

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма	Економічна кібернетика
Галузь знань	05 Соціальні та поведінкові науки
Спеціальність	051 Економіка

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

**на тему «Економетричний аналіз факторів, що впливають на вартість житлової
нерухомості»**

здобувача Майбороди Максима Геннадійович

Науковий керівник: доктор фізико-математичних наук, професор
Притоманова Ольга Михайлівна

(підпис)

**Робота допущена до захисту перед екзаменаційною комісією з
атестації здобувачів вищої освіти (ЕК)**

Завідувач кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

Київ 2023

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1 ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ЦІН НА ЖИТЛО.....	5
1.1. Економічний аналіз ринку житлової нерухомості	5
1.2. Вибір методів економетричного моделювання	10
РОЗДІЛ 2 ПОБУДОВА ЕКОНОМЕТРИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ЖИТЛО ЗІ СТОРОНИ ПОПИТУ	19
2.1 Опис інформаційної бази моделювання	19
2.2 Аналіз змінних	22
2.3 Оцінювання параметрів моделі та її статистичної значущості	29
2.4 Тестування гіпотез автокореляції та гетероскедастичності залишків	35
2.5 Економічна інтерпретація.....	42
РОЗДІЛ 3 ПОБУДОВА ЕКОНОМЕТРИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ЖИТЛО ЗІ СТОРОНИ ПРОПОЗИЦІЇ	45
3.1. Інформаційна база моделювання та аналіз даних	45
3.2. Оцінювання параметрів моделі та її статистичної значущості	51
3.3. Тестування гіпотез автокореляції та гетероскедастичності залишків	55
3.4. Економічна інтерпретація.....	60
ВИСНОВКИ	62
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	64

ВСТУП

Актуальність обраної теми полягає в тому, що з кожним роком в операціях з нерухомістю бере участь все більше громадян, підприємств та організацій. Оскільки на ринку нерухомості з'являється все більша конкуренція потрібно адаптуватися під ринок та визначити, які саме фактори відіграють ключову роль в формуванні ціни на ринку нерухомості [1]. Необхідно відстежувати етапи його становлення та коригувати політику розвитку для побудови правильного підходу. Також це допоможе виокремити переваги від покращення певних характеристик житлової нерухомості.

Підвищення ефективності функціонування ринку житлової нерухомості є важливим завданням економіки, оскільки воно вимагає великих фінансових, матеріальних та людських ресурсів. Тому важливо забезпечувати більшу цінність для клієнтів і залишатися гнучкими, незважаючи на глобальні виклики чи тиск ринку.

Метою роботи є визначення факторів, які впливають на вартість житлової нерухомості, шляхом застосування методів економетричного моделювання та отримання кількісної оцінки цього впливу.

Відповідно до поставленої мети визначені такі завдання:

1. Визначити принципи функціонування ринку нерухомості.
2. Проаналізувати сучасні тенденції розвитку ринку нерухомості.
3. Визначити основні техніки аналізу ринку нерухомості та математичні алгоритми для моделювання ціни на нерухомість.
4. Виконати економетричний аналіз факторів, що впливають на ціну житлової нерухомості на основі двох підходів: з боку попиту на ринку нерухомості та з боку пропозиції на ринку нерухомості.

Об'єктом дослідження є ринок житлової нерухомості та процеси формування цін на житло.

Предметом дослідження є математичні методи і моделі для виявлення факторів, що впливають на ціну житлової нерухомості.

Методологія та методи дослідження. Економетричне моделювання використовується для побудови економіко-математичної моделі та встановлення зв'язків між економічними показниками та факторами, що впливають на ціни на житло. Регресійний аналіз вимірює вплив факторів на зміну цін на житло. Кореляційний аналіз вивчає залежність між змінними, що можуть впливати на вартість житлової нерухомості. Аналіз з використанням фіктивних змінних дозволяє враховувати вплив інших факторів, що не включені безпосередньо в модель. Метод найменших квадратів використовується для оцінки параметрів моделі шляхом мінімізації відхилень між спостережуваними та розрахованими значеннями.

Теоретична, методична та практична значущість отриманих результатів полягає в тому, що техніки економетричного аналізу та моделювання ціни, які представлені в дослідженні, є корисними для різних суб'єктів ринку нерухомості. Наприклад, громадяни можуть використовувати ці дані, щоб дізнатися про залежність ціни житлової нерухомості на ринку у своїх регіонах та обрати житло з своїм бюджетом. Маркетологи, продавці та дослідники ринку нерухомості можуть дізнатися більше про поведінку та вподобання споживачів, а також про загальну ситуацію на ринку. Держава може використовувати ці дані для розробки раціональної державної політики, аналізувати, як, наприклад, тимчасове скасування певних податків або введення обмеження на певні характеристики вплине на попит та пропозицію на ринку, і проектування інфраструктури, яка враховує потреби різних сторін на ринку нерухомості.

Інформаційною базою дослідження є роботи провідних науковців та практиків зарубіжних та вітчизняних установ, статистичні дані з різноманітних спеціалізованих джерел про ринок житлової нерухомості в Інтернеті.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ЦІН НА ЖИТЛО

1.1. Економічний аналіз ринку житлової нерухомості

Поняття "нерухомість" вже включає дві основні категорії об'єктів, що мають різне походження: природні та штучні об'єкти нерухомості. Природні об'єкти включають земельні ділянки, ліси, багаторічні насадження та ділянки надр. Ці об'єкти також відомі як "природна нерухомість". Штучні об'єкти включають дві категорії: житлові та нежитлові приміщення.

Ринок нерухомості можна охарактеризувати як комплекс взаємозв'язаних відносин, пов'язаних з операціями щодо об'єктів нерухомості, таких як створення нових об'єктів або розвиток існуючих, купівля-продаж, іпотека, оренда та інші (рис. 1.1).

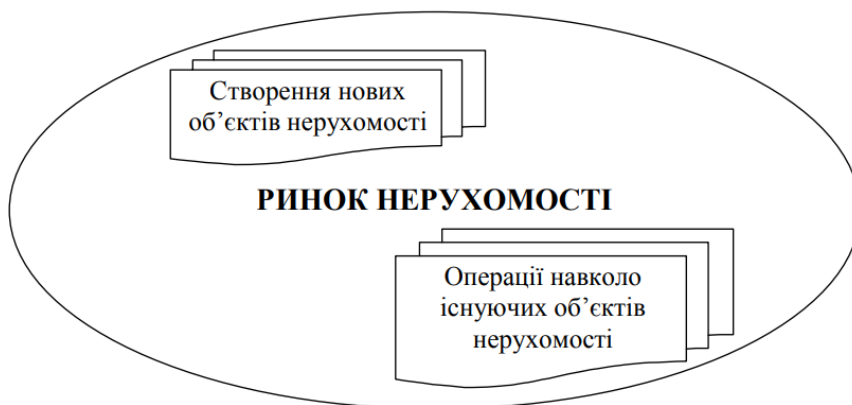


Рисунок 1.1 – Сутність ринку нерухомості [2]

Ринок нерухомості має складну й розгалужену структуру, і його можна класифікувати за різними ознаками (рис. 1.2).

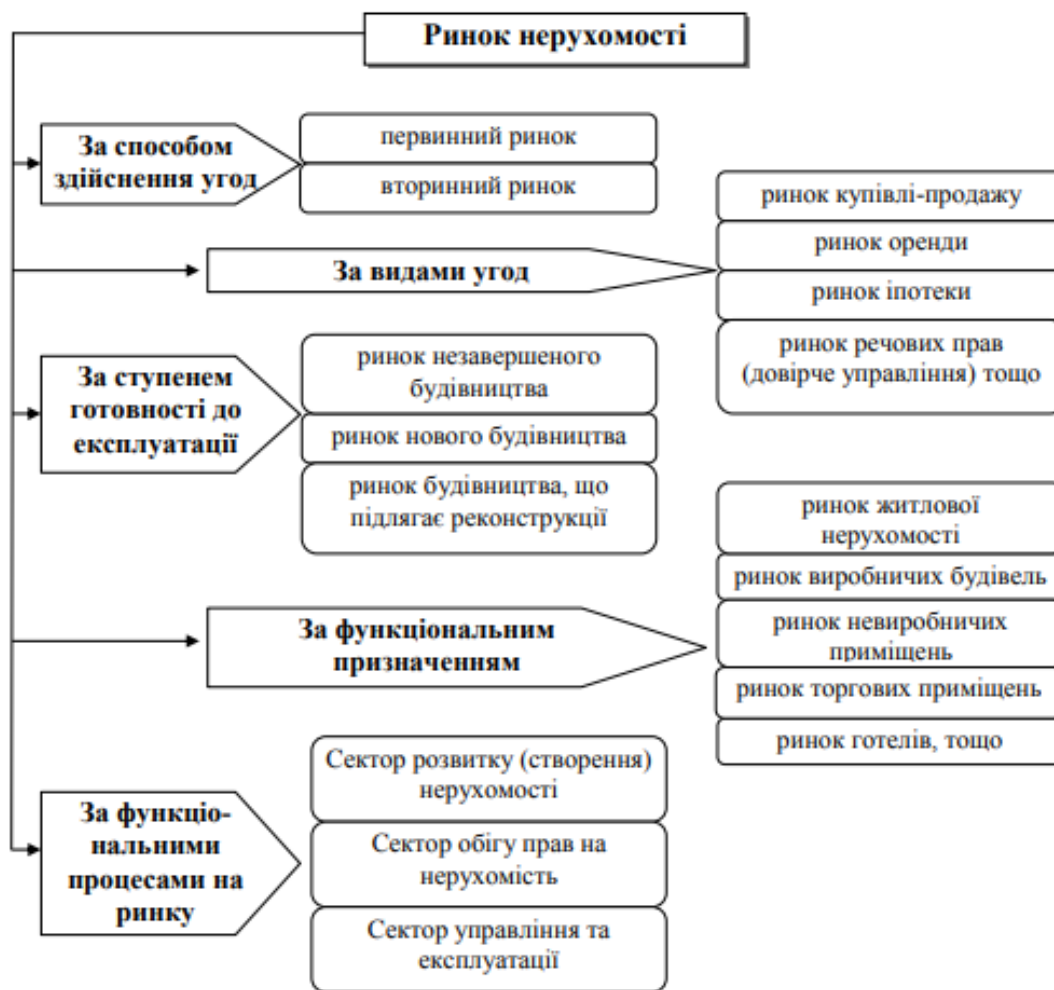


Рисунок 1.2 – Види ринків нерухомості [2]

Нерухомість може бути класифікована за різними характеристиками, що сприяє більш ефективному аналізу ринку нерухомості та спрощує розробку та використання методів оцінки різних типів нерухомості.

У цьому дослідженні ми зосередимося на аналізі ринку житлової нерухомості, що включає: будинки, котеджі, квартири. Житлова нерухомість включає: невисокоповерхові будинки (до 3 поверхів), багатоповерхові будинки (4-9 поверхів), будинки середньої висоти (10-20 поверхів), хмарочоси (понад 20 поверхів). До житлової нерухомості також можуть відноситися кондомініуми, секції, під'їзди, окремі квартири, кімнати, дачні будинки.

Житловий сектор у більшості країн є головною рушійною силою економічного зростання та є популярною темою для вивчення. З точки зору державного сектору, послаблення цього сектора має серйозні наслідки для майбутнього економічного зростання, оскільки він становить значну частину державного доходу. З точки зору домогосподарств, він становить найбільшу частину їхнього багатства, що впливає на споживчі витрати, рішення щодо заощаджень і приплив робочих місць у житловому будівництві.

Протягом тривалого часу ринок нерухомості вважали основою економічного зростання, рушійною силою розвитку національної економіки. Проте постійне зростання цін на житло спричинило багато економічних і соціальних проблем.

Загальна тенденція значного зростання цін на нерухомість на світовому ринку житла зараз різко сповільнюється. Це не дивно, враховуючи сукупні несприятливі наслідки підвищення процентних ставок, стрімкого зростання інфляції, уповільнення світової економіки та збоїв у ланцюзі поставок, спричинених пандемією Covid-19 і ще більше посилені наслідками війни в Україні.

Якщо говорити за дослідження світового ринку протягом 2022 року, то ситуація виглядає наступним чином [3]:

- У Сполучених Штатах, які завжди є ключовим ринком, імпульс різко послабився: індекс FHFA підвищився лише незначно, а індекс Кейс-Шіллера зараз падає, порівняно з кількома роками сильного зростання, на тлі падіння попиту та негативних настроїв будівельників. Канада також демонструє схожу тенденцію.

- Європейські ринки житла зараз вичерпуються, оскільки ціни на житло зростають лише у 8 із 29 європейських країн, включених до нашого глобального дослідження у 2022 році. Реальні ціни на житло на ключових європейських ринках, таких як Німеччина, Великобританія, Іспанія, Франція та Італія, є зараз падає.

- Тихоокеанські ринки житла, які раніше бурхливо розвивалися, тепер переживають проблеми. Зараз в Австралії та Новій Зеландії спостерігається різке падіння цін на нерухомість на тлі падіння попиту на нерухомість.

- Азіатські ринки житла слабшають. Реальні ціни на житло зросли лише на 6 із 14 ринків, включених до дослідження, і в більшості азіатських країн темпи також менші. Яскравим винятком є В'єтнам і Шрі-Ланка.

- Ринки житла в Латинській Америці продовжують боротися, оскільки ціни на нерухомість в Аргентині та Колумбії різко впали.

Оскільки розвиток ринку нерухомості більше сприяє заробітку людей, тому особливо необхідно вивчати фактори та коливання, що впливають на ціну нерухомості. На ціни впливають численні комплексні фактори.

В економіці ефективний розподіл ресурсів вимагає, щоб у довгостроковій перспективі ціна будь-якого ринкового товару чи послуги визначалася балансом попиту та пропозиції. Це також справедливо для ринку житла, де пропозиція будинків для продажу та попит на житло мають визначати ціни.

На добре функціонуючому ринку житла кількість нерухомості для продажу має визначатися факторами, що впливають на наявний житловий фонд, включаючи наявність землі та вартість будівництва нових будинків. Попит на житло визначається потребою в житлі, зумовленою демографічними розвитком, а також платоспроможністю (доходи, доступ до фінансування) і бажанням придбати будинок [2].

На відміну від більшості ринкових товарів і послуг, де попит потрібен сам по собі (ми вимагаємо яблуко, тому що хочемо його спожити), деякі товари або послуги натомість мають попит, тому що вони забезпечують доступ до інших товарів. Транспортні послуги, наприклад, часто потрібні для того, щоб ми могли зрештою спожити інший товар або послугу (наприклад, кіно). Житло також у багатьох відношеннях є товаром, попит на який поширений опосередковано – з точки зору

доступу до місцевих об'єктів, можливостей працевлаштування та інших послуг, які воно надає [4]. Через це бажання придбати будинок вплине на попит ринку. Високі ціни на житло в певному районі можуть відображати відносну кількість зручностей і пропонувати мешканцям високу якість життя. Велика кількість зручностей Лондона та доступ до можливостей працевлаштування в цьому сенсі можуть певною мірою компенсувати його мешканцям високу вартість житла.

Високі ціни на будь-якому ринку відображають перетин сильного попиту та обмеженої пропозиції. Якщо за будь-якої даної поточної ринкової ціни кількість попиту на будинки зростає, то ціни зростатимуть за умови фіксованої пропозиції. Такі зміни, якщо вони викликані фундаментальними детермінантами попиту та пропозиції, можуть призвести до динамічної реакції пропозиції, тобто, де це можливо, збільшиться кількість пропонованих будинків. Зростання цін, викликане цими фундаментальними факторами, має залишатися ефективним у довгостроковій перспективі [5].

Ціна на приватне ринкове житло теоретично повинна відображати баланс між пропозицією будинків і попитом на проживання в них. На практиці, однак, попит і пропозиція на житло можуть також спотворюватися іншими факторами, не пов'язаними з їх основними детермінантами [6].

З боку попиту, наприклад, використання житла як довгострокового інвестиційного активу може стимулювати додатковий попит з боку інвесторів і власників будинків, які прагнуть спекулювати на вартості нерухомості в очікуванні реалізації майбутнього приросту капіталу. У свою чергу, було стверджено, що «недоступність» на рівні міста сама по собі може розглядатися як показник бажаності району та якості життя, який він пропонує [7]. Інколи попит на житло може збільшити його понад вартості на ринку, отриманого від проживання в певному регіоні, місці, вулиці тощо (оскільки інвестори очікують майбутнього зростання цін).

Отже, ринок нерухомості має свої унікальні характеристики, структуру суб'єктів та виступає як інтегрована категорія, яка поєднує ознаки ринків товарів, інвестицій та послуг. Також в роботі буде досліджено ринок нерухомості з двох головних аспектів: попиту та пропозиції.

1.2. Вибір методів економетричного моделювання

За останні роки методологія побудови систем оцінки нерухомого майна вдосконалюється тільки за рахунок створення технологій, які враховують різні аспекти та особливості територій, по якій здійснюється оцінка. При цьому класичним математичним інструментарієм дослідження ринку житлової нерухомості й досі залишається економетричний аналіз

Учені провели чимало досліджень: Case і Shiller провели регресійний аналіз і довели, що ціна житла корелює з чисельністю населення, доходом і прибутками ринку нерухомості [8]. Quigley підкреслив, що деякі економічні зміни, такі як населення, зайнятість, індекс економіки, можуть передбачити тенденцію цін на житло. Усі вони довели, що ціни на житло можна моделювати та прогнозувати [9]. Однак, через обмеженість і однобічність своїх відповідних поглядів, вони не надали достатньої теоретичної бази щодо коливання ціни на житло та її впливових факторів. Крім того, вони не змогли забезпечити систематично прийнятну модель ціни на житло [10].

Економетрична модель – це функція або набір функцій, що описує зв'язок між економічними показниками на основі кореляційно-регресійного аналізу. Вона враховує причинно-наслідкові зв'язки між цими показниками [11].

У загальному випадку рівняння в економетричній моделі має вигляд:

$$Y = f(x_1, x_2, \dots, x_i), \quad (1.1)$$

де: Y – результат, або залежна змінна, змінювання якої описує дане рівняння;

$x_1, x_2, \dots, x_i$ – фактори, або незалежні змінні, що визначають поведінку Y .

Одним з основних підходів у вимірі зв'язку між досліджуваними показниками в економетричній моделі є кореляційно-регресійний аналіз.

Кореляційно-регресійний аналіз включає побудову та аналіз економіко-математичної моделі у формі регресійного рівняння (або рівняння кореляційного зв'язку). Це рівняння виражає залежність між результативною ознакою та однією або кількома факторними ознаками, а також надає оцінку міри інтенсивності цього зв'язку [12].

Економетрична модель призначена для аналізу і прогнозування економічних процесів і явищ, що розглядаються, в умовах невизначеності інформації за допомогою методів математичної статистики.

Економетрична модель поєднує два аспекти: теоретичний, який включає якісний аналіз зв'язків, і емпіричний, що базується на фактичних даних. Теоретична інформація відображається у специфікації моделі [13].

На основі досвіду економетричних досліджень та якісного теоретичного аналізу взаємозв'язків економічних показників можна виділити клас функцій, які можуть описувати ці зв'язки. Наприклад, лінійна функція:

$$Y = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n, \quad (1.2)$$

де: Y – залежна (пояснювальна) змінна;

$x_1, x_2, \dots, x_n$ – незалежні, або пояснювальні змінні;

a – параметри функцій.

Вибір аналітичної форми економетричної моделі залежить від конкретного набору незалежних змінних. Тому специфікація моделі передбачає вибір факторів для проведення економетричних досліджень [14].

Етапи побудови економетричної моделі:

1. Початковий аналіз економічного процесу що розглядається (його теоретичний опис з відображенням існуючих тенденцій і якісний економічний аналіз).
2. Визначення мети дослідження, введення обмежень і припущень.
3. Вибір факторів, істотних з погляду мети дослідження (специфікація змінних).
4. Вибір математичної моделі з фіксованого формою всіх структурних рівнянь.
5. Збір і попередній аналіз даних для змінних що входять у модель.
6. Вибір методу оцінювання невідомих параметрів моделі з урахуванням припущень про імовірні властивості відхилень.
7. Реалізація процедури оцінювання що забезпечує найкраще наближення модельних значень змінних до їхніх значень, що спостерігались у дійсності.
8. Перевірка якості отриманих оцінок, а також основних припущень.
9. Інтерпретація отриманих результатів, встановлення їхньої адекватності поставленим цілям.
10. Аналіз неузгодженості і коригування моделей.
11. Прийняття рішень щодо наступного дослідження з урахуванням альтернативних можливостей [15].

Для оцінки параметрів та аналізу моделі ми користуємося функцією `summary()` у середовищі програмування RStudio.

Коефіцієнт детермінації розглядають, як правило, в якості основного показника, що відображає міру якості регресійній моделі, яка описує зв'язок між

залежною і незалежними змінними моделі. Коефіцієнт детермінації показує зміну залежної змінної, що пояснюється незалежними змінними в моделі.:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_t)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.3)$$

де: y_i – значення спостережувальної змінної;

$\bar{y}$ – середнє значення по спостережуваним даними;

$\hat{y}_t$ – модельні значення, побудовані по оціненим параметрам.

Чим ближче R^2 до 1, тим вища якість моделі. У разі рівного розподілу коефіцієнта одиниці лінія регресії точно відповідає всім спостереженням [16].

Рівність коефіцієнта нулю означає, що обрані фактори не покращують якість передбачення y_i в порівнянні з тривіальним пророкуванням $\hat{y}_t = \bar{y}$. Досить якісною можна визнати модель з коефіцієнтом детермінації вище 0,8.

У нашому випадку існує 2 варіанти детермінації Multiple R-squared – звичайний коефіцієнт детермінації та Adjusted R-squared – що враховує предиктори.

Residuals – це вертикальна відстань між точкою даних та лінією регресії. Кожна точка даних має один залишок. Вони є позитивними, якщо вони знаходяться вище лінії регресії, і негативними, якщо вони знаходяться нижче лінії регресії

Estimate показує параметри кожної змінної – це коефіцієнти, що множаться предиктори у лінійному рівнянні, у даному рівнянні параметрами моделей є b_0 та b_1 .

$$\hat{Y}_t = \hat{b}_0 + \hat{b}_1 x_i \quad (1.4)$$

Std. Error – це статистичний термін, який вимірює точність, з якою розподіл вибірки представляє сукупність, використовуючи стандартне відхилення [17]. У статистиці середнє значення вибірки відхиляється від фактичного середнього

значення сукупності – це відхилення є стандартною помилкою середнього значення. Чим більше є стандартне відхилення тим, менш ефективною буде економетрична модель. Низький показник стандартного відхилення свідчить про нормальний розмах варіації.

Residual standard error – стандартна похибка залишків, що є звичайними стандартними похибками, які стосуються не змінних, а залишків.

Критерій Фішера – оцінює статистичну значущість моделі та критерія R^2 .

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad (1.5)$$

де: k – кількість змінних;

n – кількість спостережень

Якщо F розрахункове $>$ F табличне, то коефіцієнт детермінації R^2 є статистично значущим, і навпаки.

Значущість моделі за t-критерієм (критерій Стьюдента):

$$t_0 = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} \quad (1.6)$$

t_0 – критерій оцінює статистичну значущість оцінок параметрів моделі;

M_1, M_2 – значення порівнюваних середніх арифметичних;

m_1, m_2 – відповідні величини статистичних помилок середніх арифметичних.

Якщо $t_0 > t_{\text{табл}}$, то відповідний параметр моделі є статистично значущий, і навпаки.

Коефіцієнт кореляції позначається r . Варіюється в межах від 1 до -1. При значенні коефіцієнта 0, лінійної залежності між вибірками не існує [18].

Кореляційний аналіз відповідає за ступінь зв'язку між двома змінними, x і y .

$$r = \pm\sqrt{R^2} \quad (1.7)$$

$$r = \frac{\sum(x_t - \bar{x})(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2} \sqrt{\sum(y_t - \bar{y})^2}} \quad (1.8)$$

Він відображає тісноту взаємозв'язку між залежною і незалежною змінною.

Якщо зв'язок є, то збільшення одного параметра впливає на підвищення (позитивна кореляція) або зменшення (негативна) іншого. Кореляційний аналіз допомагає аналітику визначити, чи можна за величиною одного показника передбачити можливе значення іншого.

Для визначення параметрів рівняння регресії найчастіше застосовується класичний метод найменших квадратів (ІМНК) основоположниками якого є К. Гаусс та П. Лаплас.

МНК дозволяє підібрати певну неперервну аналітичну функцію для апроксимації (наближення) дискретного набору вихідних даних. Задача зводиться до знаходження такої лінії (прямої, кривої, ламаної тощо), яка мінімізує суму квадратів відхилень між фактичними y та отриманими розрахунковим шляхом за рівнянням регресії теоретичними значеннями $\hat{y}_t$. [19, 20].

Окрім описаних дій будуть проводитись дослідження на напруженість, викиди, залишки, впливові спостереження і т.д.

Residual vs. Fitted plot – почнемо з розгляду залишкової діаграми, що виходить з лінійної моделі, пристосованої до даних, які цілком задовільняють всі стандартні припущення про лінійну регресію. В ідеальному випадку ми очікуємо, що i -а точка даних буде сформована як:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i \quad (1.9)$$

ϵ_i – термін «помилки» із середнім значенням 0 та деякою дисперсією σ^2 .

Outliers and the Residuals vs Leverage plot. Немає єдиного загальноприйнятого визначення того, що є незвичним. Одне з можливих визначень полягає в тому, що відхиленням є будь-яка точка, яка недостатньо апроксимується моделлю (має великий залишок) і яка суттєво впливає на пристосованість моделі (має великий важель впливу). Тут з'являється сюжет Residuals vs Leverage. Ідеальний випадок сюжету Residuals vs Leverage – в якому ми не бачимо доказів викидів [21].

Normal QQ plot. Графік нормального QQ допомагає оцінити, чи залишки нормально розподілені. Якщо залишкові величини мають тенденцію бути більшими за величиною, ніж ми могли б очікувати від нормального розподілу, тоді наші р-значення та довірчі інтервали можуть бути занадто оптимістичними. тобто, ми не можемо належним чином врахувати повну мінливість даних. Ідеальний випадок коли залишки ідеально збігаються з діагональною лінією, тоді ці залишки є нормально розподіленими [22].

Визначення поодиночної кореляції незалежних змінних із залежної змінної. Варто розуміти наскільки ми можемо передбачити значення y в залежності від значення x .

Далі проведемо `OutlierTest()`, щоб визначити значення, які віддалені від інших спостережень, тобто ті, які суттєво відрізняються від інших точок даних, тому можуть суттєво змінювати модель.

Застосуємо функцію `influence.measures()`. Основною функцією високого рівня є, який створює стовпчик "infl", що відображає DFBETAS для кожної змінної моделі, DFFITS, коефіцієнти коваріації, відстані Кука та діагональні елементи матриці капелюхів. Випадки, які впливають на будь-який із цих заходів, позначені зірочкою [23].

Перевіримо нашу математичну модель на ряд показників:

- Global Stat – чи є лінійний зв'язок між x та y ;

- Skewness – показник, який визначає, чи містять хвости даного розподілу екстремальні значення;
- Link Function – чи розподілена змінна лінійно;
- Heteroscedasticity (Гетероскедастичність) – поняття, що означає неоднорідність спостережень, що виражається в неоднаковою дисперсії випадкової помилки регресійної моделі.

Необхідно виконати перевірку на мультиколінеарність за допомогою загальної та індивідуальної перевірки.

Загальна перевірка `omcdiag()` – ця функція виявляє наявність мультиколінеарності, використовуючи різні доступні діагностичні методи, вже наявні в літературі, такі як Детермінант $|X'X|$, Фаррар R^2 , χ^2 , Condition index, Sum of reciprocal of eigenvalues, Індикатор Тейла (Theil's indicator), Червоний індикатор (Red indicator) [24].

Функція також відображає значення діагностичних вимірів, рішення про те, що мультиколінеарність наявна, чи ні. Значення 1 вказує на те, що виявлено мультиколінеарність, а 0 – не вдається виявити мультиколінеарності за допомогою певного діагностичного методу. Чим більше тестів показало значення, тим більшою є мультиколінеарність незалежних змінних.

Індивідуальна перевірка `imcdiag()` – обчислює різні заходи діагностики мультиколінеарності для кожної незалежної змінної. Індивідуальні показники включають дисперсію коефіцієнта інфляції (VIF), F-тест Фаррара для визначення мультиколінеарності, допоміжний F-тест на взаємозв'язок між F та R-квадратом, метод Лімера, виправлений VIF (CVIF) та правило Клейна [25].

Визначається найбільше значення мультиколінеарності, які за певних умов бажано прибрати для покращення математичної моделі.

Останнє, що ми визначаємо в моделі МНК – це критерій Дарбіна-Уотсона – статистичний критерій, що використовується для тестування автокореляції першого порядку елементів послідовності [26].

РОЗДІЛ 2

ПОБУДОВА ЕКОНОМЕТРИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ЖИТЛО ЗІ СТОРОНИ ПОПИТУ

2.1 Опис інформаційної бази моделювання

Багато факторів впливають на вартість нерухомості, тому важливо визначити інформаційну базу, щоб досліджувати залежність різних чинників до ціни на житлову нерухомість [27].

У даній бакалаврській роботі використовуються два набори даних, які можна умовно розділити на дві категорії. Перший набір містить фактори, пов'язані з попитом, тоді як другий набір містить фактори, пов'язані з пропозицією. У двох датасетах залежна змінна залишається незмінною – ціна на житло.

Перший датасет має 12 незалежних змінних та включає в себе 533 спостережень. Незалежні змінні в більшості описують чинники, які впливають на житло зі сторони попиту (споживача), а особливо характеристик самого житла.

Простий, але складний проект, щоб оцінити ціну житла на основі певних факторів, таких як площа будинку, спальні, меблі, близькість до головної дороги тощо. У цьому датасеті знаходяться дані, які були зібрані в північно-східних регіонах США, такі як Нью-Йорк, Пенсильванія, Массачусетс.

Маємо наступні змінні: Price – ціна будинку (\$), Area – площа житла (m^2), Bathrooms – кількість ванних кімнат, Parking – кількість парковочних місць. Також присутні бінарні (фіктивні) змінні, які приймають значення 1 або 0: Mainroad – 1, якщо будинок знаходиться біля головної дороги, Guestoom – 1, якщо наявна вітальня, Basement – 1, якщо наявний підвал, Hotwaterheating – 1, якщо наявна гаряча вода,

Airconditioning – 1, якщо наявний кондиціонер, Prefarea – 1, якщо місцерозташування/район є привабливим. Також маємо категоріальні (факторні змінні): Bedrooms – кількість спальних кімнат в будинку: Bedrooms1 – однокімнатне житло, Bedrooms2 – двокімнатне житло, Bedrooms3 – трикімнатне житло, Bedrooms4 – чотирьохкімнатне житло, Bedrooms5 – п'ятикімнатне житло, Bedrooms6 – шестикімнатне житло, Stories – кількість поверхів будинку: Stories1 – одноповерхове житло, Stories2 – двоповерхове житло, Stories3 – триповерхове житло, Stories4 – чотирьохповерхове житло, Furnishingstatus – меблювання: Furnishingstatus1 – немебльована, Furnishingstatus2 – частково мебльована, Furnishingstatus3 – повністю мебльована.

Узагальнена інформація, щодо змінних моделі, знаходиться в табл. 2.1.

Таблиця 2.1 – Дані датасету [28]

Змінна (позначення)	Опис	Одиниця виміру
Price (y)	Ціна будинку (\$)	\$
Area (x1)	Площа (м ²)	м ²
Bedrooms (x2)	Кількість спальних кімнат	N/A
Bathrooms (x3)	Кількість ванних кімнат	N/A
Stories (x4)	Кількість поверхів	N/A
Mainroad (x5)	Місцерозташування біля головної дороги	T/F
Guestoom (x6)	Наявність вітальні	T/F
Basement (x7)	Наявність підвалу	T/F
Hotwaterheating (x8)	Наявність гарячої води	T/F
Airconditioning (x9)	Наявність кондиціонера	T/F
Parking (x10)	Кількість парковочних місць	N/A
Prefarea (x11)	Привабливе місцерозташування	T/F
Furnishingstatus (x12)	Меблювання	N/A

Джерело: розроблено автором на основі [28]

Варто зауважити, що для більш якісного аналізу було виключено певні значення для змінної bedrooms (5 та 6 спалень), адже вони дуже мало разів зустрічаються в нашому наборі даних.

На рис. 2.1 зображено перші 15 спостережень по деяких змінних.

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating
1	10150000	16200	5	3	2	1	0	0	0
2	9800000	5750	3	2	4	1	1	0	0
3	9100000	6600	4	2	2	1	1	1	0
4	8680000	7155	3	2	1	1	1	1	0
5	8400000	7950	5	2	2	1	0	1	1
6	8400000	7475	3	2	4	1	0	0	0
7	8190000	5960	3	3	2	1	1	1	0
8	7962500	6000	3	1	4	1	1	0	0
9	7700000	6480	3	2	4	1	0	0	0
10	7455000	4300	3	2	2	1	0	1	0
11	7420000	7440	3	2	4	1	0	0	0
12	7420000	6325	3	1	4	1	0	0	0
13	7350000	6000	3	1	2	1	0	0	0
14	6895000	7700	3	2	1	1	0	0	0
15	6755000	6000	4	2	4	1	0	0	0

Рисунок 2.1 – Фрагмент набору даних

Джерело: розроблено автором на основі [28]

Необхідно звести данні, які дадуть уявлення мінімальних, медіанних, середніх і максимальних значень кожної змінної (рис. 2.2). Ми можемо зауважити наприклад, якщо мінімальне значення спостережуваної ціни дорівнює нулю, це може бути невірними даними.

price	area	bedrooms	bathrooms	stories	mainroad
Min. : 1750000	Min. : 153.3	Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 0.000
1st Qu.: 3430000	1st Qu.: 334.4	1st Qu.: 2.000	1st Qu.: 1.000	1st Qu.: 1.000	1st Qu.: 1.000
Median : 4340000	Median : 423.6	Median : 3.000	Median : 1.000	Median : 2.000	Median : 1.000
Mean : 4746879	Mean : 476.9	Mean : 2.916	Mean : 1.276	Mean : 1.801	Mean : 0.863
3rd Qu.: 5740000	3rd Qu.: 590.9	3rd Qu.: 3.000	3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 1.000
Max. : 13300000	Max. : 1449.3	Max. : 4.000	Max. : 4.000	Max. : 4.000	Max. : 1.000
guestroom	basement	hotwaterheating	airconditioning	parking	prefarea
Min. : 0.0000	Min. : 0.0000	Min. : 0.00000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000
1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.00000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000
Median : 0.0000	Median : 0.0000	Median : 0.00000	Median : 0.0000	Median : 0.0000	Median : 0.0000
Mean : 0.1782	Mean : 0.3471	Mean : 0.04315	Mean : 0.3171	Mean : 0.6961	Mean : 0.2383
3rd Qu.: 0.0000	3rd Qu.: 1.0000	3rd Qu.: 0.00000	3rd Qu.: 1.0000	3rd Qu.: 1.0000	3rd Qu.: 0.0000
Max. : 1.0000	Max. : 1.0000	Max. : 1.00000	Max. : 1.0000	Max. : 3.0000	Max. : 1.0000
furnishingstatus					
Length:533					
Class :character					
Mode :character					

Рисунок 2.2 – Статистичні показники для кожної змінної в наборі даних

Джерело: розроблено автором на основі [28]

Максимальне значення спостережуваної ціни статично знижується від значень нормального розподілу та 3-го квантиля, це можна визначити як викид. Також маємо порівняно великі розмахи даних для змінних *price* та *area*, тому варто спробувати прологарифмувати ці змінні для покращення моделювання ціни на житло.

2.2 Аналіз змінних

Перед тим як будувати застосовувати техніки економетричного моделювання дуже важливим є попередній аналіз даних та визначення кореляційних коефіцієнтів [29].

Щодо нашого першого набору даних, то ми маємо одну змінну, де можна побачити кореляцію – площа.

За попереднім аналізом було визначено, що змінна *area* має коефіцієнт кореляції - 0,57.

Перейдемо до аналізу категоріальних змінних, адже вони досліджуються за допомогою Бокс-плоту та подивимося на графічний метод візуалізації розподілу даних, який показує мінімальне та максимальне значення, нижній та верхній кuartилі, медіану та викиди.

Перш за все поглянемо на однокімнатне, двокімнатне, трикімнатне, чотирьохкімнатне житло на рис. 2.3.

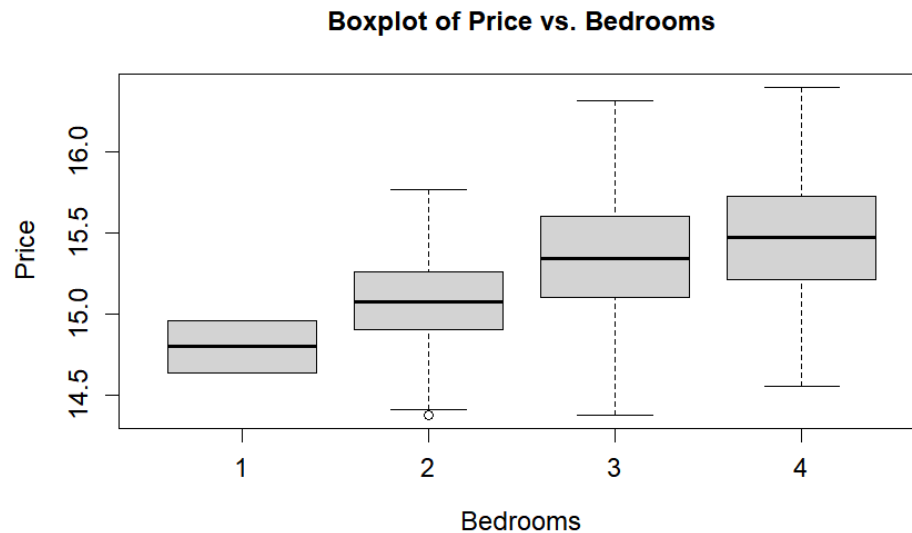


Рисунок 2.3 – Boxplot змінної Bedrooms

Джерело: розроблено автором на основі [28]

Видно чітке зростання усіх «ящиків» в залежності від кількості кімнат в будинку. Зростання є позитивним: від меншого значення до більшого. Медіана позначає середину даних і відображається лінією, яка розділяє рамку на дві частини (іноді відома як другий кuartиль) також з кожним наступним значенням зростає, що говорить про те, що змінна bedrooms добре корелює з ціною на житло.

Дослідимо змінну bathrooms зображену на рис. 2.4.

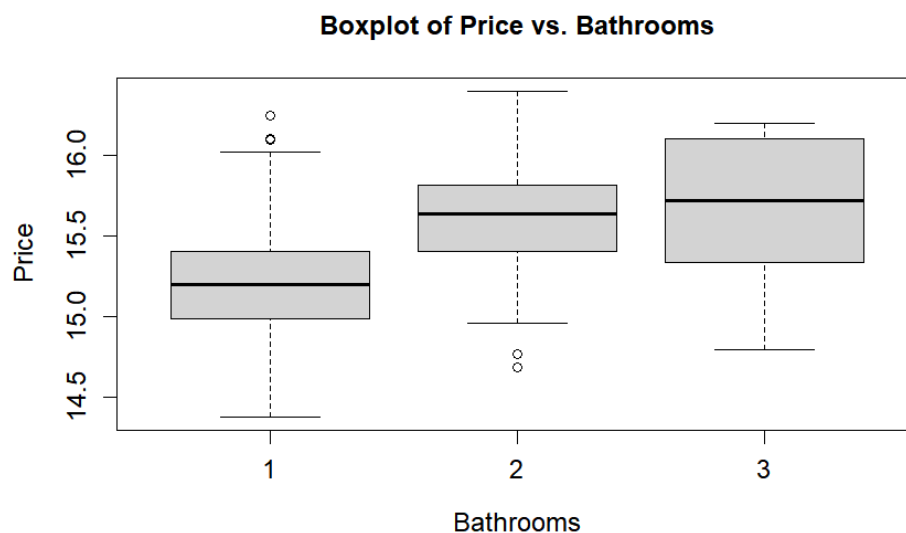


Рисунок 2.4 – Boxplot змінної Bathrooms

Джерело: розроблено автором на основі [28]

Можна спостерігати чітку різницю в кореляції між будинками з 1 та 2 ванними кімнатами. Проте наявність третьої ванної кімнати вже не має такого важливого значення, адже медіани між двома та трьома ванними кімнатами знаходяться майже на одному й тому самому рівні.

Поглянемо на змінну – stories, яка відповідає кількості поверхів в будинку на рис. 2.5.

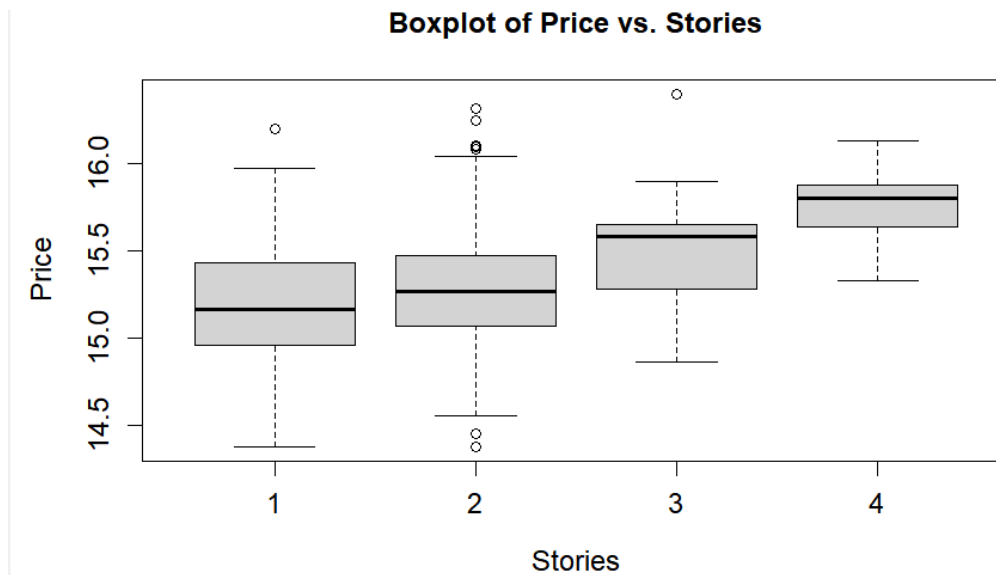


Рисунок 2.5 – Boxplot змінної Stories

Джерело: розроблено автором на основі [28]

Проаналізувавши графік, бачимо чітко виражений приріст в ціні в залежності від кількості поверхів, тому можемо впевнено стверджувати, що кореляція між кількістю поверхів та ціною в даному наборі даних присутня. Також можемо бачити викиди, які проаналізуємо пізніше.

Поглянемо на графічний розподіл даних змінної – parking, яка відповідає кількості місць для паркування на рис. 2.6.

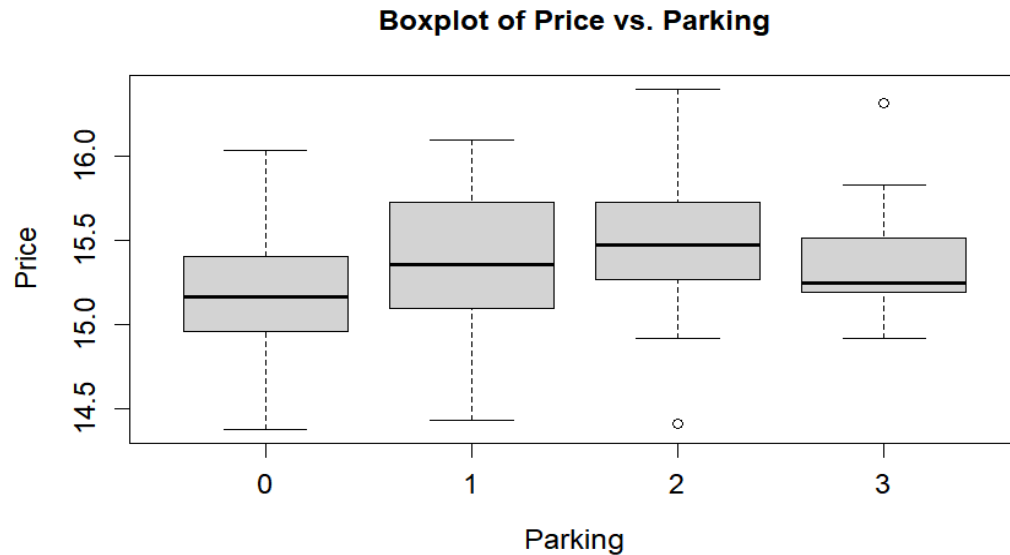


Рисунок 2.6 – Boxplot змінної Parking

Джерело: розроблено автором на основі [28]

Спостерігаємо низьку залежність ціни від кількості парковочних місць, що говорить про дуже малу кореляцію між змінною parking та price. Поглянувши більш детально, можемо побачити, що наявність трьох парковочних місць порівняно з двома, і навіть з одним має від'ємну кореляції, що інтерпретується як: більша кількість парковочних місць не збільшує ціну на житло.

Поглянемо на залежність ціни від наявності меблів у житлі на рис. 2.7.

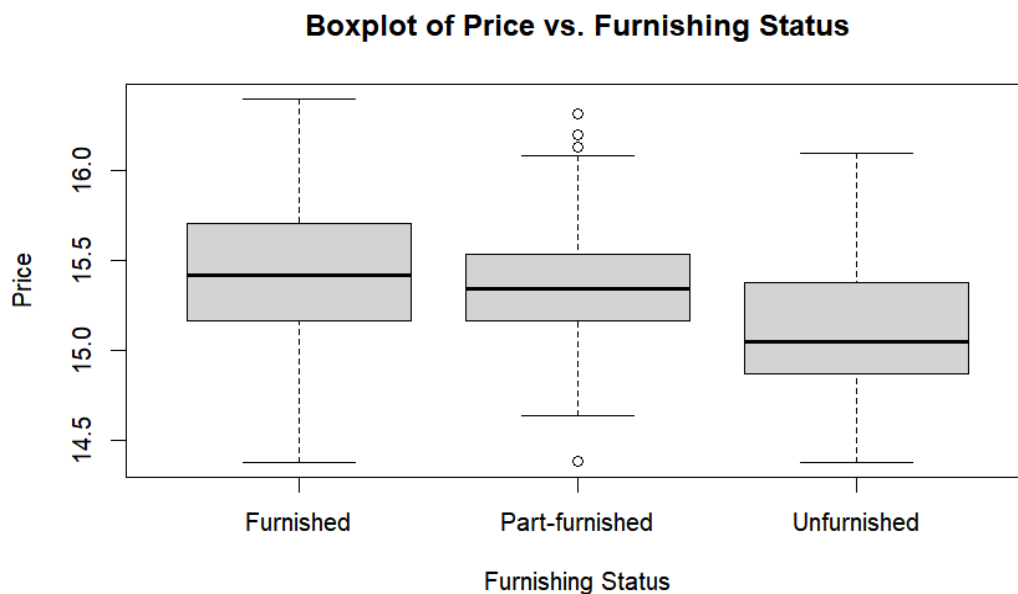


Рисунок 2.7 – Boxplot змінної Furnishing Status

Джерело: розроблено автором на основі [28]

Бачимо, що житло в якому меблі відсутні має найменшу ціну порівняно з замезбльованою та напівзамезбльованою домівкою. Важливо зазначити, що кореляція між замезбльованою та напівзамезбльованою домівкою має певну різницю, але вона доволі така мала, адже верхні кuartилі різняться, проте медіана однаковою.

Включення фіктивних змінних в модель розширює коло питань, що підлягають вирішенню на етапі специфікації моделі, тому важливо розуміти їх вплив на ціну житла. На рис. 2.8 зображені бокс-плоти фіктивних змінних.

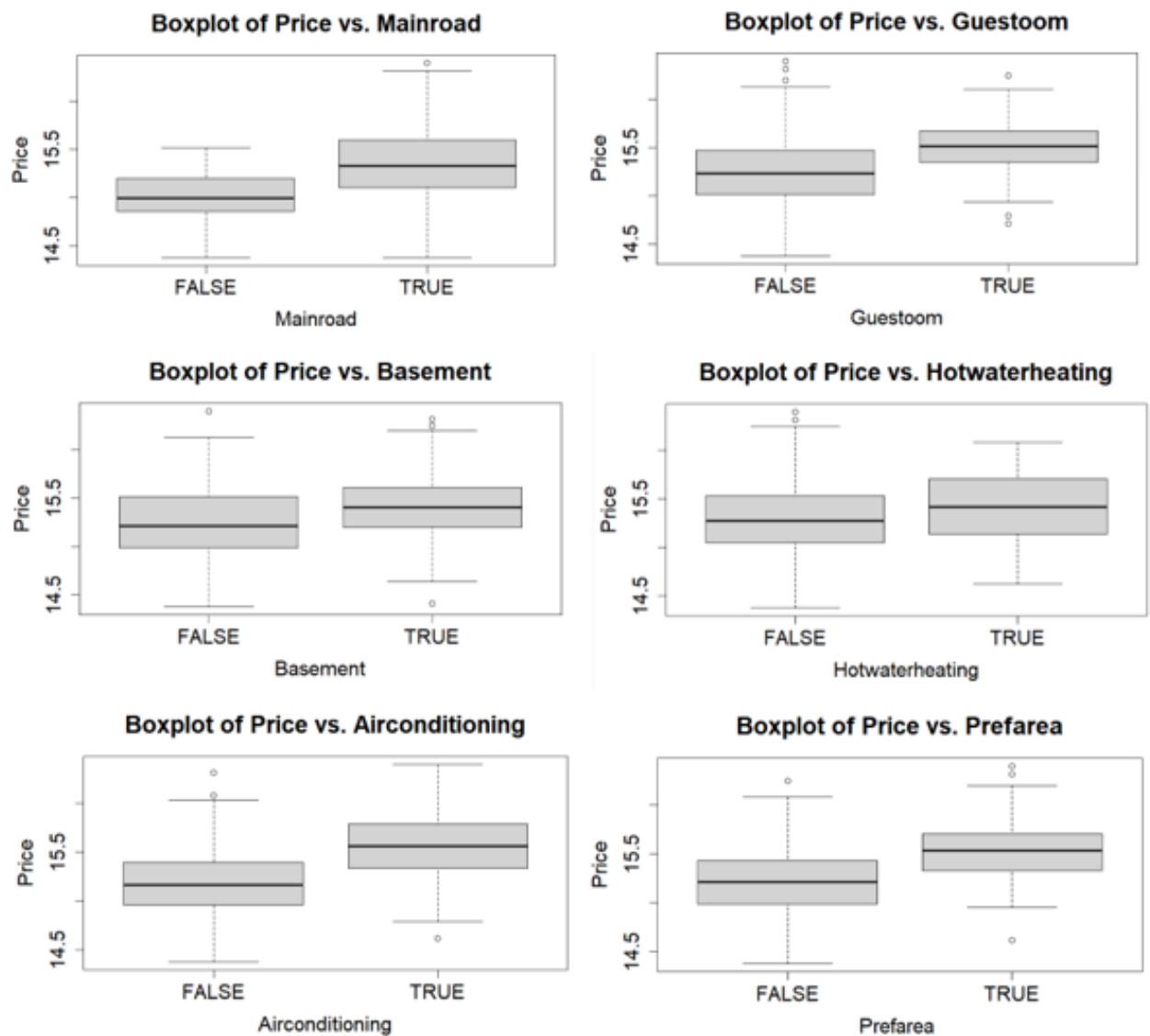


Рисунок 2.8 – Боксplot змінної mainroad, guestroom, basement, hotwaterheating, airconditioning та prefarea

Джерело: розроблено автором на основі [28]

Проаналізувавши всі графічні розподіли фіктивних змінних можна помітити спільну динаміку: наявність будь якої фіктивної змінної позитивно впливає на ціну житла, тобто збільшує її.

Якщо характеризувати все по категоріальним змінним, то можна простежити, що змінні basement та hotwaterheating мають невелику різницю між значеннями «0» та «1», що говорить про низьку кореляції з ціною житла.

Щодо інших змінних mainroad, guestroom, airconditioning та prefarea, то вони мають більший вплив на ціну житла, що говорить про те, що кореляція в них більша ніж в попередніх basement та hotwaterheating.

Також був проведений загальний аналіз впливу всіх змінних на ціну житла та виявлено гістограму залежностей на рис. 2.9, де включено лише найвпливовіші змінних.

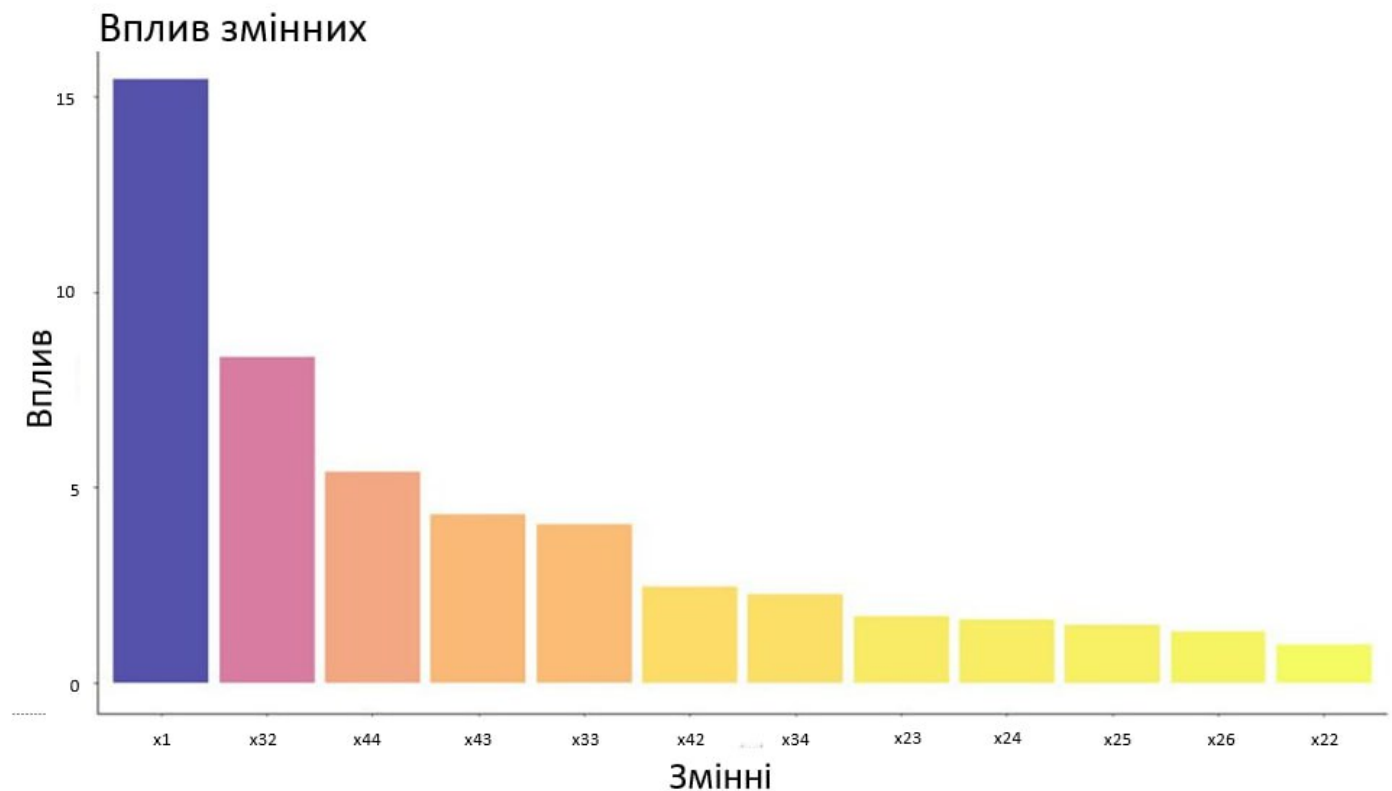


Рисунок 2.9 – Гістограма залежностей

Джерело: розроблено автором на основі [28]

Можемо стверджувати, що площа має найбільший вплив, що відповідає здорову глузду. Більша площа житла зазвичай означає більше простору для життя, що може бути більш привабливим для покупців та збільшувати попит на нерухомість. Це може призводити до збільшення ціни на нерухомість.

Більша площа житла може також вказувати на більш високу якість житла або кращі умови проживання, такі як краще оснащення, більше вікон та більше світла. Ці фактори також можуть підвищити ціну на нерухомість.

Крім того, більша площа житла може вказувати на більш високий рівень життя, що може бути пов'язано з більшими доходами та здатністю покупців платити більше за нерухомість.

Бачимо, що друге та п'яте місце посідають кількість ванних кімнат, адже вони можуть слугувати індикатором розкішності житла. Наприклад, великий і дорогий будинок може мати більше ванних кімнат, ніж менший і скромний будинок. У такому випадку, ціна буде вищою внаслідок розкішності житла.

По-друге, кількість ванних кімнат може впливати на зручність проживання. Наприклад, велика сім'я може бути зацікавлена в будинку з більшою кількістю ванних кімнат, щоб уникнути черг на ранок. У такому випадку, більша кількість ванних кімнат може збільшити ціну житла, оскільки забезпечує зручність і комфорт для мешканців.

Наступними найбільш впливовими змінними є кількість поверхів та кількість спальних кімнат.

Більші будинки зазвичай мають більше поверхів. Це може пояснити подібність в значеннях кореляції. Також це можуть бути важливими факторами при визначенні ціни на будинок, адже багато людей шукають будинок з певною кількістю поверхів, що залежить від стандартів розкішності та величини сім'ї.

2.3 Оцінювання параметрів моделі та її статистичну значущість

Економетрична модель є результатом поєднання теоретичного та емпіричного аналізу. Теоретичний аналіз полягає у вивченні взаємозв'язків між економічними змінними, які відображаються в специфікації моделі. Емпірична інформація використовується для перевірки теоретичних припущень та отримання конкретних числових результатів [30].

Перед побудовою специфікації моделі, важливо визначитися чи потрібно логарифмувати змінні price та area. Для цього нам потрібно проаналізувати за яких умов моделі будуть кращими – застосовуючи логарифмування чи залишаючи змінні в початкову вигляді, без будь-яких змін.

Побудуємо дві моделі m0 та m1 та поглянемо на значущість змінної та скоригований коефіцієнт детермінації, результати аналізу продемонстровані на рис. 2.10 та 2.11.

```
Call:
lm(formula = df$price ~ df$area)

Residuals:
    Min       1Q   Median       3Q      Max
-4813060 -1036796  -216468   685842  7509178

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2402878.5   179878.3   13.36  <2e-16 ***
df$area       4914.8      348.6    14.10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1585000 on 531 degrees of freedom
Multiple R-squared:  0.2723,    Adjusted R-squared:  0.271
F-statistic: 198.7 on 1 and 531 DF,  p-value: < 2.2e-16
```

Рисунок 2.10 – Результати функції summary() без логарифмування

Джерело: розроблено автором на основі [28]

```

Call:
lm(formula = log(df$price) ~ log(df$area))

Residuals:
    Min       1Q   Median       3Q      Max
-0.85292 -0.19426  0.00573  0.19221  0.89987

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.01881    0.20343   59.08  <2e-16 ***
log(df$area)   0.53946    0.03334   16.18  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.303 on 531 degrees of freedom
Multiple R-squared:  0.3302,    Adjusted R-squared:  0.329
F-statistic: 261.8 on 1 and 531 DF,  p-value: < 2.2e-16

```

Рисунок 2.11 – Результати функції summary() з логарифмуванням

Джерело: розроблено автором на основі [28]

За результатами даного дослідження було визначено, що значущість змінної зберігається при логарифмуванні, проте скоригований коефіцієнт детермінації R-squared суттєво збільшився та показав, що логарифмування додає кращої оцінки моделі.

При побудові специфікації моделі важливо враховувати та правильно позначити всі незалежні змінні в залежності від типу даної змінною (фіктивні, категоріальні тощо).

Набір даних містить 4 числові змінні (Price, Area, Bathrooms, Parking), 3 категоріальні змінні (Bedrooms, Stories, Furnishingstatus) та 6 фіктивних змінних (Mainroad, Guestoom, Basement, Hotwaterheating, Airconditioning, Preferarea). Отже специфікація виглядатиме наступним чином (2.1)

$$\begin{aligned}
 \text{Log}(y) = & b_0 + b_1 \text{Log}(x_1) + b_2 x_{22} + b_3 x_{23} + b_4 x_{24} + b_8 x_{32} + b_9 x_{33} + b_{10} x_{42} + \\
 & b_{11} x_{43} + b_{12} x_{44} + b_{13} x_5 + b_{14} x_6 + b_{15} x_7 + b_{16} x_8 + b_{17} x_9 + b_{18} x_{102} + b_{19} x_{103} + b_{20} x_{11} \\
 & + b_{21} x_{121} + b_{22} x_{122} + b_{23} x_{123} + \varepsilon
 \end{aligned}
 \tag{2.1}$$

Далі перейдемо до визначення параметрів та побудови моделі в середовищі RStudio з мовою програмування R. У середовищі програмування, попередньо були завантажені необхідні для роботи бібліотеки, а також самі дані.

У середовищі R правильний підхід до визначення програмою логарифмованих, фіктивних та категоріальний змінних було визначено наступним чином (рис. 2.12):

```
y<- log(df$price)
x1<- log(df$area)
x2<- as.factor(df$bedrooms)
x3<- as.factor(df$bathrooms)
x4<- as.factor(df$stories)
x5<- as.logical(df$mainroad)
x6<- as.logical(df$guestroom)
x7<- as.logical(df$basement)
x8<- as.logical(df$hotwaterheating)
x9<- as.logical(df$airconditioning)
x10<- as.factor(df$parking)
x11<- as.logical(df$prefarea)
x12<- as.factor(df$furnishingstatus)
```

Рисунок 2.12 – Переведення змінних до іншого типу даних

Джерело: розроблено автором на основі [28]

Для кращого розуміння даних, важливо проаналізувати їх структуру, яка ілюструється на рисунку 2.13. Набір даних містить 2 числові, 5 категоріальних та 6 фіктивних змінних.

x1	num [1:532] 6.54 6.83 6.55 6.54 6.55 ...
x10	Factor w/ 4 levels "0","1","2","3": 3 3 4 3 3 3 3...
x11	logi [1:532] TRUE TRUE TRUE FALSE TRUE TRUE ...
x12	Factor w/ 3 levels "Furnished","Part-furnished",...
x2	Factor w/ 4 levels "1","2","3","4": 4 3 4 4 3 4 4...
x3	Factor w/ 3 levels "1","2","3": 2 2 2 1 3 3 1 2 1...
x4	Factor w/ 4 levels "1","2","3","4": 3 2 2 2 1 4 2...
x5	logi [1:532] TRUE TRUE TRUE TRUE TRUE TRUE ...
x6	logi [1:532] FALSE FALSE FALSE TRUE FALSE FALSE
x7	logi [1:532] FALSE TRUE TRUE TRUE TRUE TRUE ...
x8	logi [1:532] FALSE FALSE FALSE FALSE FALSE FALSE ...
x9	logi [1:532] TRUE FALSE TRUE TRUE TRUE TRUE ...
y	num [1:532] 16.4 16.3 16.3 16.3 16.2 ...

Рисунок 2.13 – Структура даних

Джерело: розроблено автором на основі [28]

Починаємо аналіз моделей (m1, m3, m5, m7, m9, m11, m13, m15, m17, m19, m21, m23), які різняться між собою з постійним додаванням нового предиктора та видалення незначущих змінних.

На рис. 2.14 відображені результати моделі з додаванням усіх предикторів

```

Coefficients:
(Intercept)      13.17360    0.21114   62.392 < 2e-16 ***
x1                0.28041    0.02700   10.386 < 2e-16 ***
x22              0.05191    0.14393    0.361 0.718496
x23              0.12669    0.14444    0.877 0.380838
x24              0.12575    0.14643    0.859 0.390868
x32              0.16151    0.02338    6.909 1.46e-11 ***
x33              0.29798    0.07380    4.038 6.22e-05 ***
x42              0.05297    0.02311    2.292 0.022301 *
x43              0.17800    0.03974    4.479 9.24e-06 ***
x44              0.25971    0.04093    6.346 4.88e-10 ***
x5TRUE           0.09539    0.02778    3.433 0.000645 ***
x6TRUE           0.05266    0.02525    2.085 0.037528 *
x7TRUE           0.09153    0.02145    4.266 2.37e-05 ***
x8TRUE           0.14304    0.04415    3.240 0.001275 **
x9TRUE           0.15784    0.02100    7.517 2.54e-13 ***
x101             0.07574    0.02286    3.313 0.000987 ***
x102             0.10491    0.02490    4.212 2.99e-05 ***
x103            -0.02054    0.06348   -0.324 0.746430
x11TRUE          0.12636    0.02233    5.659 2.54e-08 ***
x12Part-furnished 0.01471    0.02247    0.655 0.512906
x12Unfurnished   -0.11746    0.02434   -4.826 1.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.201 on 511 degrees of freedom
Multiple R-squared:  0.7123,    Adjusted R-squared:  0.701
F-statistic: 63.25 on 20 and 511 DF,  p-value: < 2.2e-16

```

Рисунок 2.14 – Результати моделі з додаванням усіх предикторів

Джерело: розроблено автором на основі [28]

Критерій Стюдента говорить про те, що залежна змінна у є статистично значущою і доволі таки сильно, бо t-value є доволі таки великим значенням, x1, x32, x33, x42, x43, x44, x5TRUE, x6TRUE, x7TRUE, x8TRUE, x9TRUE, x101, x102, x11TRUE, x12Unfurnished – також статично значущі (ймовірність того, що один з параметрів моделі рівний нулю доволі низька). Останній стовпчик говорить про те, що даним коефіцієнтам можна довіряти, адже значення $\text{Pr}(>|t|)$ має доволі такі маленькі значення, а отже є меншими ніж 0.05. Довіряти даним аналізу предиктора та залежної змінної моделі можна, адже підтвердженням тому є позначення «***», «**», та «*».

Змінні x22, x23, x24, x103 та x12Unfurnished є статично незначущими, що означає, що ці параметри можуть дорівнювати нулю, тому потрібно виключити їх з даної моделі для покращення.

Multiple R-squared показує, що ціна житла пояснюється незалежними змінними в моделі на 71.23% , а Adjusted R-squared, враховуючи предиктори, показує, що ціна житла пояснюється незалежними змінними в моделі на 70,1%. F-statistic продемонстрував, що можна адекватно аналізувати дані.

Потрібно створити нову модель, яка включатиме лише значущі змінні. На рис. 2.15 зображена реалізація в середовищі RStudio:

```
m25 <- lm(y ~ x1 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x11 + x13[, unfurnished_col] +
parking_dummies[, parking_1] + parking_dummies[, parking_2])
```

Рисунок 2.15 – Реалізація моделі в середовищі RStudio

Джерело: розроблено автором на основі [28]

На рис. 2.16 відображені результати даної моделі:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.20996    0.15624   84.549 < 2e-16 ***
x1              0.29042    0.02647   10.970 < 2e-16 ***
x32             0.16731    0.02283    7.329 9.00e-13 ***
x33             0.30703    0.07357    4.173 3.53e-05 ***
x42             0.08678    0.01988    4.366 1.53e-05 ***
x43             0.21664    0.03766    5.753 1.51e-08 ***
x44             0.29141    0.03952    7.374 6.64e-13 ***
x5TRUE          0.08509    0.02760    3.083 0.00216 **
x6TRUE          0.05199    0.02530    2.054 0.04044 *
x7TRUE          0.09977    0.02136    4.671 3.83e-06 ***
x8TRUE          0.14533    0.04432    3.279 0.00111 **
x9TRUE          0.16058    0.02100    7.648 1.00e-13 ***
x11TRUE         0.13319    0.02227    5.981 4.14e-09 ***
x13[, unfurnished_col] -0.12671    0.01955  -6.480 2.14e-10 ***
parking_dummies[, parking_1] 0.07191    0.02263    3.178 0.00157 **
parking_dummies[, parking_2] 0.11338    0.02453    4.621 4.82e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2019 on 516 degrees of freedom
Multiple R-squared:  0.7066,    Adjusted R-squared:  0.6981
F-statistic: 82.87 on 15 and 516 DF,  p-value: < 2.2e-16
```

Рисунок 2.16 – Результати моделі з додаванням ефективних змінних

Джерело: розроблено автором на основі [28]

Критерій Стюдента показує, що дана модель, а особливо всі змінні є значущими, що дуже добре.

Multiple R-squared показує, що ціна житла пояснюється незалежними змінними в моделі на 70,66%, а Adjusted R-squared, враховуючи предиктори, показує, що ціна житла пояснюється незалежними змінними в моделі на 69,81%. F-statistic продемонстрував, що можна адекватно аналізувати дані.

Результати оцінки рівняння для першого датасету наведені в табл. 2.2.

Таблиця 2.2 – результати оцінки рівняння [28]

Змінна	Коефіцієнт	Стандартна похибка	t-статистика	P-значення	Pr(> t)
Const	13.20996	0.15624	84.549	<2.00E-16	***
Area (x1)	0.29042	0.02647	10.97	<2.00E-16	***
Bathrooms2 (x32)	0.16731	0.02283	7.329	9.00E-13	***
Bathrooms3 (x33)	0.30703	0.07357	4.173	3.53E-05	***
Stories2 (42)	0.08678	0.01988	4.366	1.53E-05	***
Stories3 (43)	0.21664	0.03766	5.753	1.51E-08	***
Stories4 (44)	0.29141	0.03952	7.374	6.64E-13	***
Mainroad (x5)	0.08509	0.0276	3.083	0.00216	**
Guestoom (x6)	0.05199	0.0253	2.054	0.04044	*
Basement (x7)	0.09977	0.02136	4.671	3.83E-06	***
Hotwaterheating (x8)	0.14533	0.04432	3.279	1.11E-03	**
Airconditioning (x9)	0.16058	0.021	7.648	1.00E-13	***
Prefarea (x11)	0.13319	0.02227	5.981	4.14E-09	***
Unfurnished (x123)	-0.12671	0.01955	-6.48	2.14E-10	***
Parking2 (x102)	0.07191	0.02263	3.178	0.00157	**
Parking3 (x103)	0.11338	0.02453	4.62E+00	4.82E-06	***

Джерело: розроблено автором на основі [28]

Останній стовпчик "знаки значимості" (significance codes), використовується у статистичних відомостях для відображення рівня статистичної значимості коефіцієнтів у регресійних моделях:

**** – означає дуже високу статистичну значущість (p-value < 0.001)

*** – означає високу статистичну значущість (p-value < 0.01)

** – означає помірну статистичну значущість (p-value < 0.05)

Отримавши коефіцієнти для специфікації моделі, можна вже її будувати, але краще перевірити модель на ряд гіпотез.

2.4 Тестування гіпотез автокореляції та гетероскедастичності залишків

Після побудови моделей та специфікації, важливо переконатися в тому, що модель є «адекватною». Окрім описаних дій потрібно провести дослідження на напруженість, викиди, залишки, впливові спостереження, мультиколінеарність та автокореляцію і т.д.

Почнемо з аналізу залишків, що зображений на рис. 2.17.

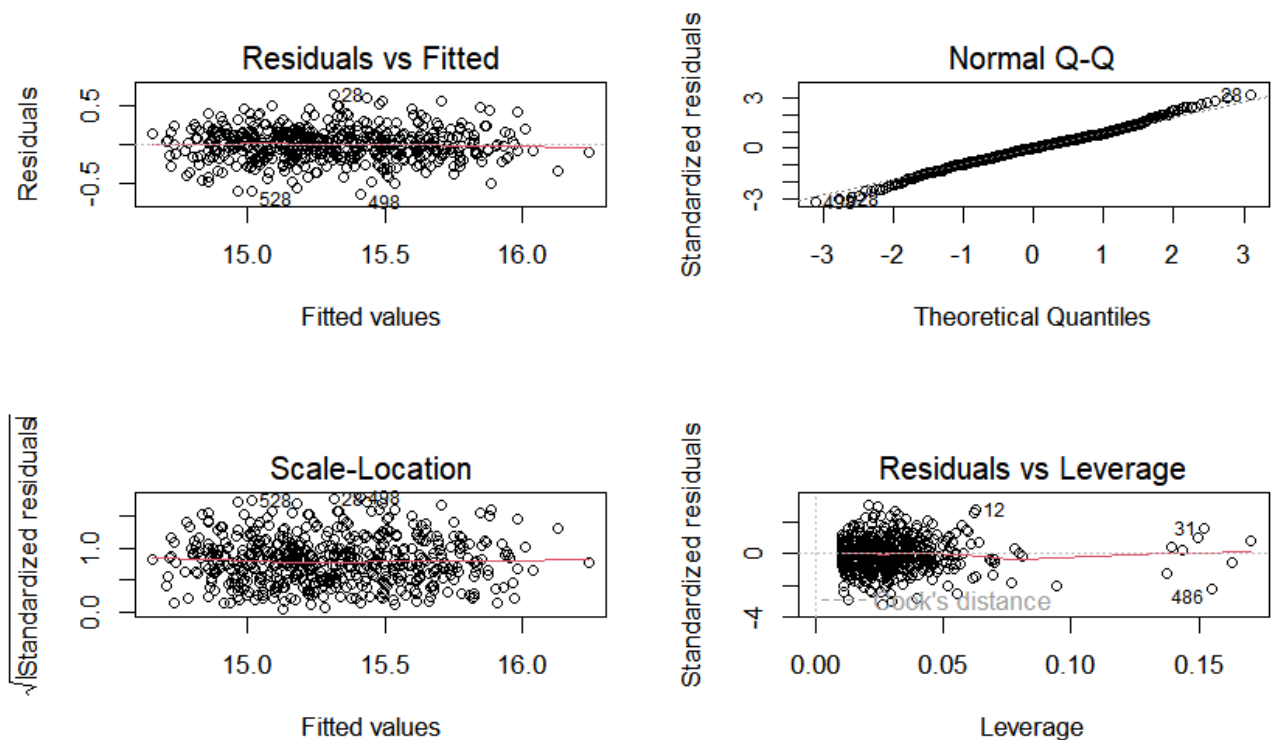


Рисунок 2.17 – Аналіз залишків

Джерело: розроблено автором на основі [28]

Residuals vs Fitted – точки розкидані по прямій лінії та навколо неї, графік розсіювання залишків на осі y та встановлених значень на осі x. У даному випадку розподіл залишків є доволі таки з «гарною поведінкою»

Перевірка на нормальність (Q-Q plot) – всі точки знаходяться більш-менш на одній лінії під кутом 30, що вказує на нормальність залишків.

Scale-Location має рівномірний розкид точок по горизонталі, це може свідчити про те, що дисперсія залишків є сталою і не залежить від значень прогнозованих змінних. Це вказує на те, що у моделі немає систематичної взаємодії між змінними, і модель може бути корисною для прогнозування значень відгуку для різних значень прогнозованих змінних.

Спостерігаємо, що точки заходяться в центрі по вертикалі та зліва по горизонталі для графіка Residuals vs Leverage, що це свідчить про те, що немає впливових спостережень (outliers), які могли би значно вплинути на оцінки регресійних коефіцієнтів. В цьому випадку модель може бути надійною та точно передбачати значення залежної змінної. Спостерігається декілька спостережень (точок), які знаходяться за пунктирною лінією, що говорить про наявність точок, які суттєво впливають на модель.

Поглянемо на розподіл залишків рис. 2.18.

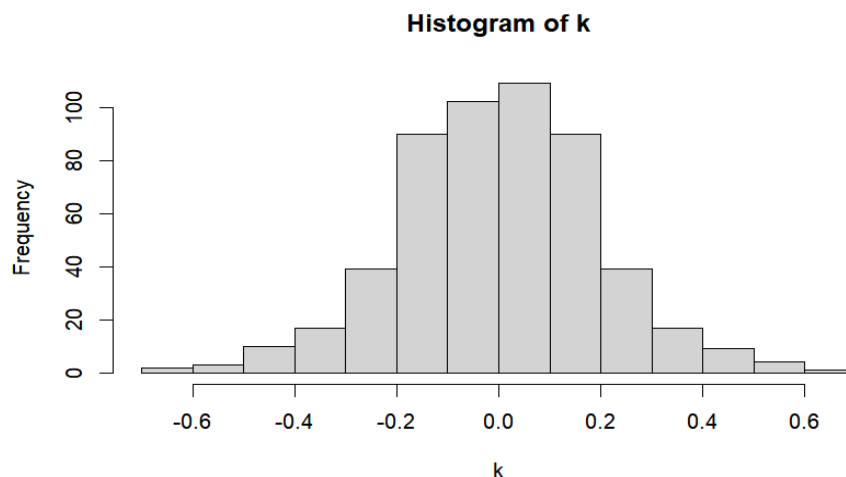


Рисунок 2.18 – Нормальність розподілу залишків

Джерело: розроблено автором на основі [28]

Маємо нормальний розподіл залишків, що є однією з передумов для правильної оцінки параметрів моделі та проведення статистичних тестів на значимість. Якщо розподіл залишків є нормальним, то це означає, що модель добре описує взаємозв'язок між залежною та незалежними змінними, і наші оцінки параметрів будуть точними.

Проведемо `OutlierTest()` для визначення викидів. За ним ми можемо визначити, що значення з найбільшим відхиленням до моделі є рядок 498 – це є викидом. Також отримали підтвердження того, що більше відхилень немає, які є більшими ніж відстань Кука, що є доволі таки позитивним для нашої ефективності економетричної моделі.

Визначимо ряд показників: відстань Кука, `DFBETAS`, `DFFITS`, коефіцієнт варіації, значення матриці проекції на простір регресорів, які можемо спостерігати на рис. 2.19.

	dfb.1_ <dbl>	dfb.x1 <dbl>	dfb.x32 <dbl>	dfb.x33 <dbl>	dfb.x42 <dbl>	dfb.x43 <dbl>
1	-2.626660e-02	2.633222e-02	1.308824e-01	3.374317e-03	-7.436293e-03	2.676672e-01
2	-1.831682e-01	1.789154e-01	1.462146e-01	-3.232092e-02	1.318407e-01	1.024562e-02
3	-1.244987e-01	1.181660e-01	1.582171e-01	4.081745e-03	9.650409e-02	-6.047871e-02
4	-1.044855e-01	9.306238e-02	-1.624477e-01	-8.047982e-02	1.675444e-01	5.111914e-02
5	-6.827672e-03	6.297224e-03	1.072511e-03	3.879918e-01	-8.299875e-02	-4.137591e-02
6	6.843103e-03	-7.379324e-03	2.880678e-02	-2.169385e-01	1.082700e-02	7.618573e-03
7	-4.020633e-02	3.559010e-02	-7.610164e-02	-3.532040e-02	7.703932e-02	1.401692e-03
8	4.209196e-02	-5.149931e-02	3.548124e-02	-4.170592e-03	-8.009845e-03	-2.379979e-02
9	-8.086476e-02	7.908832e-02	-4.513318e-02	-1.811868e-02	5.579621e-02	4.636089e-03
10	-2.330128e-02	1.681463e-02	-1.308157e-02	2.885962e-01	1.438072e-02	3.931929e-03

1-10 of 532 rows | 1-9 of 21 columns

[Previous](#)

Рисунок 2.19 – Перевірка показників «адекватності» моделі

Джерело: розроблено автором на основі [28]

Бачимо, що деякі спостереження є не типовими, вони помічені зірочками. Їхня кількість = 39.

Перевіримо нашу модель: чи виконуються вимоги, що повинні бути у лінійних моделях (рис. 2.20).

	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	7.084519897	0.13148858	Assumptions acceptable.
Skewness	0.059526720	0.80724555	Assumptions acceptable.
Kurtosis	6.020041841	0.01414431	Assumptions NOT satisfied!
Link Function	1.003010387	0.31658318	Assumptions acceptable.
Heteroscedasticity	0.001940949	0.96485961	Assumptions acceptable.

Рисунок 2.20 – Перевірка моделі на вимоги, що повинні бути у лінійних моделях

Джерело: розроблено автором на основі [28]

Global Stat є загальним описом статистичної значущості залежності між змінними. Якщо ця статистика низька, то зв'язок між змінними може бути слабким або відсутнім. За цими даними робимо висновок, що лінійний зв'язок між x та y відсутній.

Skewness – відображає ступінь асиметрії розподілу даних. У нашому випадку вона відсутня, можна припустити, що розподіл симетричний.

Kurtosis – наскільки "високими" або "низькими" є хвости розподілу порівняно з нормальним розподілом. Даний розподіл містить хвости, де присутні екстремальні значення.

Link Function – використовується для перетворення залежної змінної, щоб її розподіл відповідав вимогам лінійної регресії. Залежна змінна не розподілена лінійно.

Heteroscedasticity – чи є однакова дисперсія в різних групах даних. Перевірку на гетероскедастичність пройшов, що говорить про те, що наші дисперсії є однорідними.

Перевіримо модель на мультиколінеарність за допомогою загальної та індивідуальної перевірки, зображеної на рис 2.21.

Overall Multicollinearity Diagnostics

	MC Results	detection
Determinant $ X'X $:	0.2127	0
Farrar Chi-Square:	812.9050	1
Red Indicator:	0.1275	0
Sum of Lambda Inverse:	18.3698	0
Theil's Method:	-7.2359	0
Condition Number:	63.2313	1

Рисунок 2.21 – Перевірка мультиколінеарності

Джерело: розроблено автором на основі [28]

Ці результати свідчать про певні проблеми з мультиколінеарністю в моделі.

- Determinant $|X'X|$ значення 0,2127 що свідчить про несуттєву кореляцію між змінними, адже за цим результатом її не виявлено.
- Farrar Chi-Square та Red Indicator мають суперечливі результати, адже один показав її відсутність, інший її наявність, але проаналізувавши її показник, то можна стверджувати, що вона несуттєва.
- Sum of Lambda Inverse та Theil's Method також свідчать про відсутність проблем з мультиколінеарністю.
- Значення Condition Number більше 10, що вказує на проблеми з мультиколінеарністю.

Отже, ці результати показують, що модель має незначні проблеми з мультиколінеарністю. Це може призвести до збільшення стандартних помилок при моделювання та зниження ефективності прогнозування. Проте її вплив є замалим, щоб звертати на неї уваги.

Необхідно це підтвердити за допомогою індивідуального аналізу (індивідуально по змінним) на мультиколінеарність (рис. 2.22)

All Individual Multicollinearity Diagnostics Result

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
x1	1.4122	0.7081	15.2233	16.4261	0.8415	-0.8044	0	0.0192	1.6478
x32	1.2420	0.8052	8.9368	9.6429	0.8973	-0.7074	0	0.0218	1.0999
x33	1.0459	0.9561	1.6938	1.8276	0.9778	-0.5957	0	0.0259	0.2476
x42	1.2620	0.7924	9.6743	10.4386	0.8902	-0.7188	0	0.0215	1.1719
x43	1.2270	0.8150	8.3810	9.0432	0.9028	-0.6989	0	0.0221	1.0442
x44	1.4165	0.7060	15.3817	16.5970	0.8402	-0.8068	0	0.0191	1.6599
x5TRUE	1.1765	0.8500	6.5173	7.0322	0.9219	-0.6701	0	0.0230	0.8468
x6TRUE	1.2252	0.8162	8.3176	8.9747	0.9034	-0.6979	0	0.0221	1.0377
x7TRUE	1.3501	0.7407	12.9273	13.9486	0.8606	-0.7690	0	0.0201	1.4637
x8TRUE	1.0601	0.9433	2.2185	2.3938	0.9713	-0.6038	0	0.0255	0.3199
x9TRUE	1.2424	0.8049	8.9502	9.6573	0.8972	-0.7076	0	0.0218	1.1013
x11TRUE	1.1755	0.8507	6.4821	6.9942	0.9223	-0.6696	0	0.0230	0.8429
x13[, unfurnished_col]	1.0945	0.9137	3.4888	3.7644	0.9559	-0.6234	0	0.0247	0.4873
parking_dummies[, parking_1]	1.1872	0.8423	6.9135	7.4597	0.9178	-0.6762	0	0.0228	0.8902
parking_dummies[, parking_2]	1.2528	0.7982	9.3341	10.0716	0.8934	-0.7136	0	0.0216	1.1390

Рисунок 2.22 – Перевірка мультиколінеарності за допомогою індивідуального аналізу

Джерело: розроблено автором на основі [28]

У даній таблиці представлені результати тестів на мультиколінеарність для кожної змінної у регресійній моделі.

- VIF (variance inflation factor) показує, наскільки змінна корелює з іншими змінними у моделі. Загалом, VIF значення менше 5 вважається прийнятним для регресійної моделі, але якщо наявна висока кореляція між змінними, то значення менше 10 також може бути прийнятним. Тест успішно пройдено
- Також, відсутність мультиколінеарності можна підтвердити тим, що всі значення TOL (tolerance) для кожної змінної більше 0,2, що означає, що кожна змінна має відносно низьку кореляцію з іншими змінними в моделі.
- Wi (Variance Inflation Factor) – це міра, яка показує, наскільки дисперсія коефіцієнта оцінки регресії збільшується в результаті кореляції між змінними. Якщо Wi більше за 10, то це може свідчити про наявність мультиколінеарності.
- Fi – це міра, яка оцінює вплив кожної змінної на оцінку регресії при відсутності інших змінних. Якщо Fi більше за 10, то це може свідчити про наявність мультиколінеарності.

- Leamer – це міра, яка показує, наскільки кожна змінна взаємодіє з іншими змінними в моделі. Якщо Leamer більше за 1, то це може свідчити про наявність мультиколінеарності.
- CVIF: Діагностичний показник, який відображає вплив мультиколінеарності на оцінки параметрів моделі. Якщо CVIF більше 10, то це свідчить про проблему мультиколінеарності, а якщо більше 30 – про серйозні проблеми.
- Klein's Rule: Діагностичний показник, що вимірює, наскільки добре регресійна модель збігається з даними. Якщо найдовший шлях між кожною змінною та всіма іншими змінними менше 2,4, то можна вважати, що модель добре збігається з даними.
- IND1 та IND2: Діагностичні показники, що вимірюють важливість кожної змінної в утворенні моделі. IND1 вимірює відносний вклад кожної змінної в пояснюючу силу моделі, а IND2 – відносний вклад кожної змінної відносно усіх інших змінних. Низьке значення IND1 або IND2 може свідчити про неважливість змінної в моделі, а високе – про її важливість.

Попередньо лише 2 із 6 тестів змогли підтвердити мультиколінеарність і якщо проаналізувати показники, то можна чітко зрозуміти, що вони вже є на грані, яку можна назвати «відсутність мультиколінеарності».

Отже даний рівень мультиколінеарності можна вважати прийнятним, адже більшість тестів та змінних пройшли тести на успішно, що говорить про несуттєву наявність мультиколінеарності в моделі.

```
lag Autocorrelation D-W Statistic p-value
1      0.3677482      1.245081      0
Alternative hypothesis: rho != 0
```

Рисунок 2.23 – Тест Дарбіна-Уотсона

Джерело: розроблено автором на основі [28]

За результатами тесту Дарбіна-Уотсона (рис. 2.23) можна зробити висновок, що існує автокореляція першого порядку, тобто між сусідніми спостереженнями є певний рівень залежності. P-value дорівнює 0, що є досить низьким значенням і свідчить про статистичну значимість результатів. Автокореляція замала, щоб звертати на неї увагу.

2.5 Економічна інтерпретація

Побудуємо специфікацію моделі та дізнаємося, які коефіцієнти стоять при кожній впливовій змінній.

У даному рівнянні множник перед кожною незалежною змінною вказує на те, як змінюється залежна змінна (логарифм ціни на житло) при зміні відповідної змінної, за умови, що інші змінні залишаються незмінними.

Проаналізуємо результати, які було отримано зі специфікації моделі (2.2), розробленої на основі таблиці 2.2:

$$\begin{aligned} \text{Log}(y) = & 13.20996 + 0.29042 * \text{Log}(x1) + 0.16731 * x32 + 0.30703 * x33 + \\ & 0.08678 * x42 + 0.21664 * x43 + 0.29141 * x44 + 0.08509 * x5 + 0.05199 * x6 + 0.09977 * x7 + \\ & 0.14533 * x8 + 0.16058 * x9 + 0.07191 * x102 + 0.11338 * x103 + 0.13319 * x11 + - \\ & 0.12671 * x123 + \varepsilon \end{aligned}$$

(2.2)

Економічна інтерпретація кожної змінної в рівнянні:

- 13.20996 є константою в даному рівнянні і це означає, що якщо всі інші змінні рівняння рівні 0, очікувана (середня) ціна будинку становитиме 13.20996.
- $\text{Log}(x_1)$: коефіцієнт 0.29042 означає, що збільшення площі квартири на 1% призведе до збільшення ціни на 0.29042% (за інших рівних умов).
- x_{32} та x_{33} : коефіцієнти 0.16731 та 0.30703 означають, що наявність 2 або 3 ванних кімнат у квартирі порівняно з однією ванною збільшує ціну відповідно на 16.73% та 30.70%.
- x_{42} , x_{43} та x_{44} : коефіцієнти 0.08678, 0.21664 та 0.29141 означають, що наявність 2, 3 або 4 поверхів у будинку порівняно з одним поверхом збільшує ціну відповідно на 8.68%, 21.66% та 29.14%.
- x_5 , x_6 , x_7 , x_8 , x_9 : коефіцієнти 0.08509, 0.05199, 0.09977, 0.14533 та 0.16058 відповідно означають, що наявність таких характеристик у квартирі порівняно з їх відсутністю збільшує ціну на вказану величину відповідно на 8.51%, 5.20%, 9.98%, 14.53% та 16.06%.
- x_{102} та x_{103} : коефіцієнти 0.07191 та 0.11338 означають, що наявність 2 або 3 місць для паркування порівняно з одним місцем збільшує ціну відповідно на 7.19% та 11.34%.
- x_{11} : коефіцієнт 0.13319 означає, що хороше місцезнаходження житла збільшує ціну на 13.32%.
- x_{123} : коефіцієнт -0.12671 вказує, що за умови відсутності меблів, ціна буде нижчою на 12.67% порівняно з житлом, де меблі наявні.
- ε : це випадкова складова або помилка моделі, яка враховує вплив всіх інших факторів, які не враховані в моделі.

Отримана модель дозволяє зробити висновки про вплив різних факторів на ціни квартир. Зростання площі, наявність додаткових ванних кімнат, наявність більшої кількості поверхів у будівлі, наявність головної дороги поруч, наявність

гостьової кімнати, підвалу, системи гарячої води, кондиціонування, парковки та бажання покупців мати квартиру в певному районі позитивно впливає на ціни квартир, адже мають додатні коефіцієнти при змінних, тоді як відсутність меблів негативно впливає на ціну.

РОЗДІЛ 3

ПОБУДОВА ЕКОНОМЕТРИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ЖИТЛО ЗІ СТОРОНИ ПРОПОЗИЦІЇ

3.1. Інформаційна база моделювання та аналіз даних

Другий датасет включає 26 незалежних змінних, які в більшості описують чинники, які впливають на житло зі сторони пропозиції (будівельника) та економічного становища. Усі змінні поділяються на 2 категорії: фізичні й фінансові змінні проекту та економічні змінні й індекси.

Розглядаючи другий датасет більш детально, можемо спостерігати, що набір даних включає вартість будівництва, ціни продажу, змінні проекту та економічні змінні. Загалом 29: 7 фізичних і фінансових змінних проекту, 19 економічних змінних та індексів і дві вихідні змінні, які є витратами на будівництво та цінами продажу [31].

- **Building_Area:** Загальна площа будівлі в квадратних метрах (m^2).
- **Area:** Площа ділянки в квадратних метрах (m^2).
- **Total_preliminary_cost:** Загальна попередня кошторисна вартість будівництва на початку проекту в певних одиницях.
- **Preliminary_cost:** Попередньо орієнтовна вартість будівництва на початку проекту в певних одиницях.
- **Equivalent_preliminary_cost:** Еквівалентна попередня оціночна вартість будівництва в обраному базовому році в певних одиницях.
- **Construction_Duration:** Тривалість будівництва в кварталах.
- **Start_project_price_per_ft2:** Ціна одиниці на початку проекту за квадратний фут.

- Sales_price: Фактичні ціни продажу в певних одиницях.
- Construction_Cost: Фактичні витрати на будівництво в певних одиницях.
- Permits: Кількість виданих дозволів на будівництво.
- BSI: Індекс будівельних послуг для попередньо вибраного базового року.
- WPI: Індекс оптових цін будівельних матеріалів за базовий рік.
- Area_permits: Загальна площа дозволів на будівництво, виданих містом/муніципалітетом в квадратних метрах (м²).
- Liquidity: Сукупна ліквідність в певних одиницях.
- Investment: Інвестиції приватного сектора в новобудови в певних одиницях.
- Land_price_index: Індекс цін на землю за базовий рік.
- Number_of_extended_loans: Кількість позик, наданих банком.
- Amount_of_extended_loans: Обсяг позик, наданих банками в терміновому порядку в певних одиницях.
- Interest_rate: Відсоткова ставка за позику в терміновому порядку у відсотках.
- Average_construction_cost: Середня вартість будівництва будівель приватним сектором на момент завершення будівництва в певних одиницях виміру (наприклад, валюта або умовні одиниці) на квадратний метр (м²).
- Start_Average_construction_cost: Середня вартість будівництва будівель приватним сектором на початку будівництва в певних одиницях виміру (наприклад, валюта або умовні одиниці) на квадратний метр (м²).
- Exchange_rate: Офіційний курс обміну валютної одиниці країни до долара США.
- Exchange_rate_street_market: Неофіційний (вуличний) курс обміну валютної одиниці країни до долара США.

- CPI: Індекс споживчих цін в базовому році, що вказує на зміни рівня цін на споживчі товари та послуги порівняно з базовим роком.
- CPI_resources: Індекс споживчих цін для групи товарів, що включає житло, воду, паливо та електроенергію, в базовому році.
- SMI: Індекс фондового ринку, який відображає загальну динаміку цін акцій на фондовому ринку.
- Population: Кількість населення міста, що вказує на розмір та масштаби ринку.
- Gold: Ціна золота за унцію, що відображає вартість золота як фінансового активу.

Узагальнена інформація щодо змінних моделі знаходиться в Додатку А.

На рисунку 3.1 представлені перші 15 спостережень для певних змінних.

	var2	var3	var4	var5	var6	var7	var8	var9	var10	var11	var12
1	3150	920	598.50	190	1010.84	16	1200	2200	410	6713.00	56.2
2	7600	1140	3040.00	400	963.81	23	2900	5000	1000	3152.00	106.0
3	4800	840	480.00	100	689.84	15	630	1200	170	1627.00	41.0
4	685	202	13.70	20	459.54	4	140	165	30	2580.93	12.1
5	3000	800	1230.00	410	631.91	13	5000	5500	700	6790.00	203.8
6	2500	640	1050.00	420	647.32	12	4800	5200	700	6790.00	203.8
7	1810	492	1158.40	640	843.98	11	5700	5800	900	6458.00	224.7
8	1150	380	575.00	500	590.68	6	5300	4600	600	3131.00	259.5
9	2110	540	189.90	90	732.14	5	690	870	110	4638.00	34.6
10	3030	930	515.10	170	1007.38	3	1500	1700	190	7196.00	51.3
11	750	200	90.00	120	846.15	6	1100	1500	150	3213.00	44.5
12	4080	790	530.40	130	759.17	7	1800	2500	190	6788.00	54.2
13	5030	1540	251.50	50	667.88	3	600	780	130	2309.00	24.3
14	4040	890	1212.00	300	752.65	6	3800	5000	450	3052.00	124.4
15	4880	1070	1854.40	380	614.00	5	4100	5200	520	6370.00	190.3

Рисунок 3.1 – Фрагмент набору даних

Джерело: розроблено автором на основі [31]

Далі проведемо агрегацію даних, щоб отримати інформацію про мінімальні, медіанні, середні та максимальні значення для кожної змінної (рис. 3.2). Це дозволить отримати загальне уявлення про розподіл даних. Наприклад, якщо мінімальне значення певної змінної дорівнює нулю, це може свідчити про наявність некоректних або відсутніх даних, які можуть заважати аналізу, видавати неточні дані, тому їй важливо позбутися.

var2	var3	var4	var5	var6
Min. : 200	Min. : 60.0	Min. : 3.7	Min. : 10.0	Min. : 193.1
1st Qu.: 720	1st Qu.: 190.0	1st Qu.: 67.8	1st Qu.: 80.0	1st Qu.: 391.7
Median : 1220	Median : 300.0	Median : 164.7	Median : 140.0	Median : 522.5
Mean : 1729	Mean : 426.1	Mean : 327.9	Mean : 163.1	Mean : 554.4
3rd Qu.: 2100	3rd Qu.: 490.5	3rd Qu.: 366.1	3rd Qu.: 230.0	3rd Qu.: 667.9
Max. : 15670	Max. : 5000.0	Max. : 7208.2	Max. : 640.0	Max. : 3436.9
var10	var11	var12	var13	var14
Min. : 20.0	Min. : 1562	Min. : 12.10	Min. : 10.03	Min. : 0.920
1st Qu.: 110.0	1st Qu.: 2842	1st Qu.: 45.60	1st Qu.: 51.63	1st Qu.: 2.470
Median : 190.0	Median : 3629	Median : 74.90	Median : 79.28	Median : 3.250
Mean : 232.6	Mean : 4211	Mean : 94.43	Mean : 88.05	Mean : 3.605
3rd Qu.: 330.0	3rd Qu.: 6024	3rd Qu.: 137.40	3rd Qu.: 125.83	3rd Qu.: 4.720
Max. : 1000.0	Max. : 7196	Max. : 274.00	Max. : 225.00	Max. : 6.880
var18	var19	var20	var21	var22
Min. : 17.03	Min. : 154.4	Min. : 11.00	Min. : 170.3	Min. : 211.1
1st Qu.: 93.00	1st Qu.: 3622.2	1st Qu.: 14.00	1st Qu.: 641.5	1st Qu.: 744.5
Median : 162.75	Median : 10445.6	Median : 15.00	Median : 1023.7	Median : 1203.3
Mean : 182.00	Mean : 18861.3	Mean : 14.07	Mean : 1327.5	Mean : 1466.3
3rd Qu.: 242.27	3rd Qu.: 21723.4	3rd Qu.: 15.00	3rd Qu.: 1994.6	3rd Qu.: 2025.0
Max. : 432.40	Max. : 73143.5	Max. : 15.00	Max. : 4188.6	Max. : 4741.6
var26	var27	var28	var29	
Min. : 10.06	Min. : 354.6	Min. : 8436	Min. : 141543	
1st Qu.: 42.87	1st Qu.: 2134.5	1st Qu.: 12393	1st Qu.: 588021	
Median : 81.47	Median : 7334.8	Median : 26438	Median : 825511	
Mean : 87.07	Mean : 6604.9	Mean : 28297	Mean : 1041556	
3rd Qu.: 127.33	3rd Qu.: 10082.0	3rd Qu.: 41407	3rd Qu.: 1660444	
Max. : 222.60	Max. : 13596.4	Max. : 50928	Max. : 2606321	

Рисунок 3.2 – Статистичні показники для кожної змінної в наборі даних

Джерело: розроблено автором на основі [31]

Максимальне значення спостережуваної ціни може бути визначене як викид, оскільки воно статично знижується від значень, характерних для нормального розподілу та 3-го квантилю. Крім того, змінні "Sales_price" та мають значні розмахи даних. Для поліпшення моделювання ціни на житло можна спробувати застосувати логарифмічне перетворення цих змінних.

Перед тим як будувати модель цікаво та важливо розуміти, яка залежність між змінними, тому на основі другого датасету визначимо кореляційні коефіцієнти та зобразимо візуальний розподіл за категоріями для того, щоб розуміти різницю між ними.

Кореляційний коефіцієнт може мати значення від -1 до 1. Якщо він близький до -1, то залежні змінні взаємно зв'язані, але в протилежному напрямку, тобто коли одна змінна збільшується, інша зменшується. Якщо кореляційний коефіцієнт близький до 1, то залежні змінні зв'язані між собою в одному напрямку, тобто коли одна змінна збільшується, інша також збільшується. Якщо кореляційний коефіцієнт дорівнює 0, то між змінними немає зв'язку [32].

Побудуємо кореляційну матрицю на рис. 3.3, та дослідимо кореляційні коефіцієнти для числових змінних, окрім категоріальних та фіктивних:

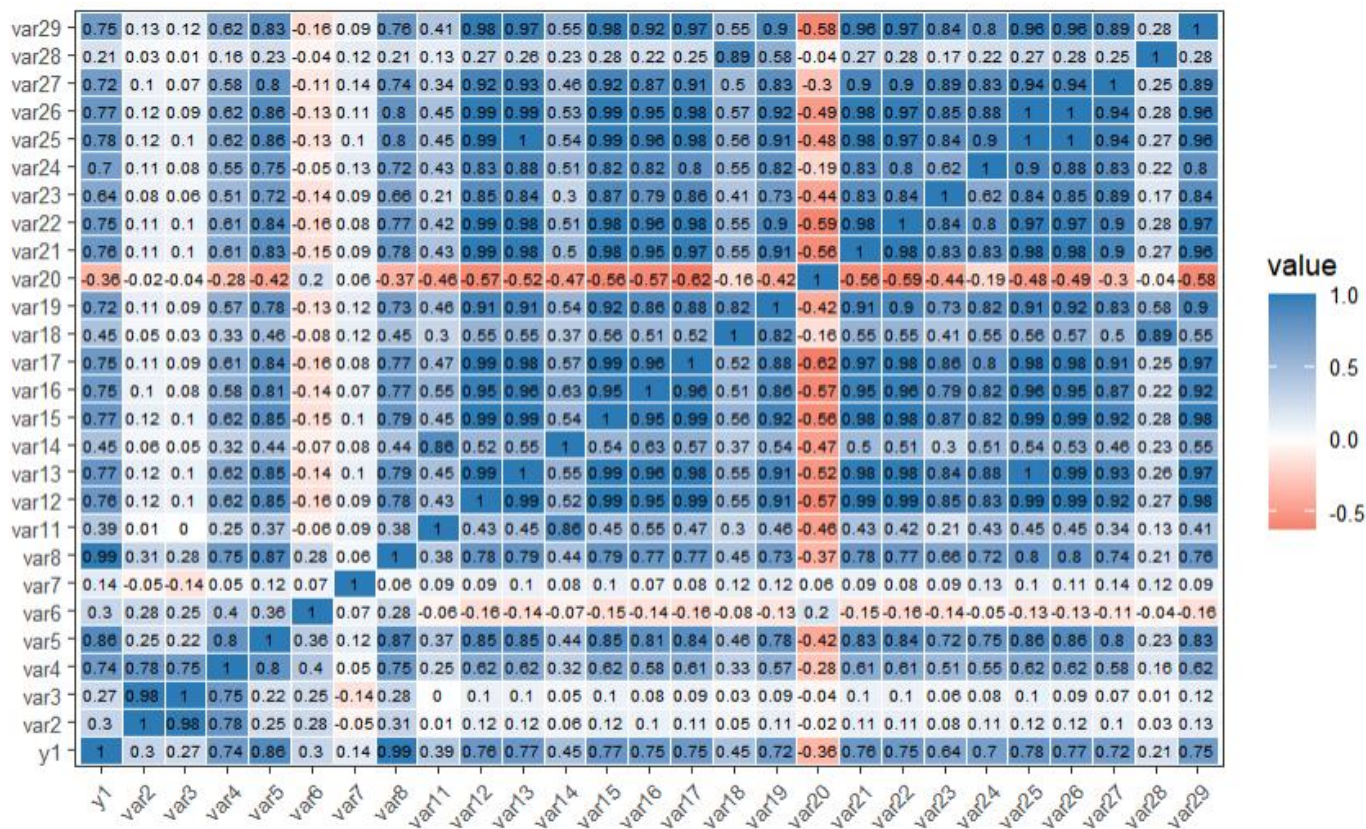


Рисунок 3.3 – Кореляційна матриця

Джерело: розроблено автором на основі [31]

Одразу можемо чітко бачити, що кількість змінних доволі таки велика та всі разом вони будуть давати великий показник мультиколінеарності, тому був проведений аналіз, щодо того, які змінні краще всього виключити з моделі дивлячись на статистичну значущість та кореляції з іншими змінними.

Також був проведений загальний аналіз впливу всіх змінних на ціну житла та виявлено гістограму залежностей на рис. 3.4.

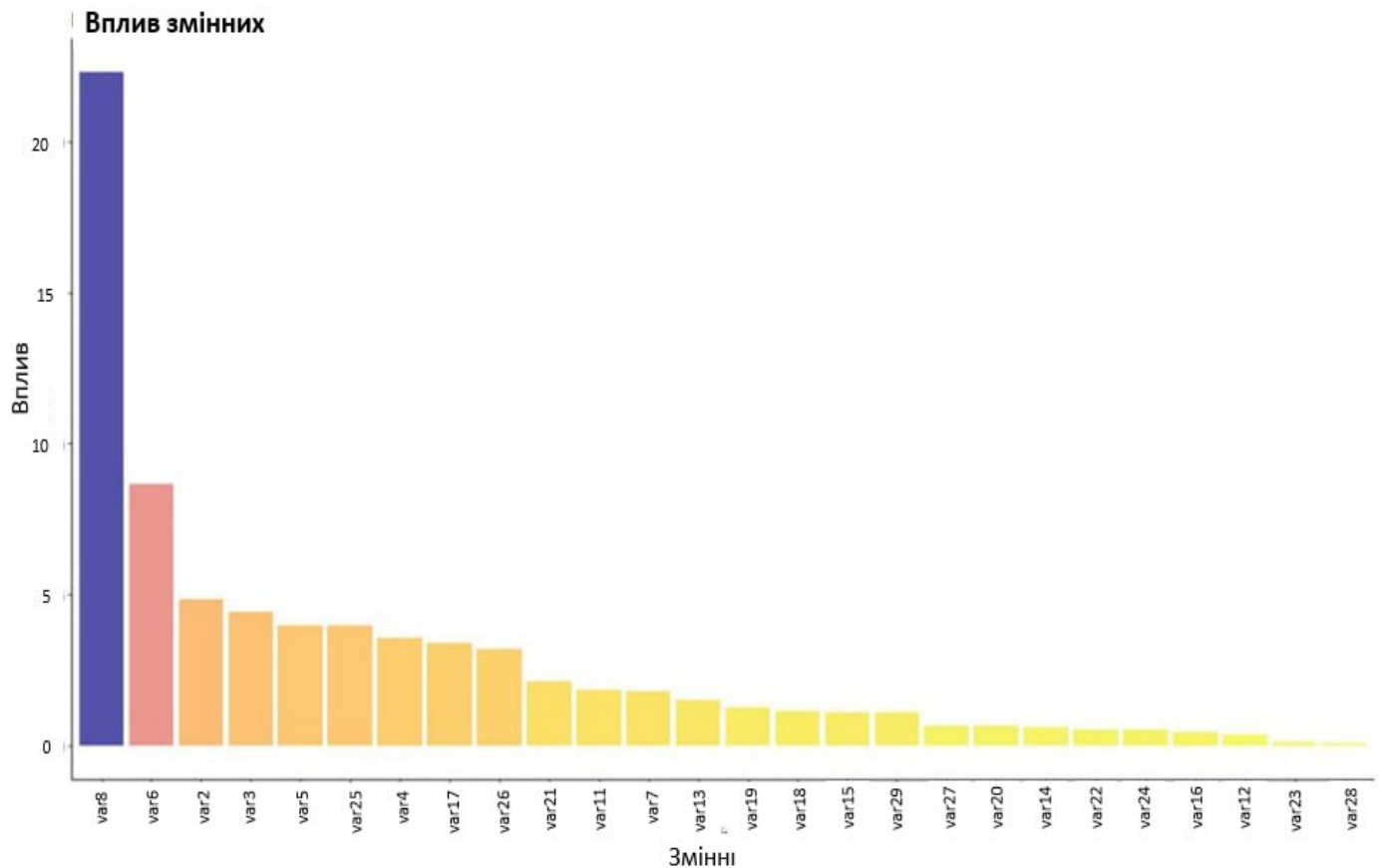


Рисунок 3.4 – Гістограма залежностей

Джерело: розроблено автором на основі [31]

Перш за все, змінна "var8", яка представляє ціну одиниці на початку проекту за квадратний метр, має найбільший вплив. Це виглядає логічним, оскільки ціна є ключовим фактором при оцінці вартості будівництва.

Друга за впливом змінна – "var6", що представляє еквівалентну попередню оціночну вартість будівництва на основі цін на початку проекту в обраному базовому році. Ця змінна також має значний вплив на ціну на житло.

Далі слід звернути увагу на загальну площу будівлі (var2) та площу житла (var3), яке купується. Ці змінні також мають помітний вплив на ціну на житло.

Нарешті, іншими важливими змінними є попередньо орієнтовна вартість будівництва на основі цін на початок проекту (var5) та індекс споживчих цін (ІСЦ) в базовому році (var25). Ці змінні також мають значний вплив на ціну на житло.

Впливовими змінними також загальна попередня кошторисна вартість будівництва виходячи з цін на початок проекту (var4), індекс цін на землю (var17) та ІСЦ житла, води, палива та електроенергії (var26).

Уже на даному етапі навіть за змінними можна передбачити високу мультикореляцію, тому в наступному розділі паралельно з моделюванням потрібно позбутися мультиколінеарності.

3.2. Оцінювання параметрів моделі та її статистичної значущості

Для поліпшення моделювання ціни на житло усі змінні були прологарифмовані задля підвищення адекватності моделі, яка була отримана в попередньому аналізі. За допомогою даного підходу R^2 був збільшений більш ніж на 15 одиниць в кінцевому результаті.

При побудові специфікації моделі важливо враховувати та правильно позначити всі незалежні змінні в залежності від типу даної змінною. Набір даних містить усі числові змінні, тому маємо наступну специфікацію (3.1) на основі додатку А:

$$\begin{aligned}
 \text{Log}(\text{var9}) = & b_0 + b_1 \text{Log}(\text{var2}) + b_2 \text{Log}(\text{var3}) + b_3 \text{Log}(\text{var4}) + b_4 \text{Log}(\text{var5}) + \\
 & b_5 \text{Log}(\text{var6}) + b_6 \text{Log}(\text{var7}) + b_7 \text{Log}(\text{var8}) + b_8 \text{Log}(\text{var11}) + b_9 \text{Log}(\text{var12}) + \\
 & b_{10} \text{Log}(\text{var13}) + b_{11} \text{Log}(\text{var14}) + b_{12} \text{Log}(\text{var15}) + b_{13} \text{Log}(\text{var16}) + b_{14} \text{Log}(\text{var17}) + \\
 & b_{15} \text{Log}(\text{var18}) + b_{16} \text{Log}(\text{var19}) + b_{17} \text{Log}(\text{var20}) + b_{18} \text{Log}(\text{var21}) + b_{19} \text{Log}(\text{var22}) + \\
 & b_{20} \text{Log}(\text{var23}) + b_{21} \text{Log}(\text{var24}) + b_{22} \text{Log}(\text{var25}) + b_{23} \text{Log}(\text{var26}) + b_{24} \text{Log}(\text{var27}) + \\
 & b_{25} \text{Log}(\text{var28}) + b_{26} \text{Log}(\text{var29}) + \varepsilon
 \end{aligned}
 \tag{3.1}$$

Визначимо параметри та побудуємо моделі в середовищі RStudio з використанням мови програмування R. Для цього були завантажені необхідні бібліотеки та дані.

В R-середовищі було застосовано логарифмування до змінних (рис. 3.5):

```

y1 <- log(df1$var9)
var2 <- log(df1$var2)
var3 <- log(df1$var3)
var4 <- log(df1$var4)
var5 <- log(df1$var5)
var6 <- log(df1$var6)
var7 <- log(df1$var7)
var8 <- log(df1$var8)
var11 <- log(df1$var11)
var12 <- log(df1$var12)
var13 <- log(df1$var13)
var14 <- log(df1$var14)
var15 <- log(df1$var15)
var16 <- log(df1$var16)
var17 <- log(df1$var17)
var18 <- log(df1$var18)
var19 <- log(df1$var19)
var20 <- log(df1$var20)
var21 <- log(df1$var21)
var22 <- log(df1$var22)
var23 <- log(df1$var23)
var24 <- log(df1$var24)
var25 <- log(df1$var25)
var26 <- log(df1$var26)
var27 <- log(df1$var27)
var28 <- log(df1$var28)
var29 <- log(df1$var29)

```

Рисунок 3.5 – Переведення змінних до іншого типу даних

Джерело: розроблено автором на основі [31]

Починаємо аналіз моделей (n1, n2, n3, n4, n5, n6, n7), які різняться між собою кількістю предикторів.

При виключенні будь-якої змінної використовувався підхід виключення змінних, які дають найбільшу мультиколінеарність, а вже наступним етапом виключалися змінні які були статистично незначущими.

Варто зауважити, що в даній роботі не буде продемонстрований процес позбуття мультиколінеарності та вибору найефективнішою моделі, адже він є доволі таки об'ємним, тому виходячи з результатів аналізу була отримана модель n7, як включає змінні: var3, var7, var23, var28, які відповідають площі ділянки, тривалості будівництва, офіційному курсу по відношенню до долара та населенню міста.

На рис. 3.6 продемонстрована реалізація моделі:

```
n7 <- lm(y1~var3+var7+var23+var28)
summary(n7)
```

Рисунок 3.6 – Реалізація моделі

Джерело: розроблено автором на основі [31]

На рис. 3.7 відображені результати даної моделі:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.11225     0.66395  -3.181  0.00159 **
var3         0.28457     0.04385   6.490 2.79e-10 ***
var7         0.31041     0.10644   2.916  0.00376 **
var23        0.63715     0.03997  15.942 < 2e-16 ***
var28        0.14434     0.05617   2.570  0.01056 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6064 on 367 degrees of freedom
Multiple R-squared:  0.4914,    Adjusted R-squared:  0.4858
F-statistic: 88.64 on 4 and 367 DF,  p-value: < 2.2e-16
```

Рисунок 3.7 – Результати моделі

Джерело: розроблено автором на основі [31]

Критерій Стюдента підтверджує, що залежна змінна є статистично значущою і має значний вплив. Також всі незалежні змінні є статистично значущими, оскільки ймовірність того, що один з параметрів моделі рівний нулю, занадто низька. Останній стовпчик інформує, що можна довіряти коефіцієнтам моделі, оскільки значення $\Pr(>|t|)$ є дуже малими, меншими за 0.05. Підтвердженням цьому є позначення "****", "***" і "**", які використовуються для позначення рівня статистичної значущості.

Multiple R-squared показує, що ціна житла пояснюється незалежними змінними в моделі на 49.14%, а Adjusted R-squared, враховуючи предиктори, показує, що ціна житла пояснюється незалежними змінними в моделі на 48.58%. F-statistic демонструє, що можна адекватно аналізувати дані.

Результати оцінки рівняння для другого датасету наведені в табл. 3.1.

Таблиця 3.1 – Результати оцінки рівняння [31]

Змінна	Коефіцієнт	Стандартна похибка	t-статистика	P-значення	$\Pr(> t)$
Const	-2.11225	0.66395	-3.181	0.00159	**
Area	0.28457	0.04385	6.49	2.79E-10	****
Construction_Duration	0.31041	0.10644	2.916	0.00376	**
Exchange_rate	0.63715	0.03997	15.942	<2.00E-16	****
Population	0.14434	0.05617	2.57	0.01056	*

Джерело: розроблено автором на основі [31]

Отримали таблицю з коефіцієнтами, стандартними похибками, t-статистиками та p-значеннями, які показують нам, що обрані змінні є ефективним и та мають вплив на ціну житла.

Наявні коефіцієнти для побудови повноцінної специфікації моделі, проте попередньо потрібно її перевірити на ряд показників.

3.3. Тестування гіпотез автокореляції та гетероскедастичності залишків

Після побудови моделей, важливо перевірити, чи є модель "адекватною". Необхідно провести аналіз на наявність напруженості, викидів, залишків, впливових спостережень, мультиколінеарності, автокореляції та інших аспектів.

Давайте розпочнемо з аналізу залишків, який зображений на рис. 3.8.

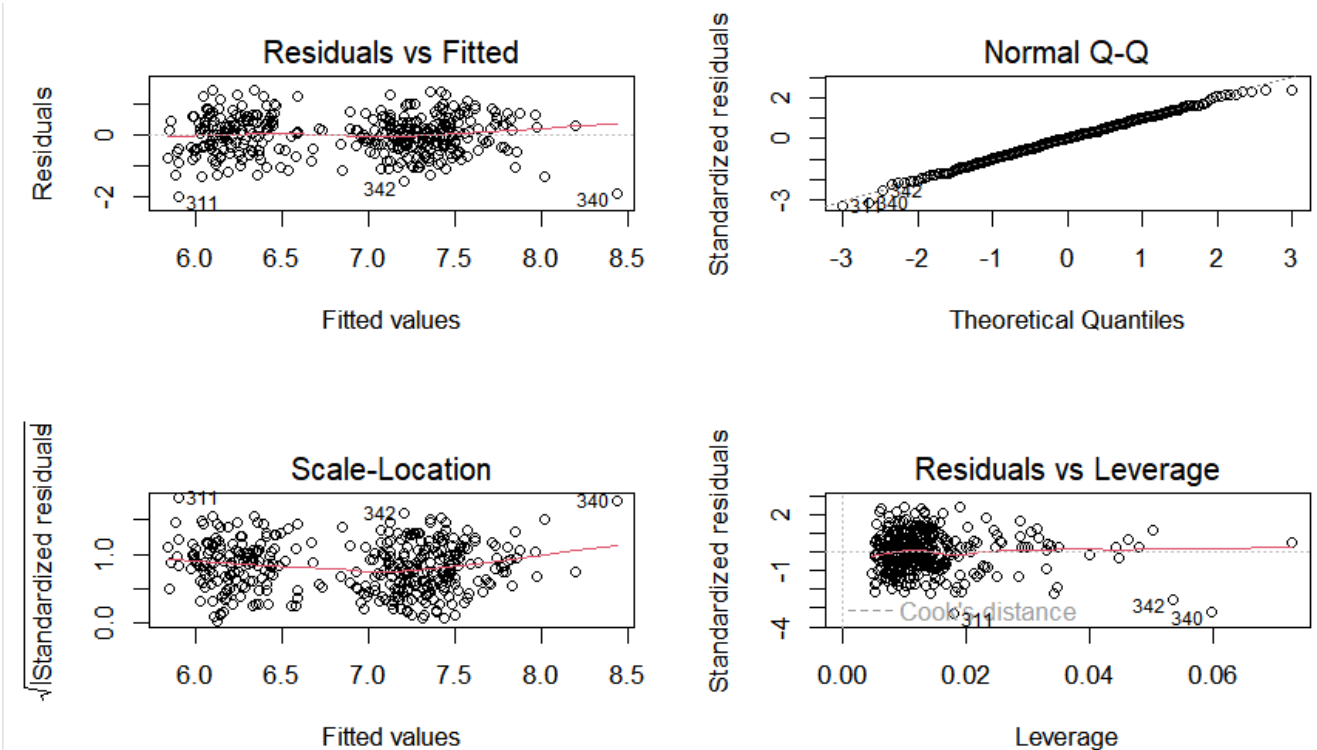


Рисунок 3.8 – Аналіз залишків

Джерело: розроблено автором на основі [31]

Графік "Residuals vs Fitted" показує, що точки розкидані навколо прямої лінії та більш менш згруповані в двох областях. Цей графік відображає розсіювання залишків на вертикальній осі (y) та встановлених значень на горизонтальній осі (x). В даному випадку розподіл залишків виглядає досить "задовільно".

Перевірка на нормальність за допомогою графіка Q-Q показує, що всі точки розташовані приблизно на одній лінії під кутом 30 градусів. Це свідчить про нормальний розподіл залишків.

Графік "Scale-Location" показує, що точки розкидані навколо прямої лінії та більш менш згруповані в двох областях. Це вказує на сталу дисперсію залишків, яка не залежить від прогнозованих значень змінних.

Графік "Residuals vs Leverage" показує, що точки знаходяться в центрі по вертикалі та переважно зліва по горизонталі. Це свідчить про відсутність впливових спостережень (викидів), які б могли суттєво вплинути на оцінки регресійних коефіцієнтів. Однак, спостерігаються декілька точок, які знаходяться за пунктирною лінією, що свідчить про наявність спостережень, які мають суттєвий вплив на модель.

Побудуємо діаграму розподілу залишків на рис. 3.9 для подальшого аналізу нормальності розподілу залишків.

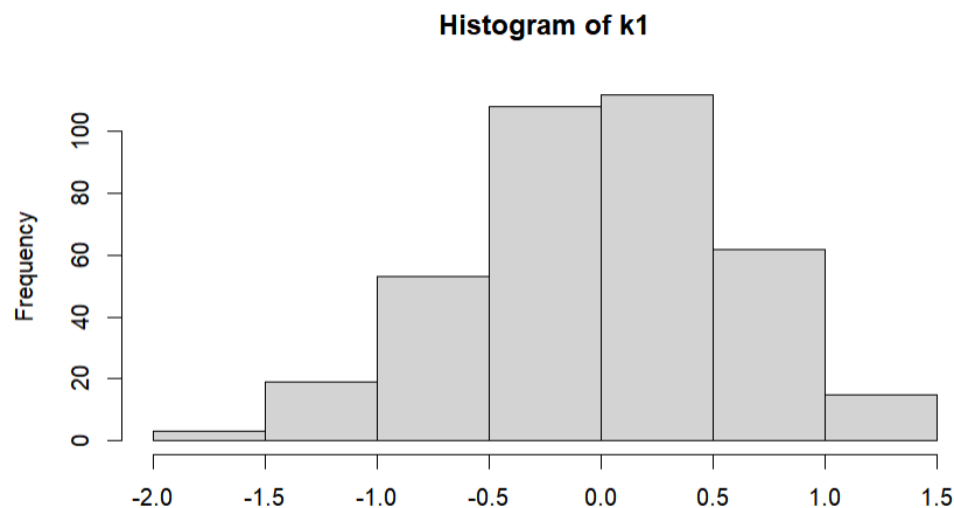


Рисунок 3.9 – Нормальність розподілу залишків

Джерело: розроблено автором на основі [31]

На даному графіку можна чітко стверджувати, що залишки мають більш менш нормальний розподіл, адже можна умовно намалювати «дзвінок» поверх стовбців. Модель демонструє високу точність у поясненні зв'язку між залежною та незалежними змінними, що підтверджує надійність оцінок параметрів.

Для виявлення викидів була використана функцію `OutlierTest()`. За допомогою цього аналізу встановлено, що спостереження з найбільшим відхиленням від моделі є рядок 311, що свідчить про наявність викиду. Крім того, підтверджено, що більше немає викидів, які мають значення, що перевищує критичну відстань Кука.

Визначено ряд показників, таких як відстань Кука, `DFBETAS`, `DFFITs`, коефіцієнт варіації та значення матриці проєкції на простір регресорів. Ці показники представлені на рис. 3.10.

	dfb.1 <dbl>	dfb.var3 <dbl>	dfb.var7 <dbl>	dfb.vr23 <dbl>
1	-9.902947e-02	1.303398e-01	2.251532e-01	-1.093419e-01
2	-7.782796e-02	6.872174e-02	1.338837e-01	5.930577e-03
3	-1.922974e-02	2.697768e-02	4.730279e-02	-2.379437e-02
4	-1.531749e-01	4.428534e-02	7.334651e-02	5.565590e-02
5	-1.267719e-01	9.132119e-02	1.491512e-01	2.405543e-02
6	-1.219828e-01	7.442273e-02	1.368240e-01	2.759518e-02
7	-4.853965e-03	8.472112e-02	2.126344e-01	8.408694e-02
8	-1.019394e-01	1.347293e-02	-1.264344e-02	5.741503e-02
9	2.191472e-02	2.669979e-02	-1.259846e-02	-4.023097e-02
10	-2.787860e-02	1.063134e-01	-1.821100e-01	-1.095431e-01

1-10 of 372 rows | 1-9 of 10 columns

Рисунок 3.10 – Перевірка показників «адекватності» моделі

Джерело: розроблено автором на основі [31]

Бачимо, що деякі спостереження є не типовими, вони помічені зірочками. Їх кількість = 25.

Протестуємо гіпотези щодо гетероскедастичності, крутизни розподілу даних, асиметрії розподілу даних, лінійності розподілу залежної змінної також перевіримо загальну статистику аналізу, адже дуже важливо переконатися в тому, що нічого не заважає точності та надійності аналізу (рис. 3.11).

	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	5.51638479	0.2382930	Assumptions acceptable.
Skewness	1.38151730	0.2398429	Assumptions acceptable.
Kurtosis	0.01052422	0.9182903	Assumptions acceptable.
Link Function	1.59281873	0.2069239	Assumptions acceptable.
Heteroscedasticity	2.53152453	0.1115924	Assumptions acceptable.

Рисунок 3.11 – Перевірка моделі на вимоги, що повинні бути у лінійних моделях

Джерело: розроблено автором на основі [31]

Дивлячись на результати аналізу, можна стверджувати, що всі показники вважаються "прийнятними" або "задовільними" з точки зору припущень моделі. Основний зміст повідомлення, що "Assumptions acceptable", вказує на те, що модель відповідає певним припущенням і може бути використана для аналізу даних.

Роблячи загальний висновок маємо наступне:

- аналіз статистичних припущень моделі є прийнятними.
- асиметрія відсутня.
- розподіл не містить хвостів, де присутні екстремальні значення.
- використана функція зв'язку між регресорами і залежною змінною є прийнятною.
- нерівномірність варіації в резидуальних даних (гетероскедастичність) відсутня

Подальшим кроком буде оцінка мультиколінеарності шляхом проведення загальної та індивідуальної перевірки. Результати цих перевірок можна побачити на рис. 3.12.

Overall Multicollinearity Diagnostics

	MC Results	detection
Determinant $ X'X $:	0.9311	0
Farrar Chi-Square:	26.3269	1
Red Indicator:	0.1093	0
Sum of Lambda Inverse:	4.1440	0
Theil's Method:	-1.3353	0
Condition Number:	58.1024	1

Рисунок 3.12 – Результати перевірок на мультиколінеарність

Джерело: розроблено автором на основі [31]

Отже, на основі отриманих результатів можна зробити висновок, що модель має незначну проблему з мультиколінеарністю. Це може призвести до певного збільшення стандартних помилок при моделюванні та зниження точності прогнозування. Проте, вплив цієї проблеми є незначним і не потребує втручання.

Для підтвердження наявності мультиколінеарності необхідно провести індивідуальний аналіз за допомогою оцінки мультиколінеарності для кожної окремої змінної. Результати цього аналізу можна побачити на рис. 3.13.

All Individual Multicollinearity Diagnostics Result

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
var3	1.0249	0.9757	3.0580	4.5994	0.9878	1.1638	0	0.0080	0.7007
var7	1.0406	0.9610	4.9762	7.4845	0.9803	1.1816	0	0.0078	1.1230
var23	1.0379	0.9634	4.6546	7.0008	0.9816	1.1786	0	0.0079	1.0531
var28	1.0406	0.9610	4.9771	7.4860	0.9803	1.1816	0	0.0078	1.1232

Рисунок 3.13 – Індивідуальний аналіз мультиколінеарності

Джерело: розроблено автором на основі [31]

Проаналізувавши всі показники для визначення мультиколінеарності, можна стверджувати, що мультиколінеарність настільки несуттєва, що на неї не варто й звертати уваги.

Попередньо лише 2 із 6 тестів змогли підтвердити мультиколінеарність і якщо проаналізувати показники, то можна чітко зрозуміти, що вони вже є на грані, яку можна назвати «відсутність мультиколінеарності».

Останній тестом буде аналіз на автокореляції за Дарбіном-Уотсоном (рис. 3.14).

```

lag Autocorrelation D-W Statistic p-value
1      0.3959201      1.204273      0
Alternative hypothesis: rho != 0

```

Рисунок 3.14 – Тест Дарбіна-Уотсона

Джерело: розроблено автором на основі [31]

Автокореляція складає 0.39, що є замалим показником, що приймати будь-які дії, тому можна стверджувати, що вона є відсутньою.

3.4. Економічна інтерпретація

Проведемо аналіз специфікації моделі для встановлення коефіцієнтів, що відповідають кожній впливовій змінній.

У цьому рівнянні на основі додатку А, кожен множник перед незалежною змінною вказує, як змінюється залежна змінна (логарифм ціни на житло) при зміні відповідної змінної, за умови, що інші змінні залишаються незмінними (3.2).

$$\begin{aligned}
 \text{Log}(\text{var9}) = & -2.11225 + 0.28457 * \text{Log}(\text{var3}) + 0.31041 * \text{Log}(\text{var7}) + 0.63715 * \\
 & \text{Log}(\text{var23}) + 0.14434 * \text{Log}(\text{var28}) + \varepsilon
 \end{aligned}
 \tag{3.2}$$

Економічна інтерпретація кожної змінної в рівнянні:

Log(var3): Коефіцієнт вказує, що збільшення площі на 1% призведе до приблизно збільшення ціни на житло на 0.28457%. Це означає, що збільшення площі має позитивний вплив на ціну на житло.

Log(var7): Коефіцієнт вказує, що збільшення тривалості будівництва на 1% призведе до приблизно 0.31041% збільшення ціни на житло. Це означає, що збільшення тривалості будівництва має позитивний вплив на ціну на житло.

Log(var23): Коефіцієнт вказує, що збільшення обмінного курсу на 1% призведе до приблизно 0.63715% збільшення ціни на житло. Це означає, що збільшення обмінного курсу має позитивний вплив на ціну на житло.

Log(var28): Коефіцієнт вказує, що збільшення населення на 1% призведе до приблизно 0.14434% одиниці ціни на житло. Це означає, що збільшення населення має позитивний вплив на ціну на житло.

Отже, змінні площі, тривалості будівництва, обмінного курсу і населення мають позитивний вплив на ціну на житло.

ВИСНОВКИ

Багатофакторна регресія широко використовується в різних аспектах життя. У цьому документі представлено процес побудови багатовимірної регресійної моделі для спрощеної задачі оцінки цін на житлову нерухомість. Це типовий процес побудови регресійних моделей, які можуть застосовуватися до багатьох програм і аспектів. Загалом, ідеї економетричного підходу на ринку нерухомості також можна використовувати для інших галузей та ринків.

Розглянуто теоретичні основи моделювання процесів на ринку житлової нерухомості, а саме: 1) проведено економічний аналіз ринку житлової нерухомості; 2) виконано огляд математичних методів та моделей дослідження ринку житлової нерухомості; 3) обґрунтовано вибір методів економетричного моделювання для проведення дослідження.

Виконано економетричний аналіз факторів, що впливають на вартість житлової нерухомості, з боку попиту на ринку нерухомості, а саме: 1) описано інформаційну базу моделювання; 2) описано специфікацію економетричних моделей; 3) виконано оцінювання параметрів моделі та їх економічну інтерпретацію; 4) розглянуто гіпотези щодо впливу якісних (дискретних, категоріальних) змінних за допомогою фіктивних змінних.

Виконано економетричний аналіз факторів, що впливають на вартість житлової нерухомості, з боку пропозиції на ринку нерухомості.

За результатами першої моделі зі сторони попиту надійними для визначення цін на житло є площа, наявність додаткових ванних кімнат, наявність більшої кількості поверхів у будівлі, наявність головної дороги поруч, наявність гостьової кімнати, підвалу, системи гарячої води, кондиціонування, парковки та бажання покупців мати квартиру в певному районі позитивно впливає на ціни житла.,

Наявність цих аспектів або змінних у будинку підвищить його ціну. Лише відсутність меблів негативно впливає на ціну.

R-квадрат, отриманий за допомогою першої моделі регресії, є досить високим, 69,81%. Тому саме такий відсоток зміни ціни на житло пояснюється незалежними змінними в моделі. Виходячи з результатів проведених аналізу, ця модель є важливою, однак є можливості для вдосконалення.

За результатами другої моделі зі сторони пропозиції надійними для визначення цін на житло є площа, тривалість будівництва, обмінний курс і чисельність населення. Усі ці фактори мають позитивний вплив на ціну на житло.

R-квадрат, отриманий за допомогою другої моделі регресії не є високим, але задовільним для реальних даних – 48,58%. Тому саме такий відсоток зміни ціни на житло пояснюється незалежними змінними в моделі. Виходячи з результатів проведених випробувань, ця модель є важливою, однак є можливості для вдосконалення.

Теоретична, методична та практична значущість отриманих результатів полягає в тому, що техніки економетричного аналізу та моделювання ціни, які представлені в дослідженні, є корисними для різних суб'єктів ринку нерухомості. Наприклад, громадяни можуть використовувати ці дані, щоб дізнатися про залежність ціни житлової нерухомості на ринку у своїх регіонах та обрати житло з своїм бюджетом. Маркетологи, продавці та дослідники ринку нерухомості можуть дізнатися більше про поведінку та вподобання споживачів, а також про загальну ситуацію на ринку. Держава може використовувати ці дані для розробки раціональної державної політики, аналізувати, як, наприклад, тимчасове скасування певних податків або введення обмеження на певні характеристики вплине на попит та пропозицію на ринку, і проектування інфраструктури, яка враховує потреби різних сторін на ринку нерухомості.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Агентство нерухомості Дуоком. Від чого залежить ціна квартир при продажу та що впливає на вартість нерухомості. URL: <https://duo.com.ua/pytannya-vidpovidi/prodazh-nerukhomosti/vid-chogo-zalezhyt-tsina-kvartyr-pry-prodazhu-ta-scho-vplyvae-na-vartist-nerukhomosti> (дата звернення: 20.05.2023).
2. Оцінка та управління нерухомістю: навчальний посібник / Кучеренко В.Р. та ін. Одеса: Видавництво ТОВ «Лерадрук», 2013. 272 с.
3. Global Property Guide. Global Residential Real Estate Market Analysis 2022. URL: <https://www.globalpropertyguide.com/investment-analysis/Global-Residential-Real-Estate-Market-Analysis-2022> (дата звернення: 20.05.2023).
4. Формування ринку нерухомості. Практичні аспекти та особливості оцінки: навчальний посібник / С.М. Кобзан та ін. Київ : Видавництво «Юрінком Інтер», 2019. 212 с.
5. Greater London Authority. House Prices in London. URL: <https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf> (дата звернення: 20.05.2023).
6. Коваль А.М. Економіко-математичне моделювання впливу макроекономічних факторів на ціноутворення ринку нерухомості. *Економіка і суспільство*, 2022. Вип. 43.
URL: <https://economyandsociety.in.ua/index.php/journal/article/view/1742/1678> (дата звернення: 20.05.2023).
7. Stephen Gibbons, Henry Overman, Guilherme Resende. Real Earnings Disparities in Britain. LSE Research Online, 2011. URL: <http://eprints.lse.ac.uk/33576/1/sercdp0065.pdf> (дата звернення: 20.05.2023).
8. Karl Case, Robert Shiller, Forecasting Prices and Excess Returns in the Housing Market. Real Estate Economics, 1990.

9. John Quigley, Real Estate Prices and Economic Cycles. International Real Estate Review, 1999.
10. Yiyang Gu. What are the most important factors that influence the changes in London Real Estate Prices? How to quantify them?. URL: <https://arxiv.org/ftp/arxiv/papers/1802/1802.08238.pdf> (дата звернення: 20.05.2023).
11. Економетрія: навчальний посібник. / Корольов О.А. та ін. Київ: Європейський університет, 2002. 660 с.
12. Кореляційно-регресійний аналіз: тези до наукової конференції. / Тутова А., Харків, Матеріали Всеукраїнської науково-технічної конференції магістрантів і студентів, 2020.
13. Економетрика: методичні вказівки для самостійної роботи. / Бегун С.І., Луцьк: Вид-во ПП Іванюк В.П., 2014. 60 с.
14. Економетрія: навчальний посібник. / Лещинський О.Л., Рязанцева В.В., Юнькова О.О., Київ: ДП «Видавничий дім «Персонал», 2008. 208 с.
15. Економетрика: навчальний посібник. / Руська Р.В., Тернопіль: Тайп, 2012. 224 с.
16. Shaun Turney. The Coefficient of Determination (R^2). Scribbr, 2022. URL: [https://www.scribbr.com/statistics/coefficient-of-determination/#:~:text=The%20coefficient%20of%20determination%20\(R%C2%B2\)%20is%20a%20number%20between%200,predicted%20by%20the%20statistical%20model](https://www.scribbr.com/statistics/coefficient-of-determination/#:~:text=The%20coefficient%20of%20determination%20(R%C2%B2)%20is%20a%20number%20between%200,predicted%20by%20the%20statistical%20model) (дата звернення: 20.05.2023).
17. Standard Error of the Mean vs. Standard Deviation: What's the Difference?. Investopedia. URL: <https://www.investopedia.com/ask/answers/042415/what-difference-between-standard-error-means-and-standard-deviation.asp> (дата звернення: 20.05.2023).
18. Correlation and Regression. BMJ, URL: <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression> (дата звернення: 20.05.2023).

19. Рябчий В.А., Рябчий В.В.. Обґрунтування принципу найменших квадратів. 2012. URL:

<http://ir.nmu.org.ua/bitstream/handle/123456789/2644/%D0%9E%D0%B1%D2%91%D1%80%D1%83%D0%BD%D1%82%D1%83%D0%B2%D0%B0%D0%BD%D1%8F%20%D0%BF%D1%80%D0%BD%D1%86%D0%B8%D0%BF%D1%83%20%D0%BD%D0%B0%D0%B9%D0%BC%D0%B5%D0%BD%D1%88%D0%B8%D1%85%20%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D1%96%D0%B2%20%5B2012%5D.pdf?sequence=1> (дата звернення: 20.05.2023).

20. Горошко Ю.В.. Метод найменших квадратів та його реалізація засобами ЧПУ. URL:

<http://erpub.chnpu.edu.ua:8080/jspui/bitstream/123456789/1464/1/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%20%D0%BD%D0%B0%D0%B9%D0%BC%D0%B5%D0%BD%D1%88%D0%B8%D1%85%20%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D1%96%D0%B2%20%D1%82%D0%B0%20%D0%B9%D0%BE%D0%B3%D0%BE%20%D1%80%D0%B5%D0%B0%D0%BB%D1%96%D0%B7%D0%B0%D1%86%D1%96%D1%8F%20%D0%B7%D0%B0%D1%81%D0%BE%D0%B1%D0%B0%D0%BC%D0%B8%20%D0%9D%D0%86%D0%A2.pdf> (дата звернення: 20.05.2023).

21. Matteo Courthoud. Outliers, Leverage, Residuals, and Influential Observations. Towards Data Science, 2022. URL: <https://towardsdatascience.com/outliers-leverage-residuals-and-influential-observations-df3065a0388e> (дата звернення: 20.05.2023).

22. Clay Ford. Understanding Q-Q Plots. University of Virginia Library, 2015. URL: <https://data.library.virginia.edu/understanding-q-q-plots/> (дата звернення: 20.05.2023).

23. Influence Measures. R Documentation, URL: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/influence.measures> (дата звернення: 20.05.2023).

24. Aniruddha Bhandari. What is Multicollinearity? Analytics Vidhya, 2023. URL: <https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#:~:text=Multicollinearity%20occurs%20when%20two%20or%20more%20independent%20variables%20have%20a,variable%20on%20the%20dependent%20variable> (дата звернення: 20.05.2023).

25. imcdiag: Individual Multicollinearity Diagnostic Measures. rdrv. URL: <https://rdrr.io/cran/mctest/man/imcdiag.html> (20.05.2023).

26. Чернушевич В.А. Масова оцінка житлової нерухомості з використанням методів машинного навчання. ЛОГОС. ONLINE. 2020 URL: <https://www.ukrlogos.in.ua/10.11232-2663-4139.16.50.html> (дата звернення: 20.05.2023).

27. Tejvan Pettinger. Factors that affect the housing market. Economics help. 2022 URL: <https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/#:~:text=The%20housing%20market%20is%20influenced,be%20determined%20by%20available%20supply> (дата звернення: 20.05.2023).

28. M YASSER H. Housing Prices Dataset. Kaggle. 2022 URL: <https://www.kaggle.com/datasets/yasserh/housing-prices-dataset> (дата звернення: 20.05.2023).

29. Економетрика: навчально-методичний комплекс. / Ставицький А.В, Київ: КНУ. 2004. 112 с.

30. Специфікація моделі. StudFiles. URL: <https://studfile.net/preview/9091091/page:2/> (дата звернення: 20.05.2023).

31. Mohammad Rafiei. Residential Building Data Set. UC Irvine. 2018. URL: <https://archive-beta.ics.uci.edu/dataset/437/residential+building+data+set> (дата звернення: 20.05.2023).

32. Pritha Bhandari. Correlation Coefficient | Types, Formulas & Examples. Scribbr. 2021 URL: <https://www.scribbr.com/statistics/correlation-coefficient/> (дата звернення: 20.05.2023).

ДОДАТКИ

Додаток А

Змінна	Опис	Одиниця виміру
Bulding_Area (var2)	Загальна площа будівлі	м ²
Area (var3)	Площа ділянки	м ²
Total_preliminary_cost (var4)	Загальна попередня кошторисна вартість будівництва виходячи з цін на початок проекту	10000000 ум.од.
Preliminary_cost (var5)	Попередньо орієнтовна вартість будівництва виходячи з цін на початок проекту	10000 ум.од.
Equivalent_preliminary_cost (var6)	Еквівалентна попередня оціночна вартість будівництва на основі цін на початку проекту в обраному базовому році	10000 ум.од
Construction_Duration (var7)	Тривалість будівництва	Квартал
Start_project_price_per_ft ² (var8)	Ціна одиниці на початку проекту за фут2	10000 ум.од
Sales_price (var9)	Фактичні ціни продажу (випуск)	10000 ум.од
Construction_Cost (var10)	Фактичні витрати на будівництво (випуск)	10000 ум.од
Permits (var11)	Кількість виданих дозволів на будівництво	N/A
BSI (var12)	Індекс будівельних послуг (BSI) b для попередньо вибраного базового року a	N/A
WPI (var13)	Індекс оптових цін (ІЦЦ) c будівельних матеріалів за базовий рік	N/A
Area_permits (var14)	Загальна площа дозволів на будівництво, виданих містом/муніципалітетом	м ²
Liquidity (var15)	Сукупна ліквідність	10000000 ум.од.
Investment (var16)	Інвестиції приватного сектора в новобудови	10000000 ум.од.
Land_price_index (var17)	Індекс цін на землю за базовий рік a	10000000 ум.од.
Number_of_extended_loans (var18)	Кількість позик, наданих банками в часовому вирішенні	N/A
Amout_of_extended_loans (var19)	Обсяг позик, наданих банками в терміновому порядку e	10000000 ум.од.
Interest_rate (var20)	Відсоткова ставка за позику в терміновому порядку e	%
Average_construction_cost (var21)	Середня вартість будівництва будівель приватним сектором на момент завершення будівництва	10000 ум.од. /м ²

Start_Average_construction_cost (var22)	Середня вартість будівництва будівель приватним сектором на початку будівництва	10000 ум.од. /м ²
Exchange_rate (var23)	Офіційний курс по відношенню до доларів	ум.од.
Exchange_rate_street_market (var24)	Неофіційний (вуличний) курс обміну до доларів	ум.од.
CPI (var25)	Індекс споживчих цін (ІСЦ) і в базовому році	N/A
CPI_resources (var26)	ІСЦ житла, води, палива та електроенергії в базовому році	N/A
SMI (var27)	Індекс фондового ринку	N/A
Population (var28)	Населення міста	N/A
Gold (var29)	Ціна золота за унцію	ум.од.