

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
“ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”»

Андреев Иван Васильевич,
Андрюхин Борис Дмитриевич,
Гусева Людмила

**“Исследование рынка недвижимости в Москве на основе объявлений
с ЦИАН”**

**Проект студентов 3 курса бакалавриата
по Эконометрике**

Москва 2025

I. Актуальность.

Рынок недвижимости Москвы - крупнейший в России, к тому же он является самым конкурентным. Цены на недвижимость в столице формируются под влиянием большого количества факторов: расположения объекта, близости к метро, площади, количества комнат, типа жилья (первичное или вторичное), инфраструктуры района, а также типа продавца (собственник, агентство, риелтор), метража объекта и др. Ввиду высокой стоимости объектов, покупатели хотят учитывать все характеристики, влияющие на цену, чтобы избежать переплат.

Платформа ЦИАН - это ведущий агрегатор объявлений о продаже жилья, представляющий обширный и репрезентативный массив данных по рынку. Она охватывает все сегменты московской недвижимости и позволяет отслеживать актуальные рыночные тенденции. Анализ таких данных дает возможность количественно оценить вклад каждого из факторов в формирование цены и построить прогнозные модели.

Актуальность исследования подтверждается рядом научных статей. В работе [Рыбнова и Швакова](#)¹ (2024 г.) обобщаются текущие тренды на рынке недвижимости Москвы, выделяются основные сегменты (первичный, вторичный, массовый, премиальный), описываются модели поведения продавцов, включая различия между агентствами, девелоперами и собственниками. Это подчеркивает высокий уровень конкуренции и разнообразие факторов, подлежащих анализу.

¹ Рыбнов Д. С., Шваков О. М., Власенкова Т. А. Анализ состояния основных тенденций развития рынка недвижимости г. Москвы и его сегментов // Вестник Академии знаний. – 2024. – № 1. – С. 314–317.

II. Гипотезы.

1. Гипотеза 1.

Цена квартиры положительно зависит от близости к станции метро.

Как показано в статье [Киселевой и Сарминой](#)² (2015 г.), транспортная доступность — один из ключевых факторов, формирующих стоимость жилья. На основе корреляционного анализа авторы приходят к выводу, что чем ближе жилье к метро, тем выше его рыночная цена.

2. Гипотеза 2.

При прочих равных условиях квартиры на вторичном рынке имеют более низкую стоимость по сравнению с объектами первичного жилья.

Согласно работе [Малкиной и Щулепниковой](#)³ (2012 г.), в периоды активной господдержки новостройки становятся особенно конкурентоспособными. Кроме того, первичное жилье, как правило, выигрывает по уровню отделки и современным решениям. Это позволяет предположить наличие ценовой премии за «первичный» статус.

3. Гипотеза 3.

Квартиры в центральных районах Москвы стоят существенно дороже, чем аналогичные объекты в удалённых районах.

[Работа Н. Н. Ноздриной и П. П. Макагонова](#)⁴ (2022 г.) демонстрирует наличие выраженной территориальной дифференциации, а именно, центральные районы характеризуются высоким уровнем развития

² Киселева С.С., Сармина Е.Ю. АНАЛИЗ СОСТОЯНИЯ И ТЕНДЕНЦИЙ РЫНКА ЖИЛОЙ НЕДВИЖИМОСТИ В РОССИЙСКОЙ ФЕДЕРАЦИИ // Экономика и социум. - 2015. - №6(19). - С. 249-254.

³ Малкина М.Ю., Щулепникова Е.А. АНАЛИЗ СОСТОЯНИЯ И ТЕНДЕНЦИЙ РЫНКА ЖИЛОЙ НЕДВИЖИМОСТИ В РОССИЙСКОЙ ФЕДЕРАЦИИ // Экономический анализ: теория и практика. - Нижний Новгород: ФИНАНСЫ и Кредит, 2012. - С. 1-12.

⁴ Ноздрина Н.Н., Макагонов П.П. Опыт исследования дифференциации районов Москвы по социально-экономическим показателям // Научные труды: Институт народнохозяйственного прогнозирования РАН. 2022. С. 204-221.

инфраструктуры, социальной привлекательностью и, как следствие, более высокими ценами. Они провели кластерный анализ районов Москвы на основе социально-экономических показателей, таких как уровень доходов населения, занятость, образовательный уровень, что также поможет лучше интерпретировать результаты не только в сравнении центра города и окраин, но и различных округов Москвы между собой.

[Работа Гончарова и Натхова](#)⁵ (2020 г.) показывает, что помимо премии за расположение в центре города, сопоставимые квартиры в западных районах будут дороже, чем в восточных. То есть помимо удалённости от Кремля важен и район.

⁵ Для цитирования: Гончаров Г.И., Натхов Т.В. Текстуальный анализ ценообразования на рынке московской жилой недвижимости. Экономический журнал ВШЭ. 2020; 24(1): 101–116.

III. Данные.

Данные представляют собой объявления о продаже жилой недвижимости в Москве, собранные с помощью парсинга сайта cian.ru. Для этого мы использовали библиотеку python "cianparser". Всего в датасете 1399 объявлений на 19.04.2025.

Также для проверки одной из гипотез нам понадобятся географические координаты станций московского метро. Они собраны с помощью обращения по [API](#) к hh.ru. Координаты квартир были получены с помощью библиотеки "geopy" преобразованием столбца `street` – улица, на которой находится квартира.

Выборка содержит следующие переменные:

1. Зависимая переменная: `price` – цена квартиры.

2. Объясняющие переменные:

- `floor` – этаж;
- `floors_count` - количество этажей;
- `rooms_count` - количество комнат;
- `total_meters` – метраж в метрах квадратных;
- `district` – район;
- `author_type` – тип продавца (агент по недвижимости, риелтор, застройщик, владелец);
- `year_of_construction` – год постройки;
- `metro_distance` – расстояние до метро в метрах – посчитано на основе координат улиц и станций метро;
- `residential_complex` – жилой комплекс (название) или его отсутствие;
- `ao` – административный округ.

Типы переменных:

Признак	Тип данных
author_type	object
floor	int64
floors_count	int64
rooms_count	int64
total_meters	float64
price	float64
year_of_construction	float64
district	object
residential_complex	object
metro_distance	float64
ao	object

Таблица 1, тип переменных

Таким образом, данные содержат четыре категориальных регрессора и шесть числовых (таблица 1).

IV. Предварительный анализ.

1. Пропуски.

Посмотрим на пропуски в датасете и решим, что с ними делать.

Признак	Пропуски
author_type	0
floor	0
floors_count	0
rooms_count	0
total_meters	0
price	20
year_of_construction	3
district	19
underground	10
residential_complex	333
metro_lon	42
metro_lat	42
street_lat	102
street_lon	102
metro_distance	138
ao	0

Таблица 2, пропуски

Ясно, что 20 наблюдений с пропущенной целевой переменной можно удалить из данных, так как они не принесут вклада в обучение модели. Со столбцом residential_complex тоже все просто: пропуск - обычный дом, НЕ жилой комплекс. Присвоим пропускам отдельную категорию no_complex. Наблюдения, для которых отсутствуют и район, и координаты можно

удалить. Для остальных попробуем найти ближайшего соседа по координатам и присвоить тот же район и округ. Наблюдения, для которых пропущены координаты улиц, придется удалить, так как присваивание координат метро бесполезно - пропадет признак `metro_distance`. Координаты районов добавят только шум. Пропуски станций метро и их координат добавим вручную. Остался год постройки - важный признак, влияние которого мы бы хотели исследовать. Заполним медианным значением.

2. Распределения.

Посмотрим на географическое распределение квартир, построенное с помощью библиотеки `folium`.

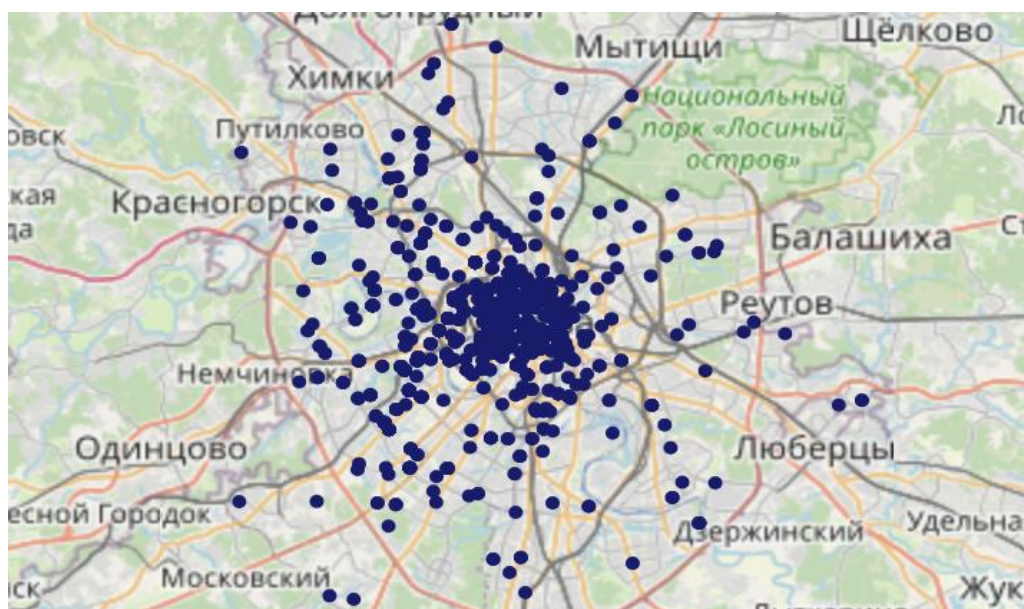


Рисунок 1, географическое распределение

Наблюдается дисбаланс (рисунок 1). Квартиры распределены географически не равномерно. Большинство квартир находятся в ЦАО.

Посмотрим на распределение таргета и его логарифма.

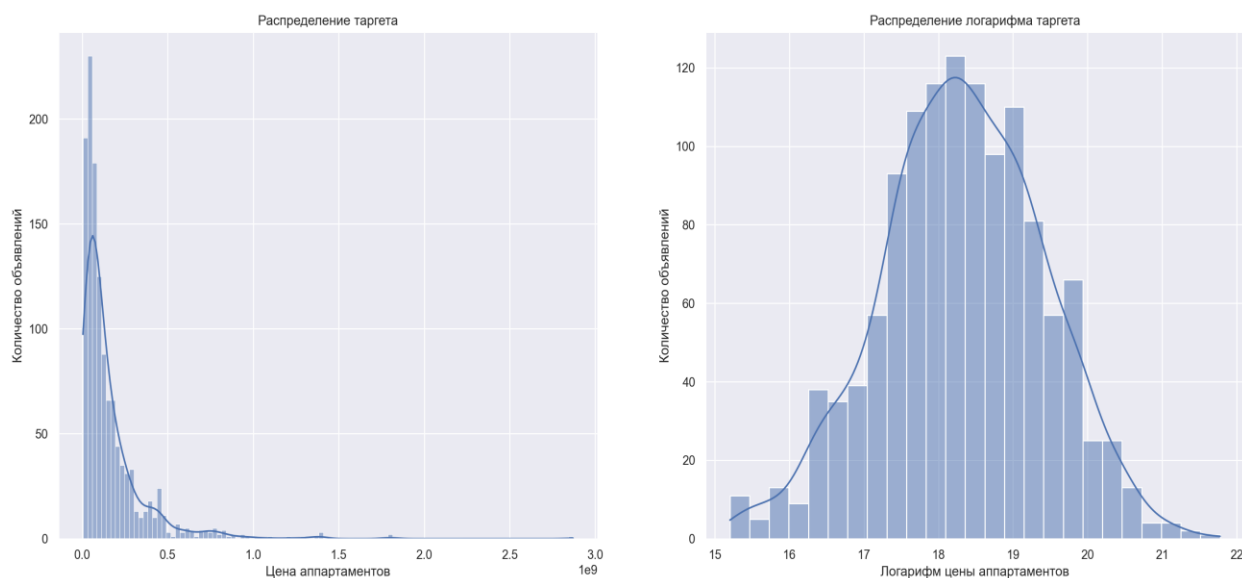


Рисунок 2, распределение таргета и его логарифма

На левом графике видно (рисунок 2), что наблюдаются тяжелые хвосты справа: большая часть наблюдений сосредоточена на низких значениях цены квартиры, а редкие дорогие квартиры растягивают распределение. Логарифм стоимости квартиры снимает асимметрию и приближает данные к нормальному распределению, уменьшая влияние выбросов.

С целевой переменной разобрались. Посмотрим на распределение категориальных признаков. Но для начала определим количество уникальных значений для каждого из них.

Признак	Уникальных значений
author_type	6
district	90
residential_complex	306
ao	9

Таблица 3, уникальные значения категориальных признаков

Признаки `district` и `residential_complex` имеют слишком большое количество уникальных значений (таблица 3). Объединим редкие категории в одну большую и построим графики распределения категориальных признаков.

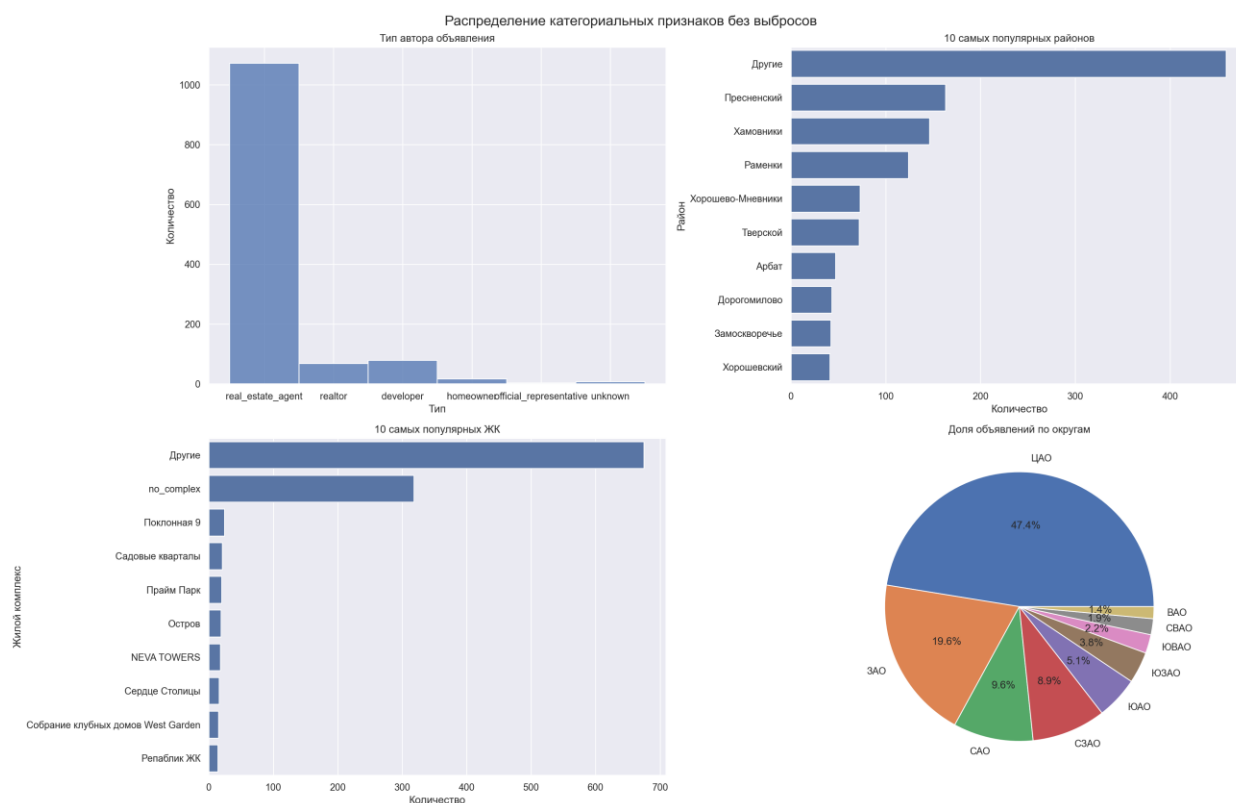


Рисунок 3, распределение категориальных признаков

Исходя из графика (рисунок 3), подавляющее большинство объявлений опубликованы агентами по недвижимости, владельцы публикуют крайне редко. Чаще девелоперы и риелторы. Опасения о географическом дисбалансе подтвердились – большинство районов входят в ЦАО. Однако в выборке представлены объявления и из других административных округов. Большинство объявлений не привязаны к какому-то жилому комплексу. Каждый жилой комплекс является достаточно редкой категорией, несмотря на то, что мы оставили десять

наиболее часто встречающихся. Возможно, данный признак окажется не столь полезным, как мы думали.

Теперь посмотрим на распределение числовых признаков.

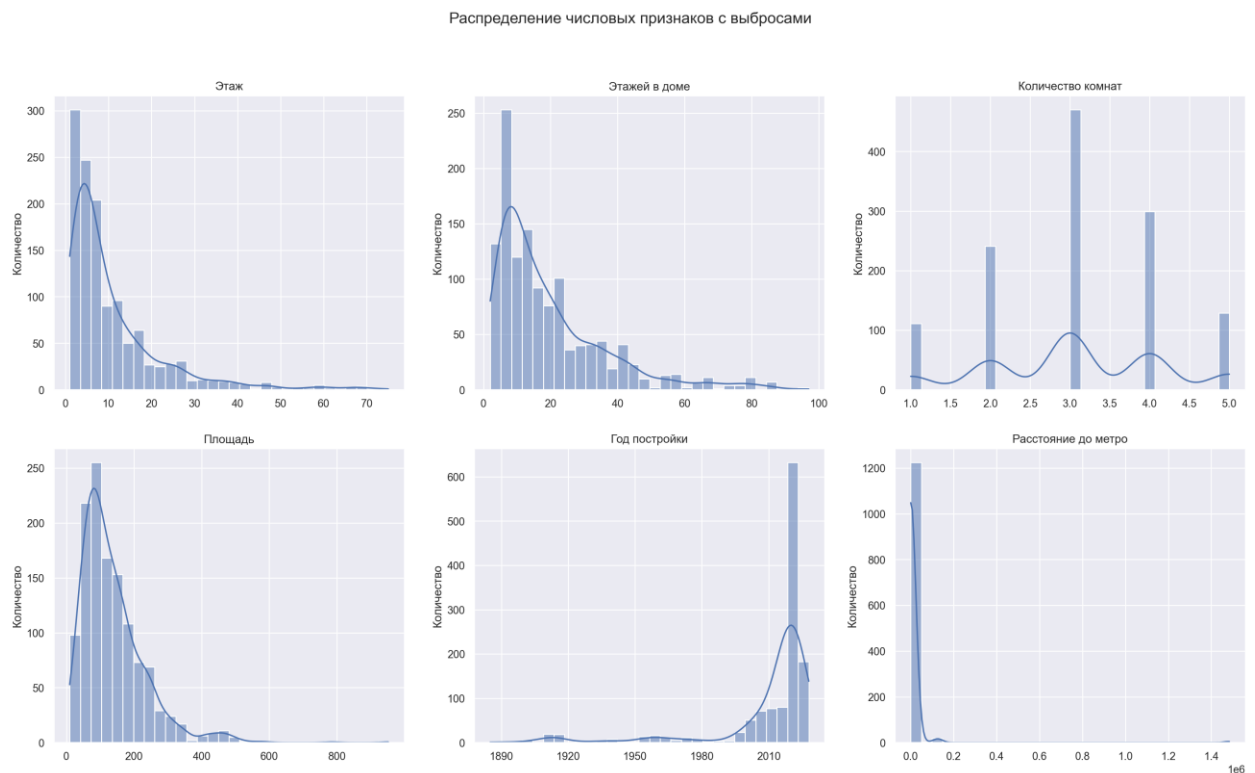


Рисунок 4, распределение числовых признаков с выбросами

Большинство признаков имеют сильную асимметрию и выбросы (рисунок 4). Исправим это с помощью межквартильного размаха.

Распределение числовых признаков (без выбросов)

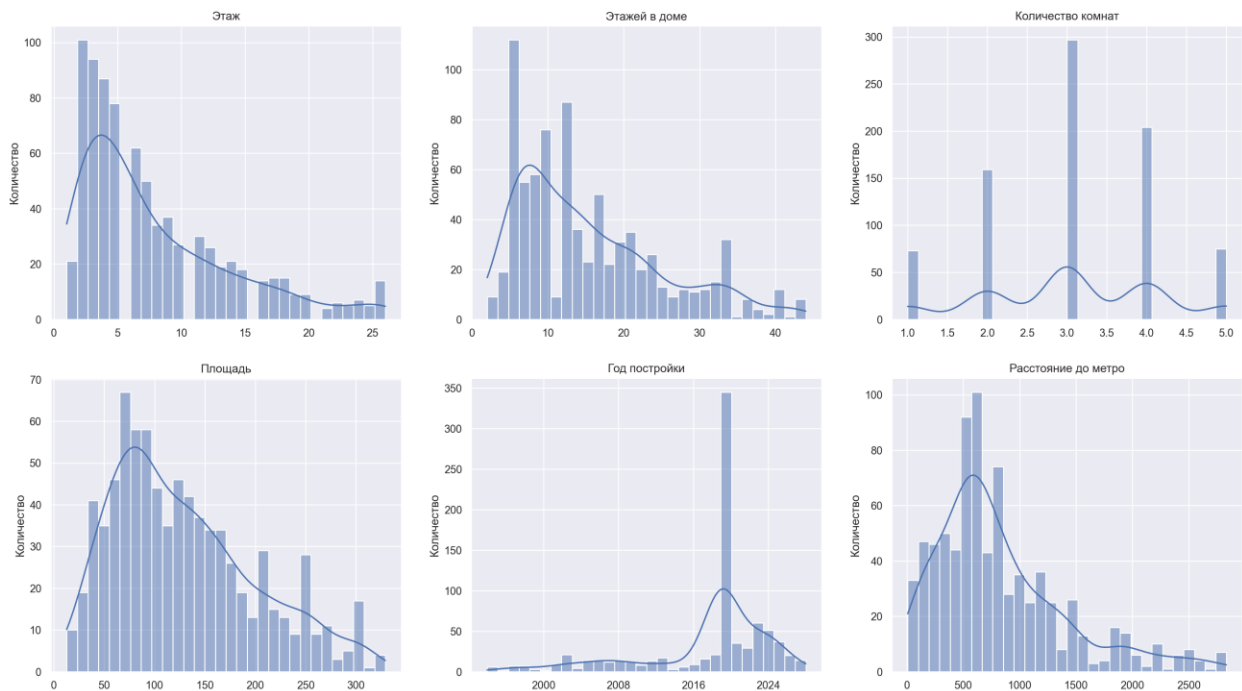


Рисунок 5, распределение числовых признаков без выбросов

Исходя из графиков (рисунок 5), большинство квартир находятся на нижних этажах в домах до двадцати этажей. Трехкомнатные квартиры доминируют, реже встречаются однокомнатные и пятикомнатные. Основная масса квартир имеет площадь до двухсот метров квадратных. Также в выборке представлены квартиры с высокими значениями площади. Большинство квартир находятся в домах, построенных после 2018 года и в пределах километра до ближайшей станции метро.

3. Зависимости.

Построим графики зависимости целевой переменной от числовых признаков.

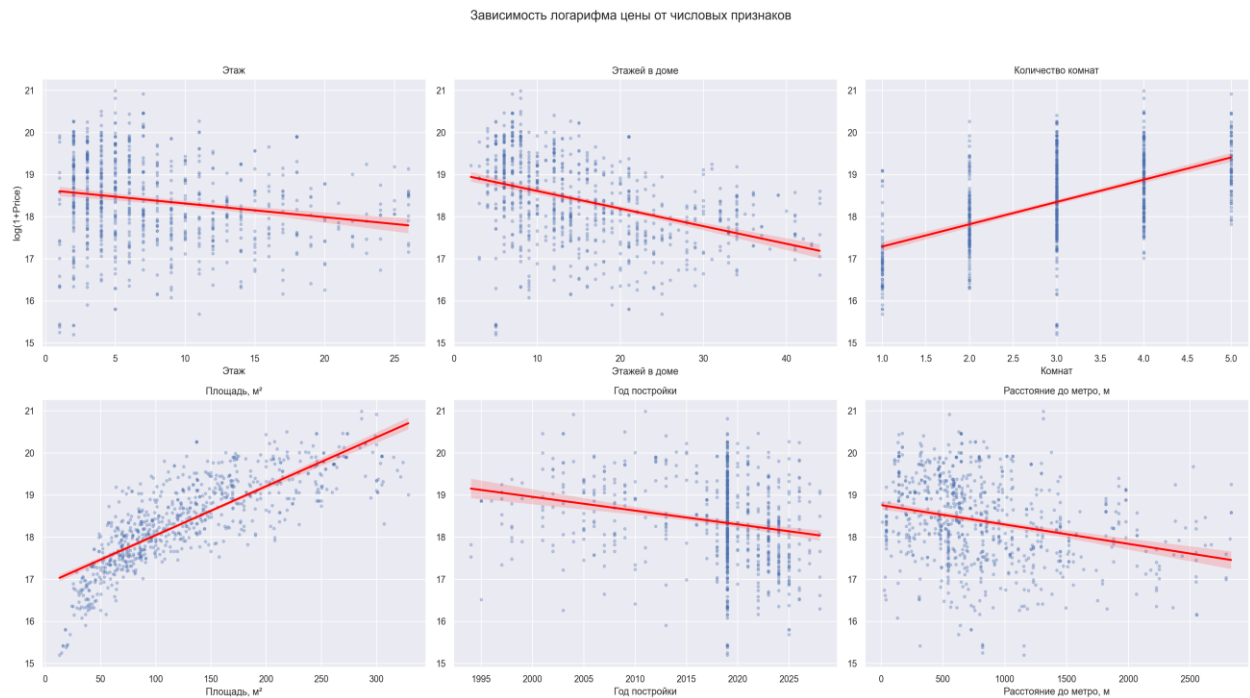


Рисунок 6, зависимость логарифма цены от числовых переменных

Исходя из графиков (рисунок 6), этаж, количество этажей в доме, год постройки здания, расстояние до метро оказывают отрицательное влияние на логарифм цены квартиры. Площадь и количество комнат оказывают более явное положительное влияние.

Построим корреляционную матрицу признаков и зависимой переменной.

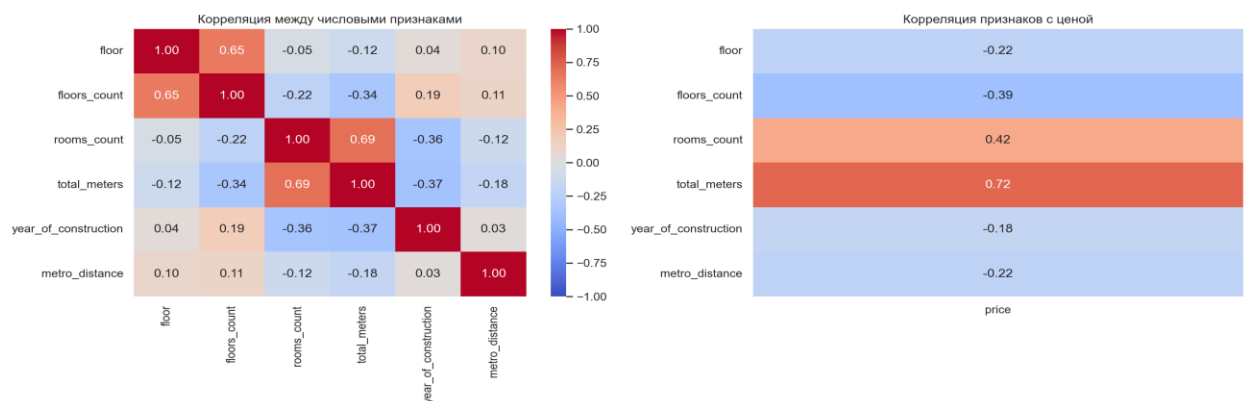


Рисунок 7, корреляционная матрица

Большинство признаков имеют слабую корреляцию между друг другом. Однако пары количества комнат и площади, этажа и количества этажей в доме имеют достаточно высокую корреляцию. Это логично. Чем больше площадь, тем больше комнат, чем больше этажей в доме, тем выше может быть этаж квартиры. Чтобы избежать мультиколлинеарности, придется исключить по одному признаку. Оставим количество этажей и площадь, так как они имеют более высокие значения корреляции по модулю с зависимой переменной. (Комментарий после сдачи proposal: была допущена ошибка с исключением переменных floor и rooms_count в связи с присутствием мультиколлинеарности. Такой подход приведет к смещенным и несостоятельным оценкам. Возникает проблема эндогенности. Все признаки, которые содержательно влияют на зависимую переменную, должны быть включены в модель.) Остальные признаки имеют слабую корреляцию с таргетом.

Наконец, построим ящики с усами для категориальных признаков.

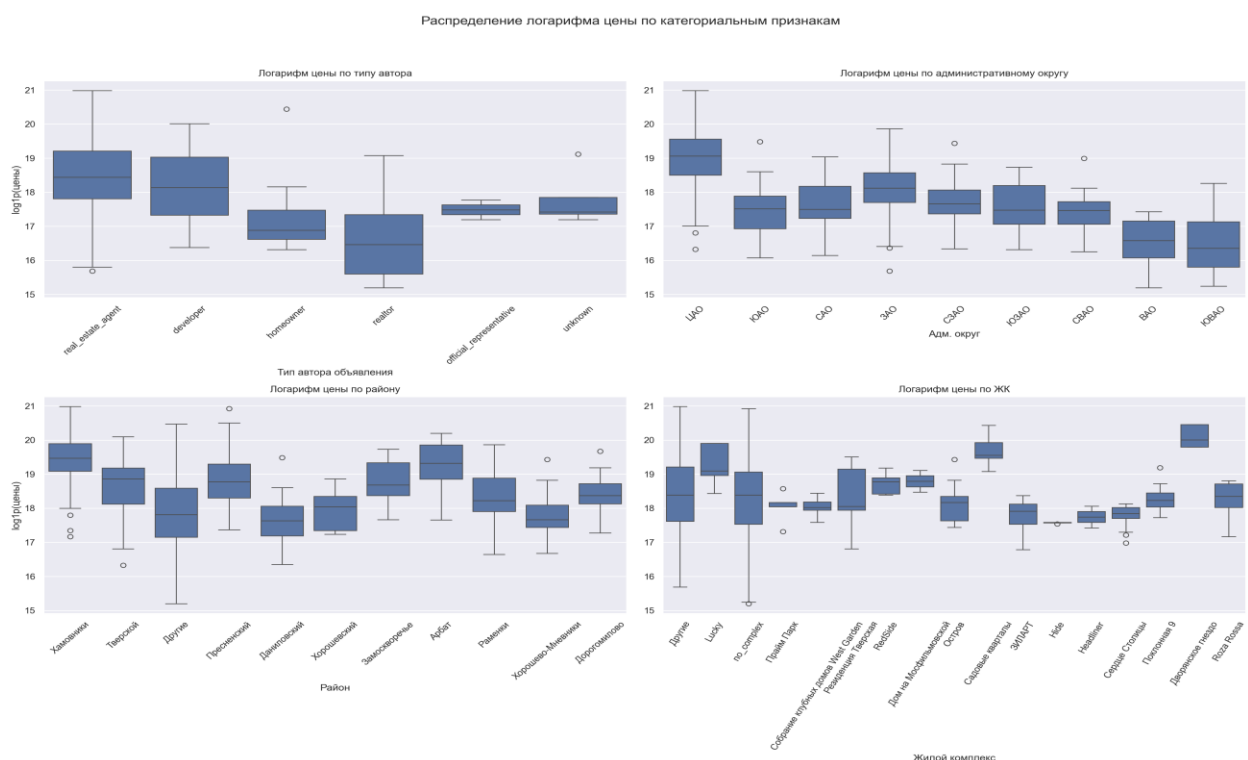


Рисунок 8, зависимость таргета от категориальных признаков

Более высокие значения логарифма цены наблюдаются у агентов по недвижимости и девелоперов, низкие – у риелторов и владельцев. По округам лидирует ЦАО – наибольшее медианное значения логарифма цены квартиры. Наиболее низкие у ВАО и ЮВАО. Хамовники, Тверской и Пресненский – самые “дорогие” районы. По жилым комплексам можно выделить сильно дорогие, однако категории “no_complex” и “Другие” охватывают слишком широкий ценовой диапазон.

V. Эндогенность.

Исходя из предварительного анализа, есть опасения о наличии проблемы эндогенности, искажающей результаты модели. Обратная причинность, пропущенные переменные, одновременность – все эти причины возникновения проблемы могут таиться в наших данных. Далее мы планируем провести тест Хаусмана и выявить наличие или отсутствие эндогенности. По результатам теста мы попробуем использовать инструментальные переменные (IV) и построим двухшаговую МНК (2SLS) (*Комментарий после сдачи proposal: допущена ошибка – тест Хаусмана сравнивает оценки IV и обычного МНК после того, как уже поборолись с эндогенностью*).

VI. МНК.

Разделим X и y , отберем финальные категориальные и числовые признаки.

Создадим дамми переменные `pervichka` и `complex`, обозначающие объявления от девелоперов с годом постройки позже 2025 года и квартиры вне жилых комплексов соответственно. Это интересные признаки для анализа.

В финальный список признаков возьмем `author_type` - тип автора объявления - набор дамми-переменных позволит нам определить присутствует ли в общем наценка различных агентов и риелторов за свою работу, `floors_count` - количество этажей в доме, `floor` - этаж, на котором располагается квартира, `total_meters` - метраж квартиры, `rooms_count` - количество комнат в квартире, `metro_distance` - расстояние от улицы, на которой располагается квартира, до ближайшей станции метро, `ao` - административный округ - набор дамми-переменных позволит нам определить разницу в стоимости квартир в различных округах города и проверить гипотезу о наценке в ЦАО, новые переменные `complex` и `pervichka`.

Исключим из модели переменные `district` и `residential_complex` - районы и жилые комплексы соответственно. Вместо районов мы используем административные округа, оказывающие более явное влияние на зависимую переменную, так как большинство районов в выборке имеют слишком малое количество наблюдений. Вместо названий жилых комплексов используем одну дамми-переменную с принадлежностью к какому-либо комплексу, почти все ЖК очень редкие в нашей выборке и рассмотрение множества дамми-переменных понизит устойчивость модели.

Наконец, давайте обучим OLS с константой и робастными ошибками – предполагаем гетероскедастичность, ведь дисперсии разных ошибок разные - и посмотрим на результаты.

OLS Regression Results

Parameter	Estimate	Std Error	s-Statistic	P-Value
const	16.0195	0.2113	75.8187	0.0000
floors_count	-0.0077	0.0023	-3.3085	0.0009
rooms_count	0.0677	0.0209	3.2451	0.0012
floor	0.0019	0.0031	0.6209	0.5347
total_meters	0.0082	0.0004	22.4180	0.0000
metro_distance	-0.0001	0.0000	-2.9612	0.0031
complex	0.5678	0.0405	14.0062	0.0000
pervichka	0.1784	0.1256	1.4201	0.1556
author_type_homeowner	-0.1485	0.1916	-0.7752	0.4382
author_type_official_representative	-0.4247	0.1097	-3.8711	0.0001
author_type_real_estate_agent	-0.0184	0.0841	-0.2194	0.8263
author_type_realtor	-0.2556	0.1197	-2.1347	0.0328
author_type_unknown	0.0506	0.1124	0.4497	0.6529
ao_3AO	0.7125	0.1860	3.8310	0.0001
ao_CAO	0.4368	0.1927	2.2669	0.0234
ao_CBAO	0.3359	0.1942	1.7297	0.0837
ao_C3AO	0.4316	0.1899	2.2727	0.0230
ao_IIAO	1.1785	0.1881	6.2645	0.0000
ao_IOAO	0.3308	0.1927	1.7168	0.0860
ao_IOBAO	-0.2190	0.2090	-1.0476	0.2948
ao_IO3AO	0.3423	0.1966	1.7407	0.0817

R-squared: 0.8307

Adjusted R-squared: 0.8264

F-statistic: 194.4032

Prob (F-statistic): 1.7289e-287

Это еще не все. Возможно, оцененная модель не имеет смысла с точки зрения интерпретации - оценки МНК несостоятельны - эндогенность. Подумаем об этом.

VII. IV-2SLS.

Эндогенность - это ненулевая ковариация/корреляция одной из объясняющих переменных с ошибками. Проблема чаще всего возникает из-за пропущенных переменных - влияют на Y и связаны с оставшимися в модели X .

На самом деле в нашей модели может быть множество пропущенных переменных:

- Качество отделки, состояние квартиры - очевидно влияет на цену, но также связано, например, с метражом и числом комнат.
- Вид из окна - влияет на цену и связан этажом, на котором располагается квартира.

Это только примеры пропущенных переменных, которые могут быть в нашей модели. Эндогенность есть. Попробуем с ней побороться.

Инструментальные переменные (IV) - метод борьбы с проблемой эндогенности. Эти переменные должны быть связаны с эндогенными признаками, будучи при этом экзогенными. То есть, инструменты влияют на Y только через эндогенный признак, а не сами по себе. Подумаем, какие наши признаки могут быть эндогенными и какие могут быть для них инструменты:

1. `total_meters`: множество пропущенных переменных таких, как качество отделки, планировка, высота потолков, влияют на спрос на

большие или маленькие квартиры. Эти факторы мы не можем включить в модель, и метраж частично забирает на себя их эффект.

2. `rooms_count`: связано с ненаблюдаемыми планировочными идеями, которые могут сильно влиять на стоимость, например, евродвушки имеют метраж, сопоставимый с классической трехкомнатной квартирой, но закрытых спальных пространств меньше. Такие манипуляции явно влияют на зависимую переменную. Количество комнат частично забирает их эффект на себя.
3. `floor`: связано с ненаблюдаемыми преимуществами и недостатками каждого этажа, которые могут влиять на стоимость квартиры. Например, первый этаж обычно является дискомфортным для людей в виду необходимости ставить решетки на окна, обилия запахов с улицы и множества случайных прохожих, а высокий этаж может включать в себя красивый вид из окон. Этаж забирает на себя эффект этих факторов.

Набором инструментов для данных эндогенных признаков могут выступать среднее число этажей, квадратных метров, медиана количества комнат по метро - разные районы имеют разную типичную этажность, метраж и количество комнат. Старые спальники заполнены хрущевками, в центре наблюдается изобилие старинных четырехэтажных домиков с высокими потолками и большими комнатами, а в отдельных высоко технологичных районах типа Москвы-сити находится множество высоток со студиями и небольшими квартирами. Такие средние величины не должны оказывать влияния на стоимость квартиры напрямую, так как не зависит ни от ремонта, ни от состояния квартиры и прочего. Также в качестве дополнительного

инструмента возьмем `floors_count` – само по себе количество этажей в доме не влияет на стоимость, но связано с метражом – в высокоэтажных домах часто располагаются более маленькие квартиры, так как целью таких зданий является размещение большего количества людей - и этажом (по определению высокая корреляция)

Таким образом, получаем три группы признаков:

1. экзогенные переменные: дамми `pervichka`, расстояние до метро `metro_distance`, дамми `complex`, набор дамми типов продавца `author_type_`, набор дамми административных округов `ao_`;
2. эндогенные признаки: метраж `total_meters`, количество комнат `room_count` и этаж `floor`;
3. инструменты: средние этаж и метраж, медиана комнат по метро `mean_floor_by_underground`, `mean_m2_by_underground`, `median_rooms_by_underground` соответственно, количество этажей в доме `floors_count`.

Наконец, давайте обучим IV-2SLS с константой и робастными ошибками, посмотрим на результаты.

IV-2SLS Estimation Summary

Parameter	Estimate	Std Error	T-statistic	P-Value
const	15.7967	0.1907	82.8451	0.0000
pervichka	0.2172	0.1311	1.6567	0.0976
metro_distance	-0.0001	0.0000	-2.2954	0.0217
complex	0.5886	0.0446	13.2122	0.0000
author_type_homeowner	-0.1678	0.1456	-1.1521	0.2493
author_type_official_representative	-0.5605	0.1351	-4.1478	0.0000
author_type_real_estate_agent	-0.1388	0.0995	-1.3951	0.1630
author_type_realtor	-0.2279	0.1201	-1.8981	0.0577
author_type_unknown	-0.0324	0.1132	-0.2860	0.7749
ao_3AO	0.5704	0.1557	3.6624	0.0002
ao_CAO	0.3546	0.1616	2.1945	0.0282
ao_CBAO	0.3280	0.1591	2.0624	0.0392
ao_C3AO	0.3229	0.1578	2.0468	0.0407
ao_ЦАО	0.9934	0.1605	6.1886	0.0000
ao_ЮАО	0.2954	0.1601	1.8446	0.0651
ao_ЮБАО	-0.1845	0.1729	-1.0676	0.2857
ao_Ю3АО	0.2500	0.1695	1.4754	0.1401
total_meters	0.0106	0.0008	12.6646	0.0000
rooms_count	0.0882	0.0507	1.7397	0.0819
floor	-0.0062	0.0044	-1.3895	0.1647

R-squared: 0.8070

Adjusted R-squared: 0.8023

F-statistic: 3149.5441

P-value (F-statistic): 0.0000e+00

Интерпретация результатов:

- $R^2 = 0.807$, модель объясняет около 80.7% дисперсии логарифма цены.
 - `total_meters` - увеличение площади квартиры на 1 м² повышает её цену примерно на 1.06%, значим
 - `rooms_count` - каждая комната увеличивает стоимость на 8.8%, не значим на 5% уровне, только на 10% уровне значим
 - `floor` - влияет слабо и не значим
 - `metro_distance` - увеличение расстояния до метро на 1000 метров снижает цену квартиры на 5.6%, значим
 - `complex` - статус ЖК дает прибавку примерно 80% к стоимости, значим
 - `ao-ЦАО` - квартиры в центре стоят примерно в 2.7 раза дороже, значим. (ЦАО, ЗАО, СВАО, СЗАО дают наибольшие прибавки к стоимости из оставшихся и значимы)
 - `pervichka` - первичное жилье дороже примерно на 24%, не значим на 5% уровне, только на 10% уровне значим
 - `author_type_official_representative` - объявления от представителей сопровождаются меньшей ценой, значим
 - `author_type_realtor` - объявления от риелторов сопровождаются меньшей ценой, значим только на 6% уровне
-

Сравнение OLS и 2SLS:

- Обе модели хорошо объясняют дисперсию цены, у 2SLS показатель немного ниже из-за корректировки на эндогенность
 - Влияние площади выросло с 0.0082 до 0.0106 в 2SLS
 - Количество комнат стало не значимым на 5% уровне, только на 10% значимо, этаж не значим в двух моделях
 - Расстояние до метро стало иметь меньший эффект, но остался значимым и отрицательным
 - Статус ЖК дает примерно одинаковую прибавку в моделях и является значимым в обоих
 - Статус первичного жилья стал значимым на 10% уровне значимости в 2SLS в отличие от OLS
 - Надбавка за местонахождение в ЦАО немного снизилась
-

Посмотрим подробнее так же на первую стадию IV-2SLS.

Диагностика первой стадии IV-2SLS

Endog Variable	rsquared	partial.rsquared	shea.rsquared	f.stat	f.pval	f.dist
total_meters	0.4605	0.2558	0.2010	320.9832	0.0000	chi2(4)
rooms_count	0.3588	0.2103	0.1661	241.2762	0.0000	chi2(4)
floor	0.5561	0.4611	0.4564	730.0555	0.0000	chi2(4)

Все три эндогенные переменные значимо объясняются инструментами, поскольку F-статистики достаточно большие, p-value нулевые, значения Partial R^2 и Shea's R^2 достаточные. Shea's R^2 на

нормальном уровне, поэтому инструменты стабильны в объясняющей силе, невзирая на мультиколлинеарность. Следовательно, можно работать с 2SLS.

Тест Ву-Хаусмана:

Н ₀	Н ₁	Тестовая статистика	P-value
Эндогенные переменные (total_meters, rooms_count, floor) являются экзогенными	Переменные эндогенные	$\chi^2 = 12.59$	0.000

Таблица 4, Тест Ву-Хаусмана

Следовательно, отвергаем Н₀ на уровне значимости 5%.

В модели есть эндогенность, использование модели OLS будет сопровождаться смещёнными оценками.

Тест Саргана:

Н ₀	Н ₁	Тестовая статистика	P-value
Все инструменты валидны, т.е. они не связаны с остатками	Хотя бы один инструмент коррелирует с ошибкой	$\chi^2 = 0.68$	0.411

Таблица 4, Тест Саргана

Следовательно, не отвергаем H_0 на уровне значимости 5%.

Все инструменты, которые используются в модели (mean_floor_by_underground, mean_m2_by_underground, median_rooms_by_underground, floors_count), валидны и не создают смещения.

Проверим устойчивость модели.

Результаты устойчивости IV-2SLS по подвыборкам

Подвыборка	n_obs	coef_m2	p_m2	coef_floor	p_floor	coef_rooms	p_rooms
Первичка	26.000	0.019	0.071	-0.018	0.623	0.012	0.986
Вторичка	779.000	0.010	0.000	-0.007	0.102	0.099	0.047
Низкоэтажные дома (≤ 9)	294.000	0.010	0.000	-0.026	0.471	0.259	0.054
Многоэтажные дома (> 9)	511.000	0.010	0.000	-0.003	0.587	0.102	0.070
Риелтор	26.000	0.040	0.000	0.009	0.172	-0.062	0.210
Не риелтор	779.000	0.010	0.000	-0.010	0.030	0.137	0.012
ЦАО	408.000	0.010	0.000	-0.053	0.000	0.122	0.350
Не ЦАО	397.000	0.009	0.000	0.010	0.028	0.131	0.004

Проверка устойчивости:

1. Площадь:

p-value нулевые во всех подвыборках, кроме первички, p-value=0,07 в этой выборке, но это можно объяснить тем, что в ней всего 26 наблюдений. Во всех подвыборках total_meters стабильно увеличивает цену (коэф. кроме первички в диапазоне 0.009-0.011, в первичке 0.0189). Поэтому можно сказать, что коэффициент устойчив во всех сегментах.

2. Этаж

Неустойчивое влияние и по большей части зависит от географии, поскольку в ЦАО и не в ЦАО значим и коэффициенты противоположных знаков. В ЦАО -0.053, не в ЦАО 0.009. Возможно, причиной служит то, что в центре по большей части дома старые и лифты отсутствуют, в то время как на окраинах дома новее и ценится вид с более высокого этажа. В остальных сегментах площадь чаще не значима.

3. Количество комнат

Зависимость положительная, однако неустойчивая. На вторичке значима и дает по 10% к цене за комнату. Также имеет коэффициенты 0.26 и 0.102 у низкоэтажных и высокоэтажных домов, но p -value 0.054 и 0.07, соответственно. Также имеет коэффициент 0.131 при p -value = 0.004 не в ЦАО, в то время как в ЦАО не значим. Можно сделать вывод о том, что количество комнат важно в массовом семейном сегменте.

VIII. Вывод.

Таким образом, в результате работы над проектом нам удалось собрать данные, провести предварительный анализ, изучая влияние независимых признаков на зависимую переменную и избавляясь от пропусков и выбросов, обучить МНК и IV-2SLS, при работе с которой были найдены подходящие инструментальные переменные для борьбы с эндогенностью, сопоставить результаты двух моделей и проинтерпретировать их, провести статистические тесты Ву-Хаусмана и Саргана и получить удовлетворительные результаты для использования модели, проверить ее устойчивость. Теперь мы можем вернуться к гипотезам из первого раздела:

1. Гипотеза 1 подтверждена: действительно, увеличение расстояния до метро на 1000 метров снижает цену квартиры на 5.6%, коэффициент значим.
2. Гипотеза 2 подтверждена, но только на 10% уровне значимости: первичное жилье дороже примерно на 24%. Коэффициент не значим на 5% уровне.
3. Гипотеза 3 подтверждена: действительно, квартиры в центре стоят примерно в 2.7 раза дороже, коэффициент значим.

Список литературы

1. Рыбнов Д. С., Шваков О. М., Власенкова Т. А. Анализ состояния основных тенденций развития рынка недвижимости г. Москвы и его сегментов // Вестник Академии знаний. – 2024. – № 1. – С. 314–317. – URL: <https://cyberleninka.ru/article/n/analiz-sostoyaniya-osnovnyh-tendentsiy-razvitiya-rynka-nedvizhimosti-g-moskvy-i-ego-segmentov> (дата обращения: 20.04.2025).
2. Киселева С.С., Сармина Е.Ю. АНАЛИЗ СОСТОЯНИЯ И ТЕНДЕНЦИЙ РЫНКА ЖИЛОЙ НЕДВИЖИМОСТИ В РОССИЙСКОЙ ФЕДЕРАЦИИ // Экономика и социум. - 2015. - №6(19). - С. 249-254. URL: <https://cyberleninka.ru/article/n/analiz-vliyaniya-udobstva-raspolozheniya-kvartir-na-ih-stoimost-po-gorodu-moskva/viewer> (дата обращения: 20.04.2025).
3. Малкина М.Ю., Шулепникова Е.А. АНАЛИЗ СОСТОЯНИЯ И ТЕНДЕНЦИЙ РЫНКА ЖИЛОЙ НЕДВИЖИМОСТИ В РОССИЙСКОЙ ФЕДЕРАЦИИ // Экономический анализ: теория и практика. - Нижний Новгород: ФИНАНСЫ и Кредит, 2012. - С. 1-12. URL: <https://cyberleninka.ru/article/n/analiz-sostoyaniya-i-tendentsiy-rynka-zhiloy-nedvizhimosti-v-rossiyskoy-federatsii/viewer> (дата обращения: 20.04.2025).
4. Ноздрина Н.Н., Макагонов П.П. Опыт исследования дифференциации районов Москвы по социально-экономическим показателям // Научные труды: Институт народнохозяйственного прогнозирования РАН. 2022. С. 204-221. URL: <https://cyberleninka.ru/article/n/opyt-issledovaniya-differentsiatsii-rayonov-moskvy-po-sotsialno-ekonomicheskim-pokazatelyam> (дата обращения: 20.04.2025).
5. Для цитирования: Гончаров Г.И., Натхов Т.В. Текстуальный анализ ценообразования на рынке московской жилой недвижимости. Экономический журнал ВШЭ. 2020; 24(1): 101–116. URL: <https://cyberleninka.ru/article/n/tekstualnyy-analiz-tsenoobrazovaniya-na-rynke-moskovskoy-zhiloy-nedvizhimosti/viewer> (дата обращения: 20.04.2025).