МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

МОДЕЛИРОВАНИЕ СТОИМОСТИ КОНТРАКТОВ НА АРЕНДУ ЖИЛЬЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Заярного Артема Сергеевича

Научный руководитель
старший преподаватель _____ А. Д. Луньков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2024

Введение

Исследование вопросов, связанных с определением стоимости аренды жилья, имеет большое значение. Аренда жилья является важной частью жизни многих людей. Понимание факторов, которые влияют на стоимость аренды, позволяет более эффективно управлять рынком аренды. Это помогает прогнозировать изменения и разрабатывать подходящие стратегии для всех заинтересованных сторон.

Регрессионные модели являются одним из наиболее эффективных инструментов для анализа и прогнозирования стоимости аренды. Они позволяют выявить зависимости между различными факторами, такими как расположение жилья и его характеристики. Понимание этих факторов помогает как арендаторам, так и арендодателям принимать обоснованные решения. Это может быть полезно при выборе жилья или определении оптимальной цены аренды.

Целью бакалаврской работы является регрессионное моделирование стоимости контрактов на аренду жилья в городе Саратов.

Для достижения данной цели были поставлены следующие **задачи**:

- Изучить классические регрессионные модели и модели с географическим взвешиванием;
- Автоматизировать получение выборки о предложениях на рынке арендуемого жилья;
- Проанализировать выборку, единицы наблюдения в которой — выставленные предложения по аренде;
- Оценить параметры регрессионных моделей на основе собранных данных;

Структура бакалаврской работы

Работа состоит из введения, трех разделов, заключения и приложения.

— В первом разделе рассматриваются основные аспекты регрессионного анализа, в частности, модели множественной регрессии и модели с географически взвешенными коэффициентами.

— Во втором разделе рассматривается полученная выборка, описывается подготовка данных для статистического анализа и проводится предварительный анализ данных.

— В третьем разделе проводится построение регрессионных моделей и представлены результаты моделирования стоимости контрактов на аренду жилья, полученные с помощью программных средств языка Python.

Основное содержание работы

Первый раздел. Основные сведения из регрессионного анализа.

В регрессионном анализе ключевыми понятиями являются:

1. Зависимая переменная — это параметр, который исследователь пытается объяснить или предсказать. В контексте ценообразования на рынке аренды жилья зависимой переменной может выступать цена аренды жилья.
2. Независимые переменные – это факторы, которые используются для объяснения изменений в зависимой переменной. Примерами независимых переменных могут служить местоположение жилья, его площадь и другие параметры, которые могут влиять на ценообразование.
3. Регрессионная модель — это математическое выражение, которое описывает связь между зависимой переменной и независимыми переменными. В простейшей форме регрессионная модель выглядит так.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon,$$

здесь Y - зависимая переменная, $X_1, X_2, \ldots, X_k$ - независимые переменные, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ - коэффициенты регрессии, и ε - случайная ошибка.

4. Коэффициенты регрессии — это числовые значения $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, которые показывают, насколько изменение в независимой переменной влияет на изменение зависимой переменной. Например, коэффициент β_1 показывает, на сколько единиц изменится зависимая переменная при изменении независимой переменной X_1 на одну единицу, при условии, что остальные переменные постоянны.
5. Ошибки представляют разницу между фактическими наблюдаемыми значениями зависимой переменной и значениями, которые предсказаны моделью. Эти ошибки ε считаются случайными и обладают определен-

ными вероятностными свойствами.

Линейная множественная регрессия

Эта глава полностью основана на концепциях и обсуждениях, представленных в [1].

В качестве первого эконометрического метода в данном исследовании используется модель множественной регрессии.

Многомерная регрессионная модель (multiple regression model), или *модель множественной регрессии* имеет вид:

$$y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, n,$$

или

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, n, \tag{1}$$

где x_{tp} - значения регрессора x_p в наблюдении t, а $x_{t1} = 1, \quad t = 1, \ldots, n$.
С учетом этого замечания не будем далее различать модели вида (1) со свободным членом или без свободного члена.

Общее назначение множественной регрессии состоит в анализе связи между несколькими независимыми переменными (регрессорами) и зависимой переменной.

Наиболее распространенным методом оценивания параметров линейных эконометрических моделей является метод наименьших квадратов. Целью метода является выбор вектора оценок $\hat{\beta}$, минимизирующего сумму квадратов остатков e_t(т.е. квадрат длины вектора остатков e):

$$e = y - \hat{y} - X\hat{\beta}, \quad ESS = \sum e_t^2 = e'e \to min.$$

Оценка метода наименьших квадратов является оптимальной.

Квантильная регрессия

В качестве второго эконометрического метода в данной работе используется квантильная регрессия. Модель квантильной регрессии используется для анализа влияния независимых переменных на различные квантили (процентили) зависимой переменной.

Квантиль — это некоторое числовое значение, которое делит набор дан-

ных на части заданных размеров. В контексте распределения данных, квантиль определяет значение, ниже которого находится определённый процент данных.

Главным отличием классической регрессии от квантильной является вопрос, которым задается исследователь. В классической регрессии главный вопрос: какие факторы связаны с изменением среднего значения Y? А в квантильной регрессии вопрос задан так: от чего зависит условный квантиль y_i? Частным случаем квантильной регрессии является медианная регрессия (квантиль 0,5).

Пусть $F(y|X)$ – функция условного распределения случайной величины Y при заданном векторе регрессоров X. Квантилем уровня $\tau(0 < \tau < 1)$ называют функцию [2]

$$Q_\tau(X) = F^{-1}(\tau|X) = inf\{y : F(y|X) \geq \tau\}. \tag{2}$$

Пусть условные квантили заданных значений переменной Y линейно зависят от вектора объясняющих переменных X, тогда

$$Q(\tau|X_i'\beta(\tau)) = X_i'\beta(\tau),$$

где $\beta(\tau)$ – вектор коэффициентов при X, соответствующих квантилю τ. Исходя из этого задача безусловной минимизации будет выглядеть следующим образом: [3]

$$\hat{\beta}_n(\tau) = arg\min_{\beta(\tau)} \left\{ \sum_i p_\tau(Y_i - X_i\prime\beta(\tau)) \right\}. \tag{3}$$

Модель квантильной регрессии полезна, когда интерес к анализу зависимости между переменными распространяется на различные квантили зависимой переменной, а не только на среднее значение. Это позволяет учесть неоднородность в данных и оценить влияние независимых переменных в разных сегментах распределения.

Модель с географически взвешенными коэффициентами

Географически взвешенная регрессия - это географический метод, который моделирует пространственно изменяющиеся зависимости. По сравнению

с классической регрессией, коэффициенты в географически взвешенной регрессии зависят от точки, в которой проводятся измерения. Точку измерения назовем местоположением.

Модель географически взвешенной регрессии имеет вид: [4]

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) \cdot x_{ik} + \varepsilon_i, \tag{4}$$

где (u_i, v_i) координаты точки i, $i = 1, \ldots, n$;
y_i - зависимая переменная в точке i;
$x_{i1}, \ldots, x_{ip}$ - независимые детерминированные регрессоры, $k = 1, \ldots, p$, p - число регрессоров;
$\beta_k(u_i, v_i)$ - неизвестные, подлежащие оценке, коэффициенты, $k = 0, \ldots, p$;
ε_i - ошибки измерений.

Оценки коэффициентов модели $\hat{\beta}_k(u_i, v_i), k = 0, \ldots, p$, вычисляются в каждом местоположении.

Для вычисления оценок коэффициентов в местоположении i применяют метод наименьших квадратов с географическим взвешиванием.

Предполагается, что регрессионные модели для соседних точек схожи, но могут варьироваться по территории. Степень близости объектов учитывается с помощью весов w_{ij} . Вектор оценок коэффициентов для каждого местоположения i вычисляется по формуле:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y, \tag{5}$$

где $W(u_i, v_i)$ - диагональная матрица весовых коэффициентов размерности $(n \times n)$.

$$W(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & w_{in} \end{bmatrix}. \tag{6}$$

Элемент матрицы $w_{ij}, (i, j = 1, \ldots, n)$ определяет степень влияния процессов, протекающих в местоположении с номером j, на зависимую переменную в местоположении с номером i. Под процессами понимаем изменения

значений регрессоров и зависимой переменной.

Наиболее употребляемые методы вычисления весовых коэффициентов:

1. Административно-территориальное деление:

Элемент весовой матрицы принимают равным единице для точек, принадлежащих району A с местоположением i и равным нулю в противном случае:

$$\begin{aligned} w_{ij} &= 1, \quad \text{если } (i,j) \in A; \\ w_{ij} &= 0, \quad \text{если } (i,j) \notin A. \end{aligned} \tag{7}$$

2. Метод движущегося окна:

В данном случае вес принимают равным единице, если расстояние d_{ij} между объектами i и j не превосходит заданного расстояния b, и равным нулю в противном случае:

$$\begin{aligned} w_{ij} &= 1, \quad \text{если } d_{ij} < b; \\ w_{ij} &= 0, \quad \text{если } d_{ij} \geq b. \end{aligned} \tag{8}$$

Величина b фиксирована и называется шириной окна или шириной полосы пропускания.

3. Фиксированные ядра:

Подход, в котором веса строятся с учетом непрерывного изменения расстояния между исследуемыми объектами, называют ядерным. Веса являются убывающими функциями расстояния, и называются ядрами. Наиболее часто применяют ядра Гаусса:

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]. \tag{9}$$

В местоположении i вес равен единице, а при удалении объектов от него быстро уменьшается согласно функции Гаусса.

4. Адаптивные ядра:

Существует несколько способов создания таких ядер. Часто веса рассчитывают с учетом рангов. Ближайшим соседям присваивают нулевой ранг и вес равный единице. При удалении объектов от местоположения ранг, как

и расстояние, увеличивается, а вес уменьшается:

$$w_{ij} = \begin{cases} exp(-\frac{d_{ij}}{R_m}) & \text{, если } d_{ij} < R_m, \\ 0 & \text{, если } d_{ij} > R_m, \end{cases} \tag{10}$$

где R_m - расстояние до m-го ближайшего соседа.

Если ширину полосы пропускания определить как расстояние до m-го соседа, то получим ядро с изменяющейся шириной полосы пропускания. В таком случае полоса автоматически меняется в зависимости от скученности точек измерения. В более густых местах - сужается, а в более разряженных - увеличивается.

Второй раздел. Сбор и анализ данных

Сбор данных

Данные по объявлениям города Саратова и Энгельса, размещенным на сервисе объявлений об аренде и продаже недвижимости Яндекс.Недвижимость, были получены с официального сайта онлайн-платформы. Для получения данных была написана программа на языке программирования Python с использованием библиотеки Requests и модуля JSON.

Всего было получено 546 объявлений об аренде жилья, которые были размещены на сервисе с февраля по май 2024 года.

Подготовка данных

В ходе подготовки данных были удалены параметры объявлений, которые содержали большое количество пустых значений, были созданы новые параметры, также были удалены объявления, которые содержали пустые значения в столбце builtYear.

Было проведено целочисленное кодирование значений в столбцах renovation и bathroomUnit.

В ходе рассмотрения данных были обнаружены выбросы в столбцах builtYear, price и area. Для их выявления были построены коробчатые диаграммы, которые позволили наглядно представить распределение данных и выделить аномальные значения рисунок 1.

Чтобы избавиться от выбросов, были удалены объявления, у которых

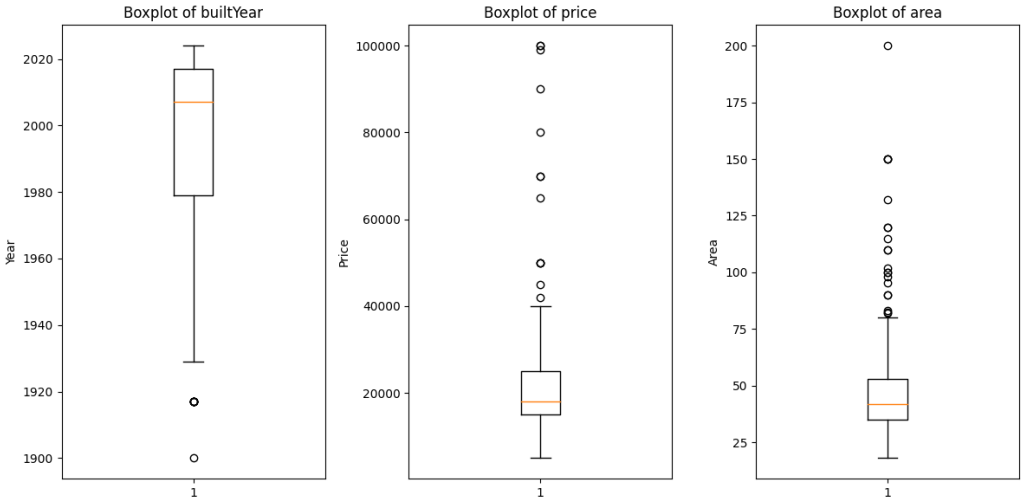

Рисунок 1 – коробчатые диаграммы builtYear, price и area

цена (price) выше 40 000, площадь (area) больше 70 квадратных метров или год постройки (builtYear) ранее 1930 года.

После подготовки данных был получен CSV-файл prepared_offers, содержащий информацию о 458 объявлениях. Параметры после подготовки представлены в таблице 5 приложения А.

Предварительный анализ данных

Предварительный анализ данных был проведён для понимания общих тенденций и особенностей собранных данных по объявлениям об аренде жилья в городе Саратове. Этот этап анализа включает в себя описание распределений параметров, анализ корреляций между параметрами, а также выявление возможных особенностей данных.

Статистические данные о характеристиках объявлений об аренде жилья в Саратове представлены в таблице 6 приложения А.

Данная таблица показывает, что большинство квартир были построены относительно недавно, что подтверждается медианным годом постройки 2007. Цены варьируются от 5000 до 38000 рублей, при этом средняя цена находится около 18720 рублей, а медиана составляет 17000 рублей. Площадь квартир варьируется от 18 до 68 квадратных метров, средняя площадь составляет 42.54 квадратных метра, а медиана — 40 квадратных метров.

Среднее значение для типа санузла составляет 0.31, что указывает на преобладание квартир со совмещёнными санузлами. Среднее значение показателя ремонта составляет 1.31, что означает, что большинство квартир

имеют косметический ремонт (1). Количество комнат в среднем составляет 1.37, с разбросом от 0 (студии) до 3 комнат. Средняя высота потолков — 2.595 метра, а средний номер этажа, на котором находятся квартиры — 5.82, при этом общее количество этажей в домах в среднем составляет 10.73. Лифт имеется в 83.4% домов, охрана — в 49.2% домов. В 73.1% квартир есть мебель, а дополнительные удобства присутствуют в 68.05% случаев. На первых этажах расположено 9.2% квартир, на последних этажах — 5.9%. Большинство зданий построены из кирпича (57.5%), 13.3% зданий монолитные и 29.1% панельные.

Для выявления взаимосвязей между различными параметрами были рассчитаны коэффициенты корреляции. Основные наблюдения:

1. Год постройки умеренно положительно коррелирует с высотой потолков, этажностью и общим количеством этажей в доме. Умеренно отрицательно коррелирует с количеством комнат и типом здания, построенного из кирпича.
2. Цена сильно положительно коррелирует с площадью квартиры и видом ремонта. Умеренно положительно коррелирует с общим количеством этажей, количеством комнат, и наличием лифта. Сильная отрицательная корреляция с дистанцией до центра. Отрицательная корреляция с типом здания, построенного по панельной технологии.
3. Площадь сильно положительно коррелирует с количеством комнат.
4. Ремонт умеренно положительно коррелирует с высотой потолков, этажностью, общим количеством этажей, наличием лифта , и наличием мебели.

Была построена карта концентрации объявлений по районам города Саратова с использованием Plotly и GeoPandas.

На рисунке 2 видно, что самое большое количество объявлений содержится в Заводском районе Саратова (98 объявлений). Наименьшее же количество объектов расположено во Фрунзенском районе (30 объявлений).

Также были созданы карты, на которых можно увидеть среднюю стоимость квартир и средний возраст зданий в разных районах рисунок 3 и 4.

На этих картах видно, что во Фрунзенском и Октябрьском районах средние цены на жильё самые высокие — 21 500 и 23 900 рублей соответ-

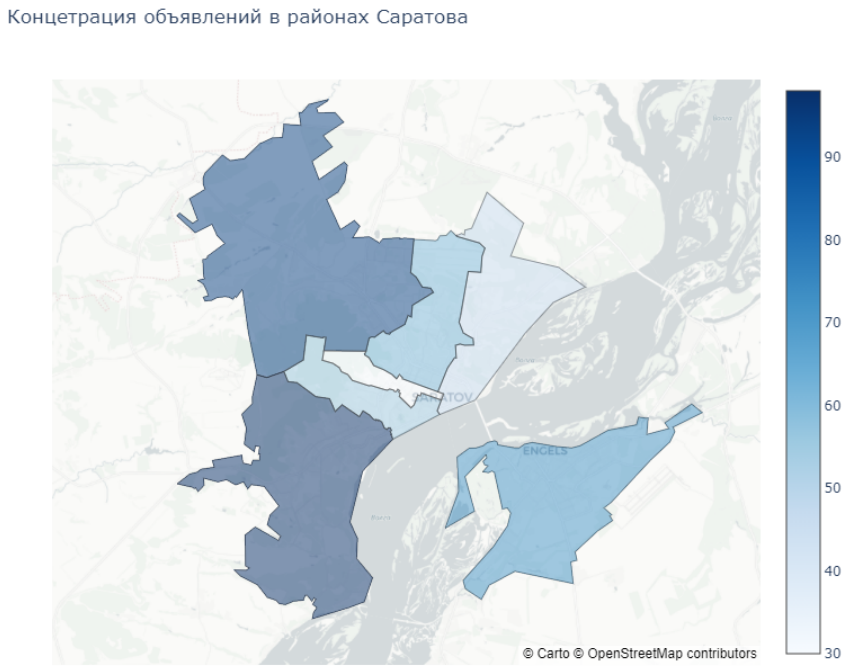

Рисунок 2 – Концентрация объявлений по районам

ственно. Однако в этих же районах средний год постройки самый низкий — 1985 и 1989 годы. В Энгельсе средний год постройки самый высокий, но средняя цена — 17 000 рублей, что является средней ценой по Саратову в целом.

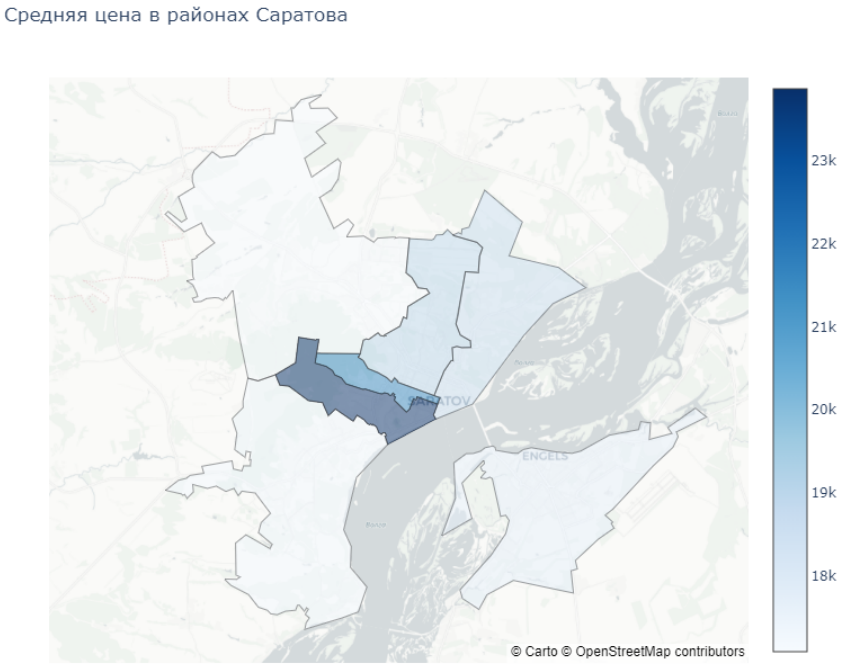

Рисунок 3 – Средняя цена по районам

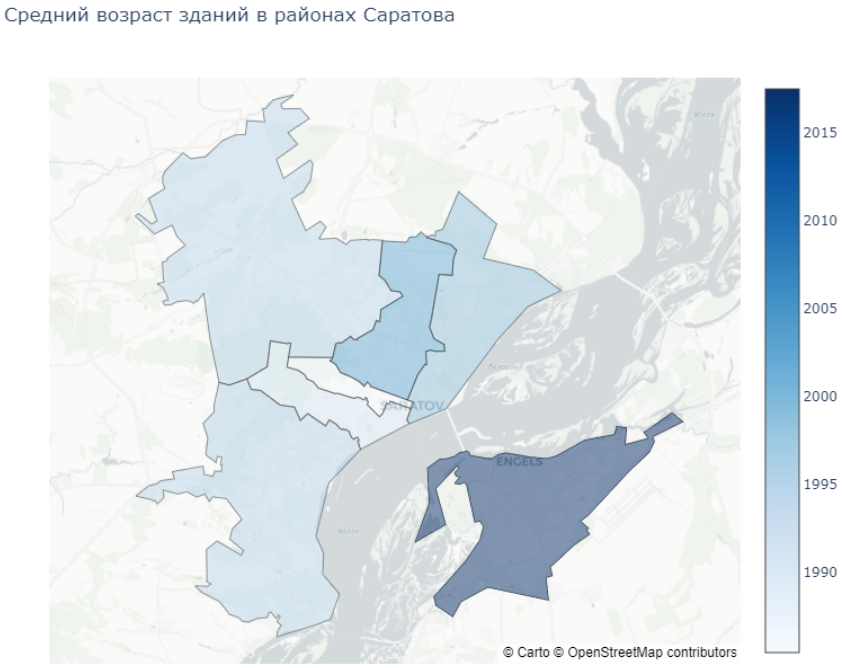

Рисунок 4 – Средний год постройки по районам

Отсюда можно сделать следующие выводы:

— Высокие цены на жильё в районах с более старым фондом недвижимости (Фрунзенский и Октябрьский) могут указывать на значимость других факторов, таких как престиж района, наличие инфраструктуры, удобства и транспортная доступность.

— Более новые здания в Энгельсе не приводят к значительному увеличению средней стоимости жилья, что может свидетельствовать о различиях в восприятии ценности новизны жилья между районами или о наличии других факторов, сдерживающих рост цен.

Третий раздел.Практическое использование регрессионного анализа для предсказания стоимости аренды жилья

Моделирование стоимости контрактов на аренду жилья с помощью множественной регрессии

Для построения регрессионных моделей использовались библиотеки *statsmode* и *matplotlib* языка *Python*.

Для построения модели множественной регрессии были выбраны следующие признаки: area, renovation, total_floor, improvements, buildingType_PANEL, distance_to_center. Признаки отбирались путем использования рекурсивного отбора признаков.

После построения модели множественной регрессии на основе подготовленных данных был получен результат, который представлен в таблице 1.

OLS Regression Results						
Dep. Variable:	price		R-squared: 0.645		Adj. R-squared: 0.639	
Model:	OLS		F-statistic: 108.5		Prob (F-statistic): 1.93e-77	
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2023	0.006	342.619	0.000	2.190	2.215
area	0.0015	0.000	13.210	0.000	0.001	0.002
renovation	0.0128	0.002	5.838	0.000	0.008	0.017
total$_floor$	0.0017	0.000	7.751	0.000	0.001	0.002
improvements	0.0076	0.002	3.069	0.002	0.003	0.012
buildingType_PANEL	-0.0100	0.003	-3.852	0.000	-0.015	-0.005
distance_to_center	-0.0036	0.000	-9.783	0.000	-0.004	-0.003
Omnibus:	4.453		Durbin-Watson:	2.103		
Prob(Omnibus):	0.108		Jarque-Bera (JB):	4.851		
Skew:	-0.144		Prob(JB):	0.0884		
Kurtosis:	3.486		Cond. No.:	276		

— MSE: 0.0005790135615343962

— Breusch-Pagan test results:

– F-statistic: 10.357518431417184

– P-value: 0.11038126221583296

Таблица 1 – Результаты модели множественной регрессии

Коэффициент детерминации R-squared равен 0.645, что говорит о том, что модель объясняет 64.5% дисперсии зависимой переменной. Но значительная часть вариаций цен остаётся необъяснённой.

F-statistic = 108.5, p-value = 1.93e-77, что указывает на значимость модели в целом.

Площадь, ремонт, количество этажей, улучшения значимо и положительно влияют на цену. Тип здания «панельный» и расстояние до центра значимо снижают цену.

Durbin-Watson statistic: 2.067. Значение около 2 указывает на отсутствие автокорреляции остатков.

Omnibus test: p-value = 0.814, Jarque-Bera test: p-value = 0.773. Высокие значения p-value указывают на то, что распределение остатков не отклоняется от нормального.

Breusch-Pagan test: F-statistic = 13.184, p-value = 0.067. P-value близко к значимости, но не достигает 0.05. Это указывает на слабые признаки

гетероскедастичности.

Низкое значение MSE указывает на хорошее качество предсказаний модели.

Модель множественной регрессии также была построена на наборе данных, который содержал только однокомнатные квартиры. Данный набор данных содержал 277 объявлений. Результаты модели представлены в таблице 2.

OLS Regression Results						
Dep. Variable:	price		R-squared: 0.697		Adj. R-squared: 0.690	
Model:	OLS		F-statistic: 98.93		Prob (F-statistic): 8.90e-54	
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2018	0.010	231.509	0.000	2.183	2.221
area	0.0014	0.000	5.225	0.000	0.001	0.002
renovation	0.0147	0.003	4.755	0.000	0.009	0.021
total_floor	0.0020	0.000	6.709	0.000	0.001	0.003
improvements	0.0084	0.003	2.766	0.006	0.002	0.014
distance_to_center	-0.0051	0.000	-11.021	0.000	-0.006	-0.004
Omnibus:	15.599		Durbin-Watson:	2.129		
Prob(0.000):	0.626		Jarque-Bera (JB):	18.683		
Skew:	-0.540		Prob(JB):	8.77e-05		
Kurtosis:	3.929		Cond. No.:	280		

— Среднеквадратичная ошибка (MSE): 0.0003648

— Breusch-Pagan test results:

– **F-statistic:** 6.348602523717644

– **P-value:** 0.2737605062250495

Таблица 2 – Результаты модели множественной регрессии для однокомнатных квартир

R-squared: 0.697 — это значение указывает на то, что примерно 69.7% вариации в цене может быть объяснено данной моделью.Высокий F-статистический показатель и практически нулевая вероятность указывают на значимость модели в целом. Модель хорошо объясняет вариацию цен на аренду однокомнатных квартир, что подтверждается высокими значениями R-squared и Adj. R-squared. Большинство коэффициентов значимы и имеют ожидаемые знаки, что подтверждает надежность модели для предсказания цен на аренду.

Так же были построены модели множественной регрессии для разных районов. Результаты построения моделей множественной регрессии для За-

водского (98 объявлений) и Ленинского (94 объявления) районов приведены в таблицах 3 и 4.

OLS Regression Results						
Dep. Variable:	price		R-squared: 0.507			Adj. R-squared: 0.478
Model:	OLS		F-statistic: 17.96			Prob (F-statistic): 3.44e-10
	coef	std err	t	P>|t|	[0.025	0.975]
const	8.7752	0.122	71.690	0.000	8.531	9.019
renovation	0.3058	0.064	4.810	0.000	0.179	0.433
rooms	0.2265	0.043	5.247	0.000	0.140	0.313
floor	-0.0016	0.011	-0.147	0.884	-0.024	0.020
total_floor	0.0219	0.011	2.038	0.045	0.000	0.043

MSE на тестовой выборке: 0.07561286877622918

Breusch-Pagan test results on test data: F-statistic: 9.133358860441884, P-value: 0.05785122173128896

Таблица 3 – Результаты модели множественной регрессии Ленинского района

OLS Regression Results						
Dep. Variable:	price		R-squared: 0.617			Adj. R-squared: 0.596
Model:	OLS		F-statistic: 29.39			Prob (F-statistic): 1.45e-14
	coef	std err	t	P>|t|	[0.025	0.975]
const	8.6005	0.111	77.325	0.000	8.379	8.822
renovation	0.3451	0.055	6.269	0.000	0.235	0.455
rooms	0.2601	0.036	7.247	0.000	0.189	0.332
floor	-0.0306	0.010	-2.947	0.004	-0.051	-0.010
total_floor	0.0466	0.010	4.555	0.000	0.026	0.067

MSE на тестовой выборке: 0.06949932998346432

Breusch-Pagan test results on test data: F-statistic: 8.578326642828259, P-value: 0.07254826239719524

Таблица 4 – Результаты модели множественной регрессии Заводского района

Обе модели были построены с одинаковым набором регрессоров. Модель Заводского района имеет более высокие значения R-squared и Adj. R-squared, что говорит о лучшей подгонке данных. Коэффициенты модели Заводского района также показывают более значимые и сильные связи между зависимой переменной (ценой) и независимыми переменными (особенно это касается переменных, связанных с реновацией и количеством комнат).

На основе проведённого анализа результатов множественной регрессии для разных районов становится очевидно, что для повышения точности и адекватности моделей ценообразования на недвижимость важно учитывать

региональные особенности. Необходимо тщательно подбирать набор регрессоров для каждого конкретного района, учитывая его уникальные характеристики и тренды. Это может включать в себя не только пересмотр существующих переменных, но и введение новых, более релевантных для конкретного контекста.

Моделирование стоимости контрактов на аренду жилья с помощью квантильной регрессии

Квантильная регрессия была применена к данным для различных квантилей (0.25, 0.5 и 0.75) с целью анализа влияния различных факторов на цену недвижимости. Были рассмотрены коэффициенты и их значимость на каждом квантиле. Результаты модели квантильной регрессии представлены в таблице 7, таблице 8 и таблице 9 приложения А. В результате анализа полученных результатов таблиц были сделаны следующие выводы:

1. Существенность факторов: факторы, такие как площадь, ремонт, общее количество этажей и улучшения, оказывают стабильное и значимое влияние на цену на всех уровнях квантилей;
2. Различия между квантилями: влияние ремонта становится более сильным по мере увеличения квантиля.
3. Качество моделей: MSE указывает на лучшее качество модели для медианного квантиля (0.5);

Эти результаты подчеркивают важность учета разных уровней цен при анализе влияющих факторов, так как значимость и величина эффектов могут существенно различаться в зависимости от ценового сегмента.

Моделирование стоимости контрактов на аренду жилья с помощью регрессии с географически взвешенными коэффициентами

Модель географически взвешенной регрессии была построена с использованием библиотек Geopandas, PySAL, PyShp.

В данной модели используется гауссово ядро, которое является одним из наиболее часто используемых пространственных ядер в географически взвешенной регрессии. Ширина окна определяет гладкость пространственной адаптации модели. В данной модели ширина окна выбирается автоматически с использованием метода Sel_BW, который подбирает оптимальное значение ширины окна, минимизируя критерий Акаике (AIC).

Критерий информации Акаике (AIC) — это мера относительного качества статистических моделей для набора данных.

Формула для расчета AIC:

$$\mathrm{AIC} = 2k - 2\ln(L),$$

где k — количество параметров в модели, а L — значение функции правдоподобия модели.

Результат использования метода Sel_BW — число bw, равное 70. Так как использовался метод для адаптивных ядер, bw — это число ближайших соседей. Найденное оптимальное число ближайших соседей было использовано как параметр в методе GWR, который использовался для построения модели.

Регрессорами были взяты параметры объявлений area, renovation, total_floor, improvements, buildingType_PANEL, distance_to_center.

Результаты географически взвешенного метода представлены в таблице 10 приложения А.

Коэффициент детерминации равен 0.691 . Таким образом, объясняется 69% дисперсии зависимой переменной.

Для удобства анализа данных были построены карты зависимостей коэффициентов регрессии при различных переменных от координат.

На рисунке 5 представлена зависимость коэффициента регрессии при площади от координат.

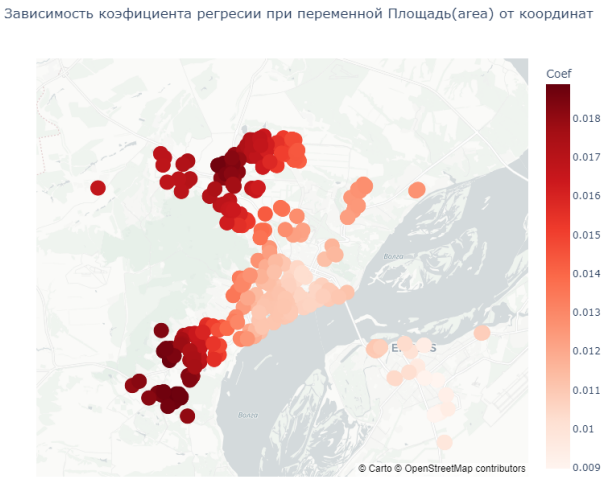

Рисунок 5 – Зависимость коэффициента для площади

На рисунке наглядно показано, в каких районах города площадь квартиры ценится выше. В Заводском и Ленинском районах заметен высокий коэффициент, в то время как в Энгельсе он минимальный. Можно заметить тенденцию: чем ближе к юго-востоку, тем меньше ценится площадь квартиры. А чем ближе к северо-западу, тем больше.

На рисунке 6 представлена зависимость коэффициента регрессии при ремонте от координат. На нем видно, что ремонт в наибольшей степени ценится в Заводском районе Саратова, в центральных районах Саратова, а именно в Октябрьском и Фрунзенском районах, коэффициент близок к минимальному значению.

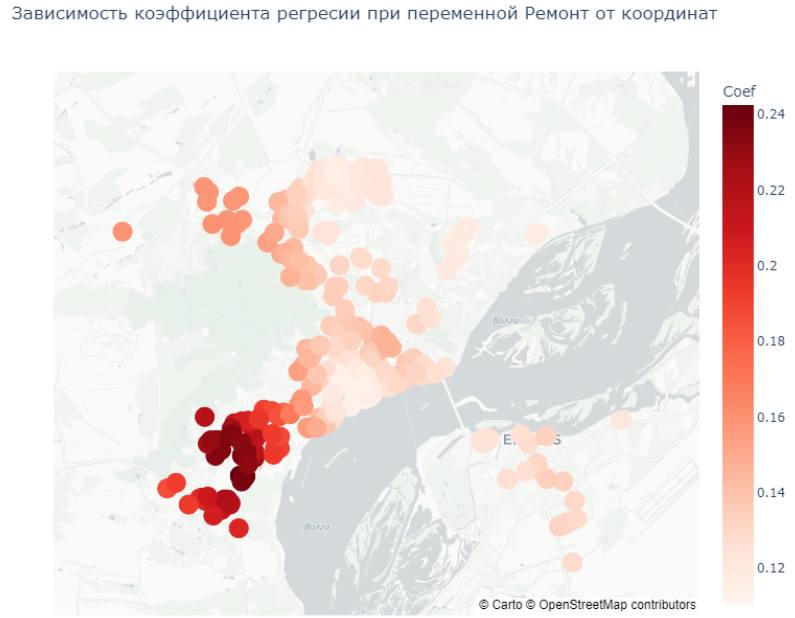

Рисунок 6 – Зависимость коэффициента для ремонта

На рисунке 7 представлена зависимость коэффициента регрессии при удобствах от координат. Коэффициент при этой переменной в заводском районе значительно ниже, чем в других районах. Наибольшее значение коэффициента для этой переменной зафиксировано в Ленинском районе.

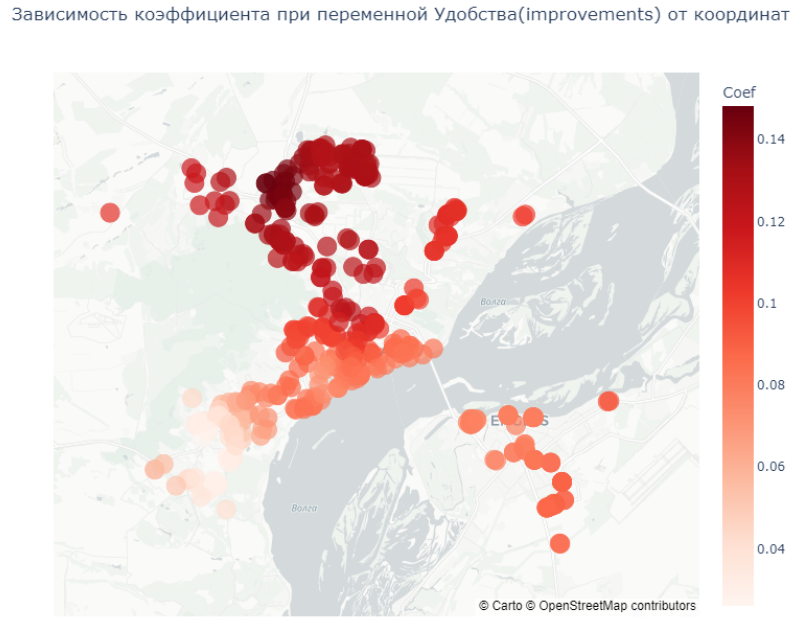

Рисунок 7 – Зависимость коэффициента для удобств

Исходя из вышесказанного, можно сделать следующий вывод: географически взвешенная регрессия позволяет определить локальные изменения коэффициентов, что обеспечивает более детальное понимание пространственных взаимосвязей.

Сравнение моделей

Линейная регрессия

1. **Простота:** Простая и понятная модель.
2. **Глобальные параметры:** Параметры оцениваются для всей выборки.
3. **Качество модели:** $R^2 = 0.645$, MSE $= 0.000579$.

Квантильная регрессия

1. **Гибкость:** Позволяет анализировать с разных сторон распределение зависимой переменной.

2. **Существенность факторов:** Факторы, такие как площадь и ремонт, оказывают значительное влияние на цену на всех уровнях квантилей.
3. **Качество моделей:**
   — 0.25 квантиль: MSE = 0.0702
   — 0.5 квантиль: MSE = 0.0593
   — 0.75 квантиль: MSE = 0.1064

Географически взвешенная регрессия

1. **Учет пространственной неоднородности:** Позволяет выявить локальные вариации.
2. **Результаты глобальной регрессии:**
   — $R^2 = 0.646$, Adj.$R^2 = 0.641$, RSS = 19.015.
3. **Результаты географически взвешенной регрессии:**
   — $R^2 = 0.691$, Adj.$R^2 = 0.672$, RSS = 16.578.
4. **Качество модели:** Модель с географическим взвешиванием показывает улучшенные результаты по сравнению с глобальными.

Из всего этого можно сделать следующие выводы:

— Линейная регрессия удобна для базового анализа, но может не учитывать сложные зависимости и гетероскедастичность;
— Квантильная регрессия предоставляет более детализированное понимание распределения зависимой переменной, что важно при наличии выбросов и асимметрии;
— Географически взвешенная регрессия лучше подходит для данных с выраженной пространственной неоднородностью, предоставляя более точные локальные оценки;

Каждая модель имеет свои преимущества и ограничения, и выбор подходящей модели зависит от специфики данных и целей анализа.

ЗАКЛЮЧЕНИЕ

В данной бакалаврской работе было исследовано моделирование стоимости контрактов на аренду жилья в городе Саратов с использованием регрессионного анализа.

В ходе работы были решены следующие задачи:

— Изучены классические регрессионные модели и модели с географическим взвешиванием.

— Автоматизировано получение выборки предложений на рынке арендуемого жилья.

— Проведен анализ выборки, включающей выставленные предложения по аренде.

— Оценены параметры регрессионных моделей на основе собранных данных.

Результаты проведенного исследования подтвердили значимость различных факторов, таких как расположение жилья и его характеристики, на стоимость аренды. Применение географически взвешенных моделей позволило более точно учитывать пространственную неоднородность данных, что улучшило качество прогнозирования.

В заключение, регрессионное моделирование показало свою эффективность в анализе рынка аренды жилья, предоставив ценную информацию для принятия обоснованных решений как арендаторами, так и арендодателями

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Магнус, Я. Р.* Эконометрика. Начальный курс / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. — 6-е изд., перераб. и доп. изд. — М.: Дело, 2004. — 576 с.

2 *Meinshausen, N.* Quantile regression forests [Электронный ресурс] / N. Meinshausen // *Journal of Machine Learning Research [Электронный ресурс].* — 2006. — .- URL: http://jmlr.org/papers/v7/meinshausen06a.html (дата обращения: 25.03.2024).- Загл. с экрана.- Яз. англ.

3 *Бекирова, О. А.* Квантильная регрессия [Электронный ресурс]. — Большая российская энциклопедия [Электронный ресурс]: научно-образовательный портал. — 2023. — .- URL: https://bigenc.ru/c/kvantil-naia-regressiia-87e087/?v=5807621 (дата обращения: 25.03.2024).- Загл. с экрана.- Яз. рус.

4 *Балаш, В.* Особенности построения географически взвешенной регрессии для моделирования рынка недвижимости / В. Балаш, О. Балаш, А. Харламов // *Промышленность: экономика, управление, технологии.* — 2008. — С. 125–127.

5 *Вуколов, Э. А.* Регрессионный анализ. Методические указания по курсу Статистика / Э. А. Вуколов. — М.: МИЭТ, 2000. — 52 с.: ил.

6 *Демиденко, Е. З.* Линейная и нелинейная регрессии / Е. З. Демиденко. — М.: Финансы и статистика, 1981. — 302 с.: ил.

7 *Балаш, В.* Эконометрический анализ геокодированных данных о ценах на жилую недвижимость [Электронный ресурс] / В. Балаш, О. Балаш, А. Харламов // *Научная электронная библиотека «КиберЛенинка»[Электронный ресурс] : Прикладная эконометрика.* — 2011. — .- URL: https://cyberleninka.ru/article/n/ekonometricheskiy-analiz-geokodirovannyh-dannyh-o-tsenah-na-zhiluyu-nedvizhimost (дата обращения: 25.03.2024).- Загл. с экрана.- Яз. рус.

8 *Бобровская, Е. Д.* Детерминанты цены на краткосрочную аренду жилья в экономике совместного потребления (на примере airbnb в г. Москве) /

Е. Д. Бобровская, А. В. Полбин // *Прикладная эконометрика.* — 2022. — С. 5–28.

9 *Fotheringham, A. S.* Geographically weighted regression with a non-euclidean distance metric: a case study using hedonic house price data / A. S. Fotheringham, B. Lu, M. Charlton, P. Harris // *National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland.* — 2014. — Pp. 1—19.

10 *Pereira, P. T.* Does education reduce wage inequality? quantile regressions evidence from fifteen european countries: Tech. Rep. 120 / P. T. Pereira, P. S. Martins. — Bonn, Germany: IZA, 2000. — Pp. 1—22.

11 requests [Электронный ресурс]. — URL: https://requests.readthedocs.io/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

12 numpy [Электронный ресурс]. — URL: https://numpy.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

13 pandas [Электронный ресурс]. — URL: https://pandas.pydata.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

14 statsmodels [Электронный ресурс]. — URL: https://www.statsmodels.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

15 sklearn [Электронный ресурс]. — URL: https://scikit-learn.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

16 geopandas [Электронный ресурс]. — URL: https://geopandas.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

17 plotly [Электронный ресурс]. — URL: https://plotly.com/python/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

18 pysal [Электронный ресурс]. — URL: https://pysal.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

19 Matplotlib [Электронный ресурс]. — URL: https://matplotlib.org/ (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

20 json [Электронный ресурс]. — URL: https://docs.python.org/3/library/json.html (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

21 csv [Электронный ресурс]. — URL: https://docs.python.org/3/library/csv.html (Дата обращения: 27 апреля года), - Загл. с экрана. - Яз. англ.

ПРИЛОЖЕНИЕ А
Таблицы

Таблица 5 – Параметры объявлений после первичной подготовки данных

Название параметра	**Описание**	**Тип данных**
builtYear	Год постройки здания.	Целое число
price	Стоимость объекта недвижимости.	Дробное число
area	Общая площадь объекта недвижимости.	Дробное число
bathroomUnit	Санузел (совмещенный = 0, раздельный = 1, два санузла и больше = 2).	Целое число
renovation	Состояние ремонта (требуется = 0, косметический = 1, евроремонт = 2, дизайнерский = 3).	Целое число
rooms	Количество комнат.	Целое число
ceilingHeight	Высота потолков.	Дробное число
floor	Этаж, на котором находится объект.	Целое число
total_floor	Общее количество этажей в здании.	Целое число
latitude	Широта местоположения объекта.	Дробное число
longitude	Долгота местоположения объекта.	Дробное число
LIFT	Наличие лифта.	Булево
SECURITY	Общая система безопасности.	Булево
furniture	Наличие мебели.	Булево
improvements	Наличие удобств в квартире.	Булево
first_floor	Квартира на первом этаже.	Булево
last_floor	Квартира на последнем этаже.	Булево
buildingType_BRICK	Тип здания: кирпичное.	Булево
buildingType_MONOLIT	Тип здания: монолитное.	Булево
buildingType_PANEL	Тип здания: панельное.	Булево

Таблица 6 – Статистические данные о характеристиках объявлений

	mean	std	min	25%	50%	75%	max
builtYear	1998.042	20.472	1956.000	1979.000	2007.000	2017.000	2024.000
price	18720.742	6437.079	5000.000	14000.000	17000.000	23000.000	38000.000
area	42.545	10.003	18.000	35.000	40.000	49.000	68.000
bathUnit	0.309	0.467	0.000	0.000	0.000	1.000	2.000
renov	1.311	0.542	0.000	1.000	1.000	2.000	3.000
rooms	1.374	0.677	0.000	1.000	1.000	2.000	3.000
ceilHeight	2.595	0.120	2.000	2.500	2.600	2.614	3.300
floor	5.821	4.295	1.000	3.000	5.000	8.000	22.000
totFloor	10.731	5.525	2.000	9.000	10.000	10.000	26.000
LIFT	0.834	0.373	0.000	1.000	1.000	1.000	1.000
SECUR	0.492	0.500	0.000	0.000	0.000	1.000	1.000
furn	0.731	0.444	0.000	0.000	1.000	1.000	1.000
improv	0.681	0.467	0.000	0.000	1.000	1.000	1.000
firstFloor	0.092	0.289	0.000	0.000	0.000	0.000	1.000
lastFloor	0.059	0.236	0.000	0.000	0.000	0.000	1.000
TypeBrick	0.575	0.495	0.000	0.000	1.000	1.000	1.000
TypeMonolit	0.133	0.340	0.000	0.000	0.000	0.000	1.000
TypePanel	0.291	0.455	0.000	0.000	0.000	1.000	1.000
dist_to_cent	5.986	3.180	0.340	3.003	6.536	8.380	13.031

Таблица 7 – Результат квантильной регрессии для квантиля 0.25

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.4836	0.088	96.714	0.000	8.311	8.656
area	0.0154	0.002	8.551	0.000	0.012	0.019
bathroomUnit	0.0708	0.041	1.723	0.086	-0.010	0.152
renovation	0.1665	0.037	4.554	0.000	0.095	0.238
total_floor	0.0168	0.003	5.435	0.000	0.011	0.023
improvements	0.1097	0.039	2.807	0.005	0.033	0.187
MSE	0.0702					
Breusch-Pagan test:						
	F-statistic: 4.5908					
	P-value: 0.4678					

Таблица 8 – Результат квантильной регрессии для квантиля 0.5

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.6384	0.081	106.315	0.000	8.479	8.798
area	0.0158	0.002	9.128	0.000	0.012	0.019
bathroomUnit	-0.0084	0.039	-0.217	0.828	-0.084	0.067
renovation	0.1797	0.032	5.622	0.000	0.117	0.243
total_floor	0.0150	0.003	4.804	0.000	0.009	0.021
improvements	0.1175	0.036	3.269	0.001	0.047	0.188
MSE	0.0593					
Breusch-Pagan test:						
	F-statistic: 9.1952					
	P-value: 0.1015					

Таблица 9 – Результат квантильной регрессии для квантиля 0.75

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.8177	0.095	92.767	0.000	8.631	9.005
area	0.0145	0.002	7.033	0.000	0.010	0.019
bathroomUnit	0.0763	0.046	1.669	0.096	-0.014	0.166
renovation	0.2030	0.037	5.499	0.000	0.130	0.276
total_floor	0.0154	0.004	3.995	0.000	0.008	0.023
improvements	0.0889	0.042	2.106	0.036	0.006	0.172
MSE	0.1064					
Breusch-Pagan test:						
	F-statistic: 9.1433					
	P-value: 0.1035					

Таблица 10 – Результат построения географически взвешенной регрессии

Model Information				
Model type			Gaussian	
Number of observations:			457	
Number of covariates:			7	
Global Regression Results				
Residual sum of squares			19.015	
Log-likelihood			78.048	
AIC			-142.097	
AICc			-139.775	
BIC			-2737.092	
R^2			0.646	
Adj. R^2			0.641	
Variable	Est.	SE	t(Est/SE)	p-value
const	8.951	0.057	157.564	0.000
area	0.015	0.001	15.038	0.000
renovation	0.138	0.020	6.828	0.000
total_floor	0.016	0.002	8.424	0.000
improvements	0.102	0.022	4.648	0.000
buildingType_PANEL	-0.082	0.023	-3.576	0.000
distance_to_center	-0.034	0.003	-10.262	0.000
Geographically Weighted Regression (GWR) Results				
Spatial kernel			Adaptive Gaussian	
Bandwidth used			70.000	
Residual sum of squares			16.578	
Effective number of parameters (trace(S))			27.246	
Degree of freedom (n - trace(S))			429.754	
Sigma estimate			0.196	
Log-likelihood			109.386	
AIC			-162.280	
AICc			-158.418	
BIC			-45.776	
R^2			0.691	
Adjusted R^2			0.672	
Adj. alpha (95%)			0.013	
Adj. critical t value (95%)			2.498	
Summary Statistics For GWR Parameter Estimates				
Variable	Mean	STD	Min	Max
const	9.015	0.168	8.679	9.243
area	0.014	0.003	0.009	0.019
renovation	0.146	0.036	0.110	0.242
total_floor	0.014	0.005	0.006	0.026
improvements	0.095	0.031	0.026	0.148
buildingType_PANEL	-0.074	0.033	-0.129	-0.011
distance_to_center	-0.037	0.011	-0.057	-0.017