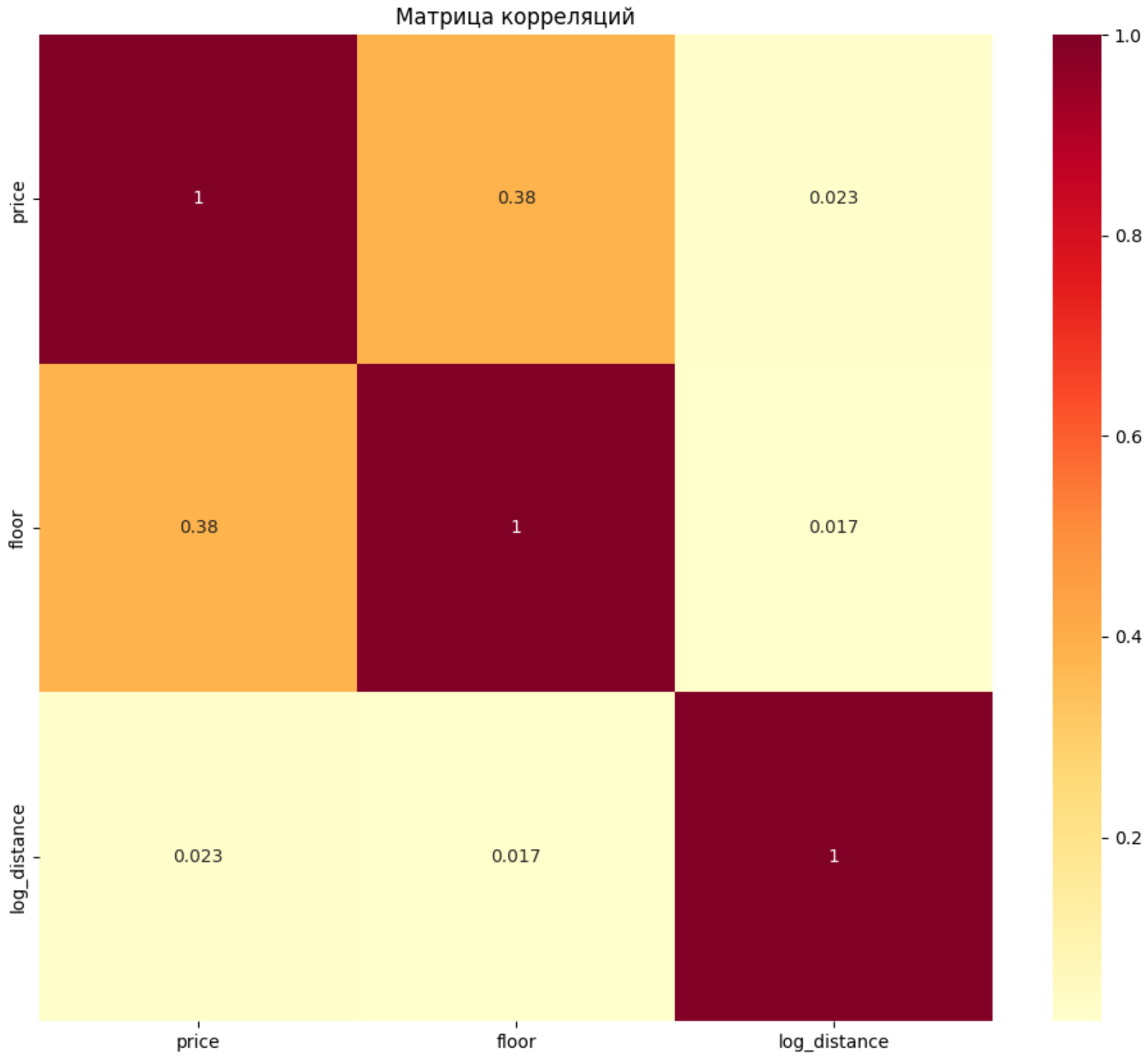

Рисунок 9 Матрица корреляций

6.8 Анализ статистической значимости корреляции

```
Корреляция (цена и square): 0.81, p-значение: 0.0000
Значимая зависимость на уровне 5%.
Корреляция (цена и floor): 0.38, p-значение: 0.0001
Значимая зависимость на уровне 5%.
Корреляция (цена и log_distance): 0.02, p-значение: 0.8158
Нет значимой зависимости на уровне 5%.
```

Рисунок 10 Данные анализа из log

Интерпретация:

Площадь (square): высокая корреляция (0.81) и значимое p-значение указывают на сильную линейную связь между ценой и площадью.

Этаж (floor): средняя корреляция (0.38) и значимое p-значение показывают умеренную, но статистически значимую связь между ценой и этажом.

Логарифм расстояния (log_distance): низкая корреляция (0.02) и высокое p-значение указывают на отсутствие значимой зависимости между ценой и расстоянием.

7. Построение регрессионной моделей стоимости жилья

7.1 Model 1

```
Результаты для Model 1:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.736
Model:                            OLS   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     138.0
Date:                Wed, 12 Mar 2025   Prob (F-statistic):           2.37e-29
Time:                        16:58:45   Log-Likelihood:                -1621.7
No. Observations:                 102   AIC:                             3249.
Df Residuals:                      99   BIC:                             3257.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.467e+06   5.72e+05     -2.565      0.012    -2.6e+06   -3.32e+05
square      1.408e+05   9480.523     14.855      0.000    1.22e+05     1.6e+05
floor        2.35e+05    4.1e+04      5.731      0.000    1.54e+05    3.16e+05
==============================================================================
Omnibus:                       10.748   Durbin-Watson:                   2.228
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               11.556
Skew:                           0.648   Prob(JB):                      0.00310
Kurtosis:                       4.018   Cond. No.                         165.
==============================================================================
```

Рисунок 11 Данные регрессионной модели 1

Интерпретация Model 1

Описание модели:

- Зависимая переменная: цена (price).
- Независимые переменные: площадь (square) и этаж (floor).

Результаты:

- R-squared: 0.736 – указывает, что 73.6% вариации в цене объясняется данной моделью. Это говорит об относительно хорошем уровне объясненной дисперсии.
- Adj. R-squared: 0.731 – корректирует R-squared с учётом числа переменных, что делает его более надёжным показателем.

Коэффициенты:

- const: -1.467e+06. Отрицательный свободный член, который указывает на начальное значение цены без учёта других факторов.

- square: 1.408e+05. Значимый положительный эффект площади на цену. Каждый дополнительный квадратный метр увеличивает цену примерно на 140,800 руб.
- floor: 2.35e+05. Значимый положительный эффект этажа на цену. Каждый дополнительный этаж увеличивает цену примерно на 235,000 руб.

P-значения: P<0.05 для всех коэффициентов, что свидетельствует о их статистической значимости.

Диагностические тесты:

- Omnibus и Jarque-Bera: показатели указывают на незначительное отклонение от нормальности распределения остатков.
- Durbin-Watson (DW): 2.228 – указывает на отсутствие автокорреляции остатков.

Вывод: модель эффективно объясняет вариации цены через площадь и этаж. Оба коэффициента значимы и оказывают положительное влияние на цену.

7.2 Model 2

Результаты для Model 2:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.648
Model:	OLS	Adj. R-squared:	0.641
Method:	Least Squares	F-statistic:	91.28
Date:	Wed, 12 Mar 2025	Prob (F-statistic):	3.39e-23
Time:	16:58:45	Log-Likelihood:	-1636.3
No. Observations:	102	AIC:	3279.
Df Residuals:	99	BIC:	3286.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.087e+05	2.09e+06	0.052	0.959	-4.05e+06	4.26e+06
square	1.469e+05	1.09e+04	13.506	0.000	1.25e+05	1.69e+05
log_distance	-4.411e+04	3.15e+05	-0.140	0.889	-6.68e+05	5.8e+05

Omnibus:	36.606	Durbin-Watson:	2.208
Prob(Omnibus):	0.000	Jarque-Bera (JB):	81.608
Skew:	1.372	Prob(JB):	1.90e-18
Kurtosis:	6.417	Cond. No.	526.

Рисунок 12 Данные регрессионной модели 2

Интерпретация Model 2
Описание модели:

- Зависимая переменная: цена (price).
- Независимые переменные: площадь (square) и логарифм расстояния (log_distance).

Результаты:

- R-squared: 0.648 – указывает, что 64.8% вариации в цене объясняется моделью, что является приемлемым показателем, но ниже, чем в Model 1.
- Adj. R-squared: 0.641 – учитывает число переменных, давая немного скорректированную оценку.

Коэффициенты:

- const: 1.087e+05. Несущественное значение, так как p-value > 0.05.
- square: 1.469e+05. Значимый положительный эффект площади на цену. Увеличение площади на квадратный метр повышает цену на 146,900.
- log_distance: -4.411e+04. Несущественный эффект (p-value = 0.889), логарифм расстояния незначим в объяснении цены.

P-значения: площадь значима ($p < 0.05$), но логарифм расстояния нет.

Диагностические тесты:

- Omnibus и Jarque-Bera: значительные отклонения от нормальности распределения остатков.
- Durbin-Watson (DW): 2.208. Отсутствие автокорреляции остатков.

Вывод: модель показывает значимое влияние площади на цену, но логарифм расстояния не оказывает статистически значимого эффекта. Остатки могут быть не нормально распределены, что стоит учитывать.

7.3 Model 3

```
Результаты для Model 3:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.148
Model:                            OLS   Adj. R-squared:                  0.130
Method:                 Least Squares   F-statistic:                     8.572
Date:                Wed, 12 Mar 2025   Prob (F-statistic):           0.000369
Time:                        16:58:45   Log-Likelihood:                -1681.5
No. Observations:                 102   AIC:                             3369.
Df Residuals:                      99   BIC:                             3377.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4.789e+06    3.2e+06      1.495      0.138   -1.57e+06    1.11e+07
floor       3.027e+05   7.32e+04      4.133      0.000    1.57e+05    4.48e+05
log_distance  8.898e+04   4.89e+05      0.182      0.856   -8.82e+05    1.06e+06
==============================================================================
Omnibus:                       24.242   Durbin-Watson:                   1.749
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.088
Skew:                           1.141   Prob(JB):                     3.96e-08
Kurtosis:                       4.677   Cond. No.                         93.7
==============================================================================
```

Рисунок 13 Данные регрессионной модели 3

Интерпретация Model 3:

Описание модели:

- Зависимая переменная: цена (price).
- Независимые переменные: этаж (floor) и логарифм расстояния (log_distance).

Результаты:

- R-squared: 0.148. Модель объясняет лишь 14.8% вариации в цене, что указывает на слабую объяснительную способность.
- Adj. R-squared: 0.130. Ещё более скорректированная низкая оценка, что свидетельствует о слабой модели.

Коэффициенты:

- const: 4.789e+06. Высокое значение, но незначительно (p-value = 0.138).
- floor: 3.027e+05. Значимый эффект этажа на цену. Дополнительный этаж увеличивает цену на 302,700.

- log_distance: 8.898e+04. Не оказывает статистически значимого влияния на цену (p-value = 0.856).

P-значения: только этаж имеет статистически значимый эффект.

Диагностические тесты:

- Omnibus и Jarque-Bera: значительные отклонения от нормальности остатков.
- Durbin-Watson (DW): 1.749. Указывает на возможную автокорреляцию.

Вывод: эта модель слабо объясняет изменения в цене, но показывает значимость влияния этажа. Логарифм расстояния не оказывает значимого эффекта, и остатки распределены ненормально. Это ограничивает полезность данной модели.

7.4 Model 4

```
Результаты для Model 4:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.736
Model:                            OLS   Adj. R-squared:                  0.728
Method:                 Least Squares   F-statistic:                     91.11
Date:                Wed, 12 Mar 2025   Prob (F-statistic):           3.08e-28
Time:                        16:58:45   Log-Likelihood:                -1621.7
No. Observations:                 102   AIC:                             3251.
Df Residuals:                      98   BIC:                             3262.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.062e+06   1.83e+06     -0.579      0.564    -4.7e+06    2.58e+06
square      1.409e+05   9532.921     14.782      0.000    1.22e+05     1.6e+05
floor       2.352e+05   4.12e+04      5.706      0.000    1.53e+05    3.17e+05
log_distance -6.378e+04   2.74e+05     -0.233      0.816   -6.07e+05     4.8e+05
==============================================================================
Omnibus:                       11.040   Durbin-Watson:                   2.231
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               11.951
Skew:                           0.660   Prob(JB):                      0.00254
Kurtosis:                       4.034   Cond. No.                         533.
==============================================================================
```

Рисунок 14 Данные регрессионной модели 4

Интерпретация Model 4:

Описание модели:

- Зависимая переменная: цена (price).
- Независимые переменные: площадь (square), этаж (floor), логарифм расстояния (log_distance).

Результаты:

- R-squared: 0.736. Указывает на то, что 73.6% вариации в цене объясняется этой моделью.
- Adj. R-squared: 0.728. Ускорректированная оценка с учетом числа переменных.

Коэффициенты:

- const: -1.062e+06. Несущественное значение (p-value = 0.564).
- square: 1.409e+05. Значимый положительный эффект площади на цену. Увеличение на квадратный метр повышает цену на 140,900.
- floor: 2.352e+05. Значимый эффект этажа на цену. Дополнительный этаж увеличивает цену на 235,200.
- log_distance: -6.378e+04. Не оказывает значимого влияния на цену (p-value = 0.816).

P-значения: площадь и этаж – значимые переменные.

Диагностические тесты:

- Omnibus и Jarque-Bera: некоторое отклонение от нормальности распределения остатков.
- Durbin-Watson (DW): 2.231. Указывает на отсутствие автокорреляции.

Вывод: эта модель хорошо объясняет изменения цены через площадь и этаж. Логарифм расстояния не оказывает значимого влияния. Модель будет полезна для прогнозирования цены на основе значимых факторов.

7.5 Model 5

```
Результаты для Model 5:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.009
Method:                 Least Squares   F-statistic:                    0.05453
Date:                Wed, 12 Mar 2025   Prob (F-statistic):              0.816
Time:                        16:58:45   Log-Likelihood:                -1689.6
No. Observations:                 102   AIC:                             3383.
Df Residuals:                     100   BIC:                             3388.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.64e+06   3.42e+06      1.943      0.055    -1.4e+05    1.34e+07
log_distance  1.231e+05   5.27e+05      0.234      0.816   -9.23e+05    1.17e+06
==============================================================================
Omnibus:                       36.610   Durbin-Watson:                   1.916
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.560
Skew:                           1.492   Prob(JB):                     1.30e-15
Kurtosis:                       5.687   Cond. No.                         59.9
==============================================================================
```

Рисунок 15 Данные регрессионной модели 5

Интерпретация Model 5:

Описание модели:

- Зависимая переменная: цена (price).
- Независимая переменная: логарифм расстояния (log_distance).

Результаты:

- R-squared: 0.001. Модель объясняет лишь 0.1% вариации в цене, практически без объяснительной силы.
- Adj. R-squared: -0.009. Указывает на то, что модель хуже, чем простое усреднение.

Коэффициенты:

- const: 6.64e+06. Высокое значение, но незначимо (p-value = 0.055).
- log_distance: 1.231e+05. Не оказывает значимого эффекта на цену (p-value = 0.816).

P-значения: ни один коэффициент не является статистически значимым.

Диагностические тесты:

- Omnibus и Jarque-Bera: значительные отклонения от нормальности распределения остатков.
- Durbin-Watson (DW): 1.916. Возможная автокорреляция остатков.

Вывод: логарифм расстояния не влияет на цену, и модель не объясняет изменений цены. Это неудивительно, так как R-squared показывает практически полное отсутствие объяснительной способности.

8. Отбор наилучшей регрессионной модели

8.1 Наилучшая модель

В результате анализа регрессионных моделей Model 1 была выбрана по следующим причинам:

- Высокое значение скорректированного R^2: Модель объясняет 73.1% вариации в цене, демонстрируя хороший баланс между сложностью модели и её объясняющей способностью.
- Статистическая значимость: Площадь и этаж показали низкие p-значения, свидетельствуя о значимом влиянии на стоимость квартиры.
- Интерпретируемость: Простота модели позволяет легко объяснить результаты, что важно для принятия решений на рынке недвижимости.

8.2 Интерпретация коэффициентов

Площадь (square): коэффициент в размере 140,800 указывает, что каждый дополнительный квадратный метр увеличивает стоимость квартиры в среднем на 140,800 единиц. Это логично, так как большая площадь часто ассоциируется с более высокой ценой на рынке.

Этаж (floor): коэффициент в размере 235,000 означает, что увеличение этажа на один связано с ростом стоимости квартиры на 235,000 единиц. Это может отражать предпочтения покупателей, такие как лучший вид или меньшее воздействие уличного шума.

8.3 Визуализация

Прогнозные vs. Фактические значения

Модель в целом демонстрирует хорошую предсказательную способность, хотя некоторые ошибки могут присутствовать.

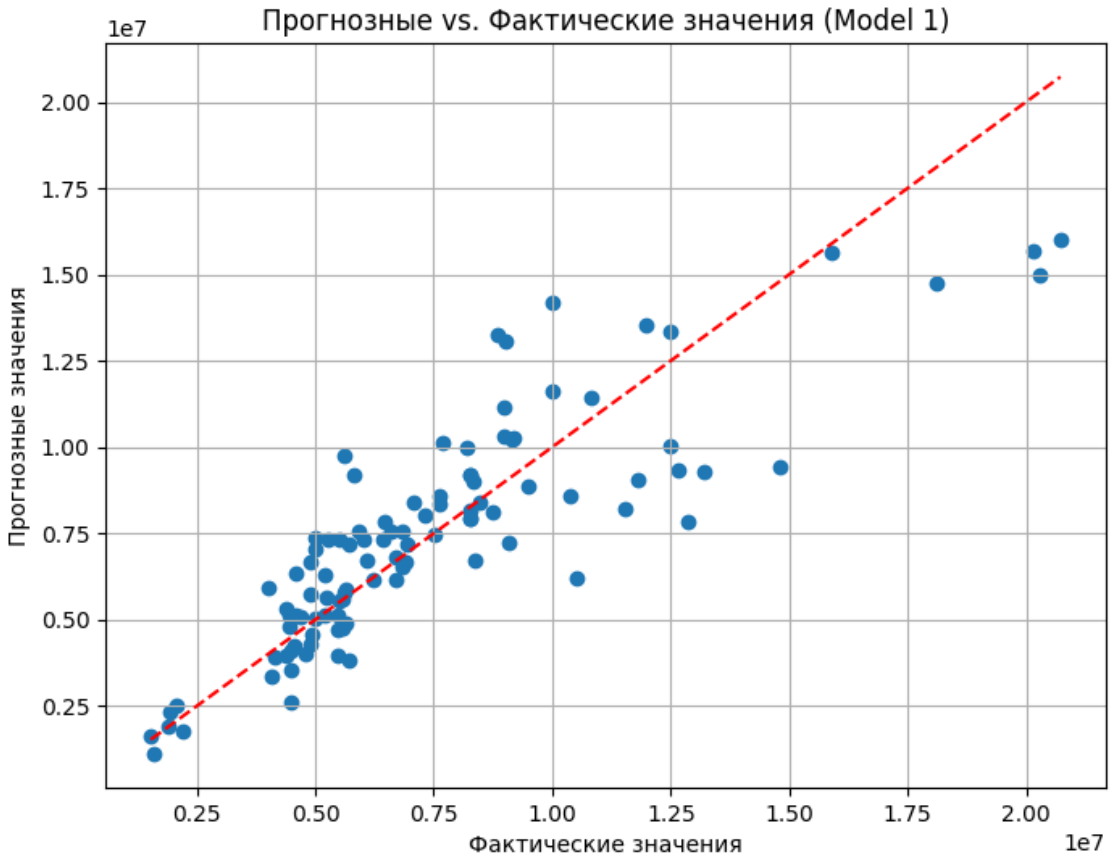

Рисунок 16 Model 1 Прогнозные vs. Фактические значения

Гистограмма ошибок.

Показала приближение к нормальному распределению, подтверждая качество модели.

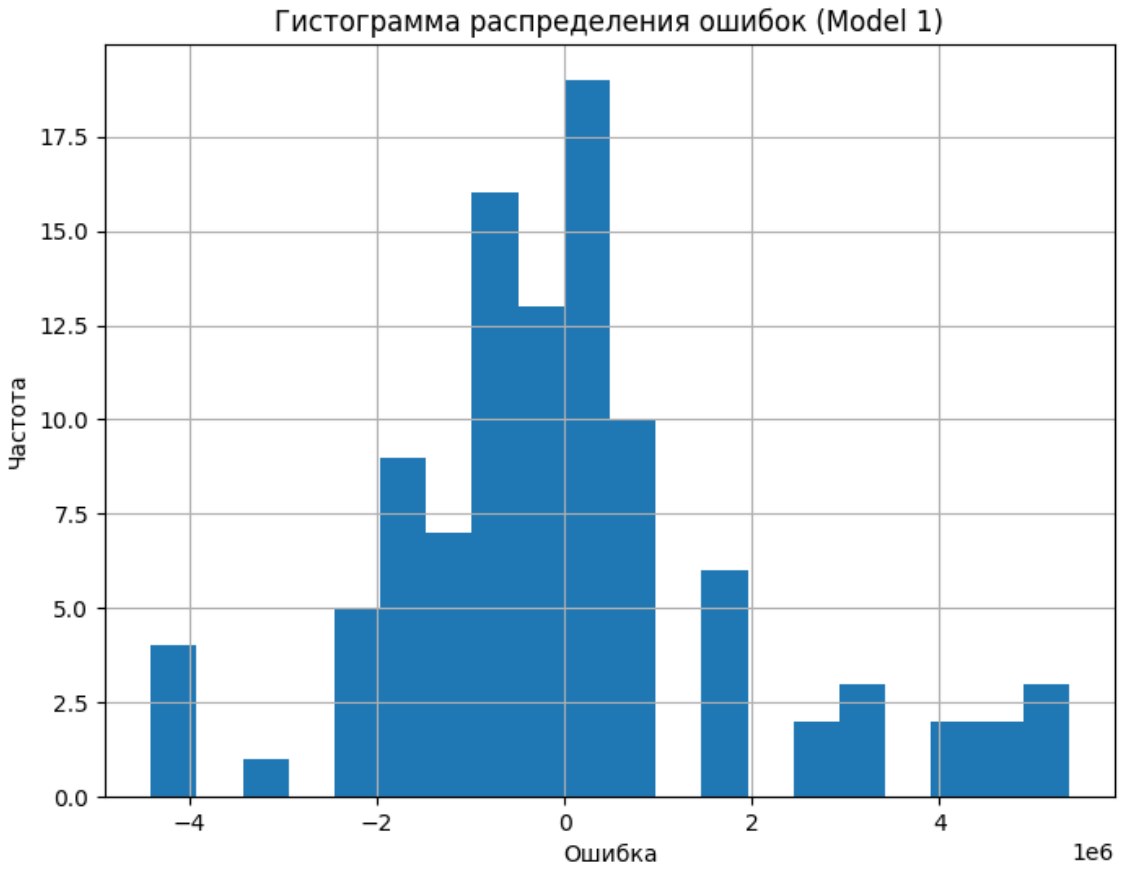

Рисунок 17 Model 1 Гистограмма ошибок

Вывод: Model 1 обеспечивает надёжные и интерпретируемые результаты, делая её оптимальным выбором для нашего анализа цены на недвижимость.

9. Применение результатов научного исследования в практической деятельности

9.1 Применение и назначение

Разработанная концепция программного продукта, основанная на регрессионной модели для оценки стоимости квартир, открывает широкие возможности для применения в различных областях:

9.1.1 Риелторская деятельность

Программный продукт позволит риелторам быть более конкурентоспособными за счет точной и оперативной оценки стоимости недвижимости. Это ускорит процесс подготовки предложений для клиентов, предоставляя персонализированные рекомендации, основанные на текущих рыночных условиях.

9.1.2 Инвестирование в недвижимость

Инвесторы смогут использовать продукт для оценки доходности и рисков потенциальных инвестиций. Возможность сравнительного анализа объектов позволит инвесторам принимать взвешенные решения, минимизируя финансовые риски.

9.1.3 Повышение прозрачности рынка

За счет интеграции с актуальными базами данных продукт будет способствовать повышению прозрачности рынка недвижимости, что устраняет информационное неравенство между участниками рынка и содействует честной конкуренции.

9.1.4 Улучшение клиентского опыта

Покупатели получат доступ к объективной информации о стоимости жилья, что упростит процесс выбора и повысит доверие к сделкам. Пользователи смогут самостоятельно осуществлять начальную оценку и анализ предложений без необходимости постоянного обращения к специалистам.

9.1.5 Образовательные и исследовательские цели

Продукт может быть использован в образовательных учреждениях и исследовательских центрах для изучения рыночных механизмов и разработки стратегий функционирования рынка недвижимости.

9.2 Основные функциональные возможности

Оценка стоимости: пользователь собирает Dataset параметров жилья в выбранном городе, такие как площадь и этаж. Модель регрессии определяет оценочную стоимость, основываясь на статистическом анализе.

Визуализация и аналитика: продукт предоставляет графическую визуализацию данных, показывающую влияние различных параметров на

стоимость недвижимости. Это включает диаграммы прогнозов и отклонений, что помогает пользователю выявлять ключевые факторы, влияющие на рынок.

Сравнительный анализ: программа позволяет пользователям сравнивать различные объекты недвижимости, предлагая рекомендации по выбору с учетом всех важных характеристик.

9.3 Дальнейшее развитие

Интеграция с рыночными базами данных: автоматическое обновление данных о ценах и объектах для более актуальной информации и анализа трендов.

Учет дополнительных параметров: введение таких факторов, как, качество инфраструктуры и управляющая компания, для расширения модели оценки.

Разработка мобильной версии: создание web приложения на React позволяющее пользователям использовать продукт в любое время и на любой платформе, повышая их мобильность и эффективность работы.

ЗАКЛЮЧЕНИЕ

Программный продукт разработан для профессионалов в сфере недвижимости, включая риелторов, инвесторов и потенциальных покупателей. Цель продукта — обеспечить точную и быструю оценку стоимости квартир с использованием передовой регрессионной модели. Это позволит пользователям эффективно анализировать рынок и принимать обоснованные решения, что увеличит их конкурентоспособность и снизит риски, связанные с непредвиденными изменениями цен.

Данный программный продукт может стать незаменимым инструментом в секторе недвижимости, предлагая пользователям аналитические возможности на высшем уровне для успешного участия на конкурентном рынке.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Коваленко, Л. П. Введение в машинное обучение [Электронный ресурс] / Л. П. Коваленко. – Электрон. текстовые данные. – Казань: Издательство КФУ, 2022. – 180 с. – Режим доступа: https://litres.com/book/elena-kapaca/mashinnoe-obuchenie-dostupnym-yazykom-69273385/read/
2. Николаев, А. Г. Программирование на Python для анализа данных [Электронный ресурс] / А. Г. Николаев. – Электрон. текстовые данные. – Владивосток: ДВФУ, 2020. – 320 с. – Режим доступа: https://bhv.ru/product/python-eto-prosto-poshagovoe-rukovodstvo-po-programmirovaniyu-i-analizu-dannyh/
3. Лебедев, Ю. В. Основы эконометрики [Электронный ресурс] / Ю. В. Лебедев. – Электрон. текстовые данные. – Самара: Самарский университет, 2019. – 110 с. – Режим доступа: https://elib.osu.ru/bitstream/123456789/13226/1/%D0%9B%D0%B5%D0%B1%D0%B5%D0%B4%D0%B5%D0%B2%D0%B01.pdf
4. Фомин, В. С. Методы многомерного анализа [Электронный ресурс] / В. С. Фомин. – Электрон. текстовые данные. – Нижний Новгород: Университет Лобачевского, 2018. – 95 с. – Режим доступа: https://www.iae.nsk.su/images/stories/5_Autometria/5_Archives/1990/2/89-91.pdf
5. Сорокина, Т. В. Статистические методы в экономике [Электронный ресурс] / Т. В. Сорокина. – Электрон. текстовые данные. – Уфа: БашГУ, 2021. – 130 с. – Режим доступа: https://stm.hse.ru/mirror/pubs/share/direct/865070119.pdf

ПРИЛОЖЕНИЯ

Для доступа к полному коду проекта практики и дополнительной документации вы можете посетить репозиторий на GitHub:

Ссылка на проект: GitHub – https://github.com/AniCatPro/NIR