



Володимир ДРУЖИНІН¹, д-р техн. наук, проф.
ORCID ID: 0009-0009-5049-0099
e-mail: volodymir.druzhynin@knu.ua

Євгеній БАТРАК², канд. техн. наук, доц.
ORCID ID: 0000-0003-3027-4507
e-mail: batrak.yevhenii@lil.kpi.ua

Наталія ЦЬОПА², канд. техн. наук, доц.
ORCID ID: 0000-0002-3215-0711
e-mail: tsopa.natalia@lil.kpi.ua

Ольга НІКОЛЕНКО², магістр
ORCID ID: 0009-0003-2304-7821
e-mail: nikolenko.olha@lil.kpi.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

МУЛЬТИДЖЕРЕЛЬНИЙ АЛГОРИТМІЧНИЙ АНАЛІЗ ЧИТАЦЬКИХ І КОМЕРЦІЙНИХ ТРЕНДІВ КНИЖКОВОГО РИНКУ

Вступ. Книжковий ринок зазнає швидких трансформацій під впливом цифрових платформ, а вчасне виявлення літературних трендів стає критично важливим для видавництва, що працюють із перекладною літературою. Мета дослідження – створення й обґрунтування інтегрованої моделі мультиджерельного аналізу ключових читацьких і комерційних показників за даними Goodreads, Amazon і The New York Times Bestsellers для забезпечення виявлення закономірностей формування літературних трендів та уточнення методологічних підходів до прогнозування їхньої динаміки.

Методи. Проаналізовано 280 книжок із Goodreads Choice Awards 2024, 225 видань зі списку Amazon Best Books 2024 і 949 книжок рейтингу NYT Bestsellers. Дані збирали Python-скриптами (Selenium, BeautifulSoup), очищали та нормалізували (pandas, numpy). Використано статистичний аналіз, PCA і кластеризацію K-Means для порівняння читацьких, комерційних і медійних показників.

Результати. Аналіз трьох джерел виявив виразні закономірності та спільні тенденції. На рівні читацьких уподобань (Goodreads) найширшу аудиторію зберігають жанри Romance і Fiction, тоді як History та Nonfiction стабільно демонструють високі рейтинги. Окремо виділяються "приховані перлини" – жанри World War II, Mental Health і Health, які поєднують високу оцінку з невеликим охопленням і відкривають можливості для цільового просування.

У комерційному сегменті (Amazon) переважають паперові формати Hardcover і Paperback, а Kindle Edition залишається найдоступнішим за ціною. Преміальні ніші, як-от Epic Fantasy та регіональна художня проза, утримують найвищі середні ціни, що підтверджує готовність читачів інвестувати у винятковий контент. Кореляційний аналіз засвідчив відсутність прямого зв'язку між вищою ціною і кращим рейтингом, проте показав помірну залежність між кількістю доступних форматів і загальною вартістю.

Медійна видимість (NYT Bestsellers) свідчить, що тривалий інтерес публіки забезпечують дитячі серії та нехудожня література: окремі видання залишаються у списку понад 300 тижнів. Сезонність дебютів має два піки – у першому та третьому кварталах, що підвищує шанси книжок на довготривалу присутність у рейтингах.

У підсумку лише вісім видань одночасно досягли високих показників за читацькими, ринковими та медійними критеріями, демонструючи збалансований успіх і найбільший потенціал для перекладу та міжнародного просування.

Висновки. Інтегрований аналіз трьох джерел показав, що комерційний успіх книжки не визначається одним фактором: потрібне поєднання стійкої читацької уваги, збалансованої цінової політики та здатності довго залишатися у публічному полі. Таке комплексне бачення дає змогу видавцям планувати портфель не лише за жанровими пріоритетами, а й за довгостроковою економічною доцільністю. Важливо враховувати й часові коливання попиту, адже ринкові цикли впливають на ефективність маркетингових і перекладацьких кампаній. Зіставлення читацьких, комерційних і медійних критеріїв створює основу для точнішого прогнозування та мінімізує ризик одноразових успіхів. У підсумку отримані дані формують методологічну базу, яка допомагає приймати стратегічні рішення щодо перекладу, інвестицій і розвитку міжнародних проєктів.

Ключові слова: інтегрований аналіз даних, кроссплатформне порівняння, відкриті дані, алгоритмічний аналіз, кластеризація, моделювання читацької поведінки.

Вступ

Книжковий ринок є важливим елементом культурного й економічного життя суспільства. З розвитком цифрових технологій змінюються формати і способи поширення літератури, однак інтерес до книг залишається стабільним, трансформуючись відповідно до сучасних умов. Для видавництва, що орієнтуються на якісну перекладну літературу, важливо вчасно розпізнавати актуальні літературні тренди, які забезпечують не лише комерційний успіх, але й тривалу читацьку підтримку. Важливим інструментом у цьому процесі є аналіз відкритих даних, що дає змогу системно оцінювати читацькі вподобання, ринкові тенденції та медійну видимість книжок.

Метою дослідження є створення й обґрунтування інтегрованої моделі мультиджерельного аналізу читацьких, комерційних і медійних показників книжкового ринку, що забезпечує виявлення закономірностей формування літературних трендів та уточнення методологічних підходів до прогнозування їхньої динаміки.

Огляд літератури. Goodreads і Amazon давно вважають багатим джерелом аналітики. У роботі (Kousha, Thelwall, & Abdoli, 2017) показано, що кількість рецензій і оцінок на Goodreads корелює з подальшим впливом книги,



проте дослідження не враховувало жанрову неоднорідність аудиторії. Подальші дослідження (Maity, Panigrahi, & Mukherjee, 2018) свідчать, що з точністю 89 % можна спрогнозувати Amazon-бестселери на основі поведінкових ознак із Goodreads (додавання книг на полиці, участь в обговореннях), проте представлена модель не аналізувала зміст відгуків і не охоплювала медійну видимість. Також показано, що середній рейтинг і кількість рецензій є основними драйверами потрапляння в Amazon-бестселери в період з 2009 по 2019 р., але не враховувались цінні та жанрові фактори. У 2020 р. проаналізовано сезонні "вікна" дебютів і тривалість перебування художніх творів у NYT у період з 1931 по 2020 р., однак дослідження стосувалося лише твердої палітурки й не враховує змін критеріїв формування рейтингу (напр., перехід до врахування електронних продажів і нових категорій) (Pruett, 2020). В роботі (Kousha, & Thelwall, 2015) показано слабкі кореляції ($r \approx 0.2-0.4$) між Amazon-відгуками й науковими цитуваннями академічних монографій і бестселерів, проте ця робота обмежилася тільки двома жанровими підмножинами.

Попередні роботи доводять цінність кожного джерела, але їх вузькі вибірки, різні методики нормалізації й агрегації не дають цілісної картини. Отже, огляд публікацій зазначеної тематики показав відсутність єдиного підходу, здатного одночасно врахувати жанрово-форматну різноманітність, узгодити шкали оцінок і перевірити чутливість вагових коефіцієнтів у різних контекстах. Саме ці прогалини об'єднують ідею мультиджерельного аналізу: інтеграції Goodreads Choice Awards 2024, Amazon Best Books 2024 і NYT Bestsellers 2024 в єдину модель зі спільними метриками. Такий підхід не лише дає змогу зіставити емоційну, ринкову й медійну складові популярності, а й підвищує надійність рекомендацій для видавничих стратегій, адже базується на збалансованих даних трьох незалежних платформ.

Методи

Дослідження інтегрує три незалежні джерела, що репрезентують читацький, комерційний і медійний виміри популярності книжок. Використано 280 позицій фіналістів Readers' Favorite Books зі списку Goodreads Choice Awards 2024, 2 122 записів про формати та ціни для 225 унікальних видань зі списку Amazon Best Books 2024, а також 949 книжок, що входили до щотижневих рейтингів The New York Times Bestsellers 2024 р.

Для роботи з даними створено набір програмних інструментів на основі мови Python, які охоплюють повний аналітичний цикл: від збору й очищення – до глибокого оброблення, кластеризації й побудови візуалізацій. Збір даних здійснювався Python-скриптами з використанням Selenium, BeautifulSoup і requests для охоплення динамічних і статичних сторінок, збереження "сирих" даних у форматі CSV. Етап підготовки даних включав очищення текстових полів, нормалізацію жанрів і форматів та усунення дублікатів за допомогою pandas, numpy і networkx. Візуалізації та розподіли будували у matplotlib і seaborn. Для зниження розмірності застосовували StandardScaler і PCA, для сегментації – KMeans. Оцінювання кореляцій проводили за допомогою scipy.stats.pearsonr. Кожну платформу аналізували окремо, а потім результати зіставляли для виявлення збігів і розбіжностей між читацькою, ринковою та медійною оцінкою, після чого була побудована уніфікована модель для рекомендацій видавничих рішень.

На рівні Goodreads для кожної книги k зібрано назву, автора, середній рейтинг $r_k \in [1.0; 5.0]$, кількість оцінок $c_k \in N_0$ та жанрові теги, перетворені на бінарний вектор $\vec{g}_k \in \{0,1\}^{20}$. Після очищення даних і нормалізації жанрів ця структура дозволяє розглядати книжку як точку в тривимірному просторі "якість – популярність – жанрова структура". Для жанру j обчислювали середній рейтинг і середнє охоплення за формулами відповідно

$$\bar{r}_j = \frac{1}{|K_j|} \sum_{k \in K_j} r_k \text{ та } \bar{c}_j = \frac{1}{|K_j|} \sum_{k \in K_j} c_k. \quad (1)$$

Далі кожному жанру присвоювали три ранги: за потужністю $rank_{|K|}(j)$, за середнім рейтингом $rank_r(j)$ і за середнім охопленням $rank_c(j)$. Сума рангів утворює індекс трендовості

$$T_j = rank_{|K|}(j) + rank_r(j) + rank_c(j). \quad (2)$$

Щоб виявити, які жанрові поєднання найчастіше трапляються в межах однієї книжки і створюють "вузли" читацької уваги, розраховано "взаємопроникнення" жанрів. Цей показник вимірює, скільки разів пара жанрів i та j одночасно присутня в описі тієї самої книжки

$$s_{ij} = \sum_{k=1}^K g_{ki} \cdot g_{kj}. \quad (3)$$

Для виявлення прихованої (латентної) структури застосовано зменшення розмірності PCA з подальшою кластеризацією KMeans із параметром $K = 4$, і з мінімізацією суми квадратів відстаней до центрів:

$$V = \sum_{i=1}^K \sum_{x \in c_i} \|x - \mu_i\|^2. \quad (4)$$

Стабільність жанру оцінювалась як частка "дуже високих" проти суми "дуже високих" і "дуже низьких" оцінок за формулою

$$\text{Стабільність}_j = \frac{|K_j^{(r>4.4)}|}{|K_j^{(r>4.4)}| + |K_j^{(r<2.5)}|}. \quad (5)$$

Для жанрів із невеликою кількістю книжок, менше ніж п'ятнадцять, введено два індекси. Нішевий індекс

$$N_j = \frac{\bar{r}_j}{5} + \frac{\bar{c}_j}{\max_j \bar{c}_j} \quad (6)$$

та індекс перспективності

$$P_j = \frac{\bar{r}_j}{5} + \left(1 - \frac{\bar{c}_j}{\max_j \bar{c}_j}\right), \quad (7)$$

який виділяє "приховані перлини". Зв'язок "якість проти охоплення" перевірено коефіцієнтом Пірсона

$$r = \frac{\text{Cov}(r_k, c_k)}{\sigma_{r_k} \cdot \sigma_{c_k}}. \quad (8)$$

Для Amazon кожний запис містив назву, автора, дату першої публікації, середній рейтинг r_k , формат f (Hardcover, Paperback, Kindle Edition, Audiobook тощо), ціну p , тип ціни (нова, вживана, електронна тощо), основну категорію та підкатегорію. Такий підхід дозволяє описати не обсяги продажів, а структуру ринкової пропозиції: фізичні носії – їхню логістику, цифрові – жанрові сегменти, ціна – платоспроможний попит, рейтинг – сприйняття якості. Середні ціни та стандартні відхилення за форматом і типом визначено як



$$\mu_p^f = \frac{1}{n_{fp}} \sum_{j=1}^{n_{fp}} p_{fp}^{(j)} \quad (9)$$

та

$$\sigma_p^f = \sqrt{\frac{1}{n_{fp}} \sum_{j=1}^{n_{fp}} (p_{fp}^{(j)} - \mu_p^f)^2}. \quad (10)$$

Також обчислено кореляції між середньою ціною μ_p^f , середнім рейтингом μ_p^k та кількістю форматів n_f^k за формулою (8). Для порівняння "ціни проти якості" застосовувалося індекс value-for-money

$$V_k = \frac{\mu_p^k}{\mu_p^f}. \quad (11)$$

У просторі ознак $(\mu_p^k, \mu_p^f, n_f^k)$ виконано PCA з подальшою кластеризацією KMeans ($K = 5$).

Для NYT Bestsellers опис кожної книги включав тривалість присутності у списку w_k , середню позицію p_k , квартал дебюту $q_k \in \{Q_1, Q_2, Q_3, Q_4\}$, автора a_k і широту авторського портфеля s_a , що дорівнює кількості категорій, у яких автор був представлений. Для підмножини long-lived L_j , тобто книг із перебуванням у рейтингу не менше десяти тижнів, на рівні категорій оцінено середню тривалість

$$\bar{\omega}_j = \frac{1}{|L_j|} \sum_{k \in L_j} \omega_k, \quad (12)$$

і середню позицію

$$\bar{p}_j = \frac{1}{|L_j|} \sum_{k \in L_j} p_k. \quad (13)$$

Приховані (латентні) структури на рівнях категорій, авторів і книг отримано через PCA і KMeans у векторах (ω_k, p_k, s_a) .

Для інтеграції трьох аспектів успіху сформовано вибірку зі 126 творів, для яких доступні якісні показники на всіх платформах. Кожну книгу описано семивимірним вектором нормалізованих ознак:

$$\vec{x}_k = (\tilde{r}_k^{GR}, \tilde{c}_k^{GR}, \tilde{r}_k^{AM}, \tilde{c}_k^{AM}, \tilde{p}_k^{AM}, \tilde{w}_k^{NYT}, p\tilde{o}s_k^{NYT}). \quad (14)$$

Нормалізація виконувалася min-max для приблизно рівномірних розподілів. Для експоненційно скошених змінних, наприклад, кількість оцінок і ціна, застосовувалося логарифмічне перетворення. Для метрик, де менше означає краще, таких як позиція NYT, використано інверсну нормалізацію. Ваги блоків для читачього, ринкового та медійного вимірів визначено пропорційно до дисперсій:

$$D_x = \frac{1}{|s|} \sum (x_k - \bar{x})^2 \quad (15)$$

та

$$\omega_i = \frac{D_{x_i}}{D_{GR} + D_{AM} + D_{NYT}} \text{ за умови } \sum_i \omega_i = 1. \quad (16)$$

Отримано $\omega_{GR} = 0.29$, $\omega_{AM} = 0.37$, $\omega_{NYT} = 0.34$. Підсумковий індекс видавничої доцільності обчислено як

$$Total_Score_k = 0.29 \cdot \tilde{r}_k^{GR} + 0.37 \cdot \tilde{r}_k^{AM} + 0.34 \cdot \tilde{w}_k^{NYT}. \quad (17)$$

Додатково враховано показник Source Rank, що дорівнює кількості платформ, на яких присутня книга. В усіх обчисленнях використовувалася десяткова крапка і безперервна нумерація формул.

Результати

Розподіли Goodreads виявили виразну асиметрію середніх рейтингів із піком у діапазоні від 3.8 до 4.2, причому значення $\tilde{r}_k$ змінювалися від 3.43 до 4.63. Кількість оцінок варіювала від 630 до 1.1 млн, більшість книжок мали менше ніж 30 000 відгуків, що засвідчує нерівномірну увагу аудиторії та обґрунтовує потребу комбінованих показників, а не поодиноких метрик (рис. 1 і 2).

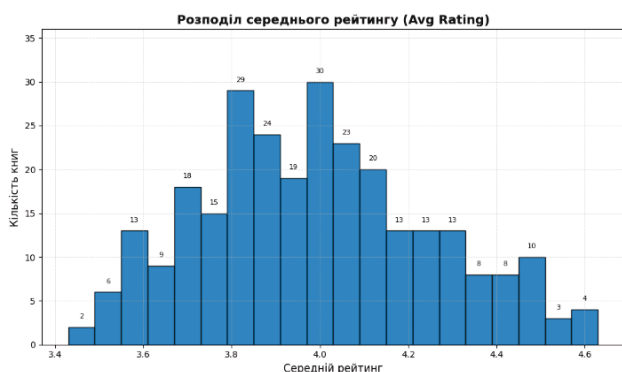


Рис. 1. Goodreads: розподіл середнього рейтингу

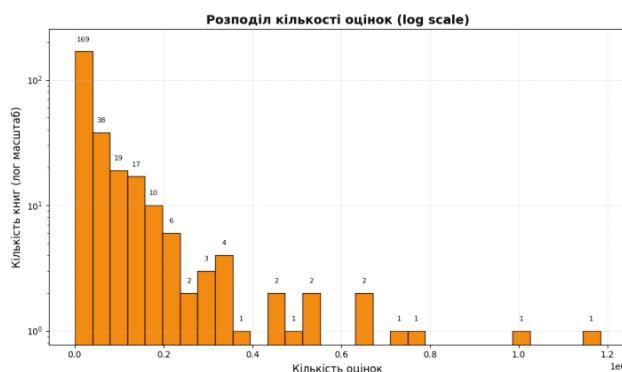


Рис. 2. Goodreads: розподіл кількості оцінок

Порівняння жанрів через індекс T_j вивело у лідери Book Club, Romantasy, Romance, Sports Romance та Audiobook, (рис. 3). Це означає, що поєднання масштабу вибірки, середніх оцінок і середнього охоплення формує стабільний попит саме в цих напрямках.

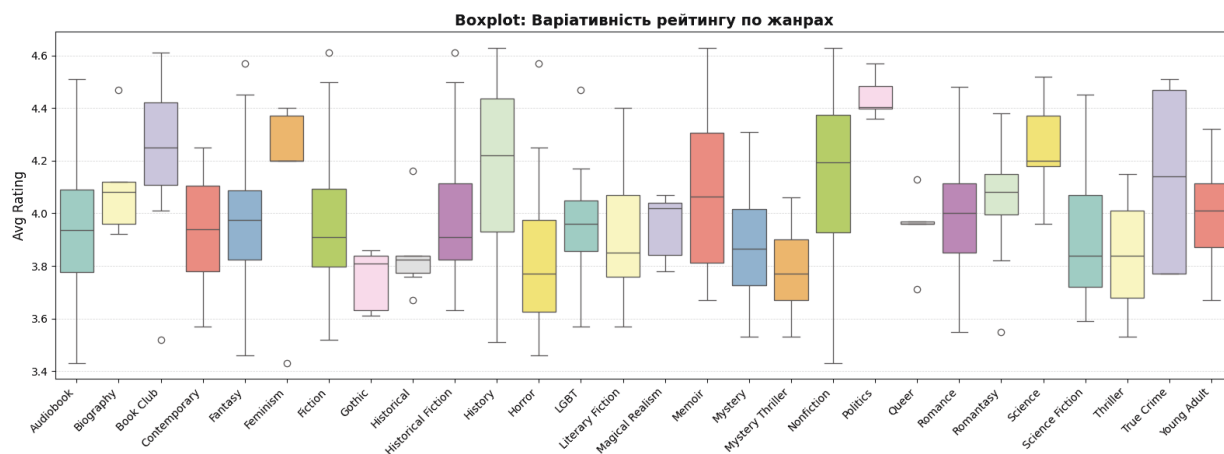


Рис. 3. Goodreads: варіативність рейтингу по жанрах

Матриця "взаємопроникнення" зафіксувала типові вузли читацьких уподобань, найчастіше разом траплялися пари Romance + Fantasy зі значенням 38, Mystery + Nonfiction зі значенням 36, Fiction + Historical Fiction зі значенням 32, а також Fiction + Fantasy зі значенням 20 (рис. 4). Такі комбінації відображають не лише тематичну близькість жанрів. Вони показують сталі очікування читача щодо набору рис твору. Наприклад, зв'язка Romance + Fantasy зазвичай означає магічний світ, емоційну лінію стосунків і високу ставку для героїв. Це корисно для планування міжжанрових лінійок і рекомендацій.

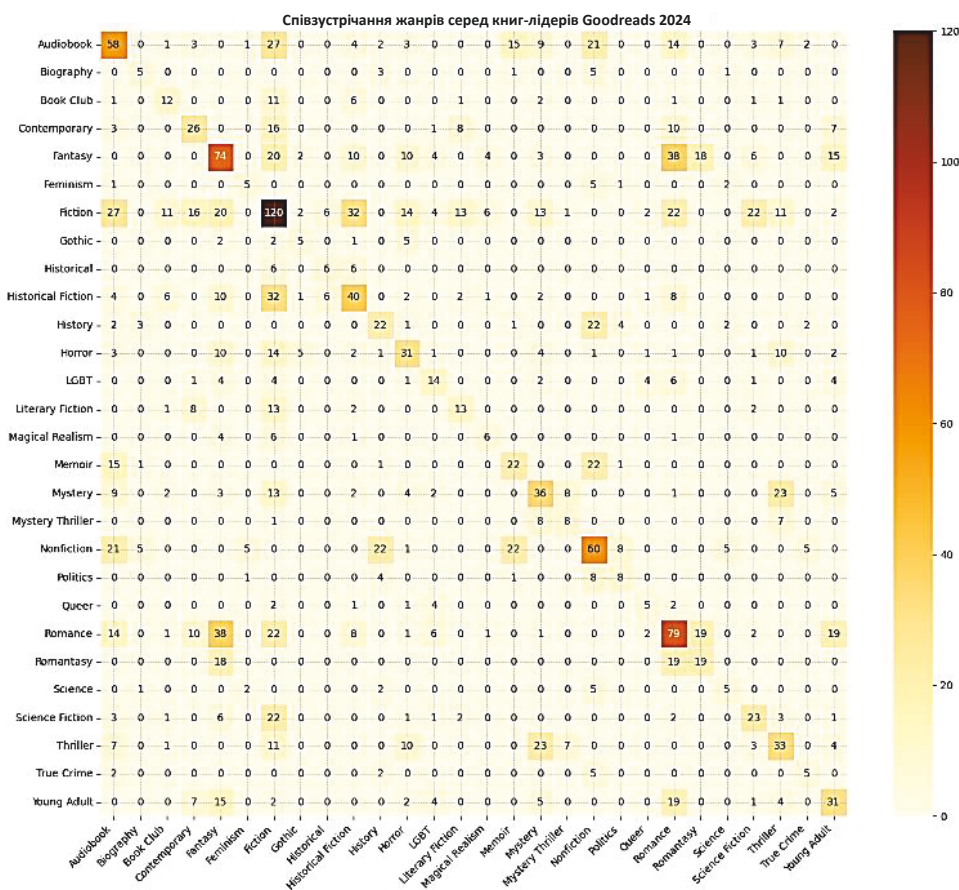


Рис. 4. Goodreads: теплова карта "співзустрічання"

Зменшення розмірності та кластеризація підтвердили сегментацію читацького поля на чотири групи. До групи складної прози увійшли History і Memoir із високими рейтингами та помірною популярністю. До емоційних жанрів належать Romance і Thriller із великими обсягами оцінок. Камерні напрями представлені Book Club і Small Town Romance з невеликою кількістю назв і стабільними показниками. Нішеві жанри, зокрема Paranormal і Poetry, мають невелику аудиторію та високі середні оцінки (рис. 5). Окрема кластеризація книг за векторами жанрової належності



дала п'ять архетипів читачьких уподобань, від преміальної прози до літнього бестселера, що дозволяє оперувати не лише жанрами, а стилістичними профілями.

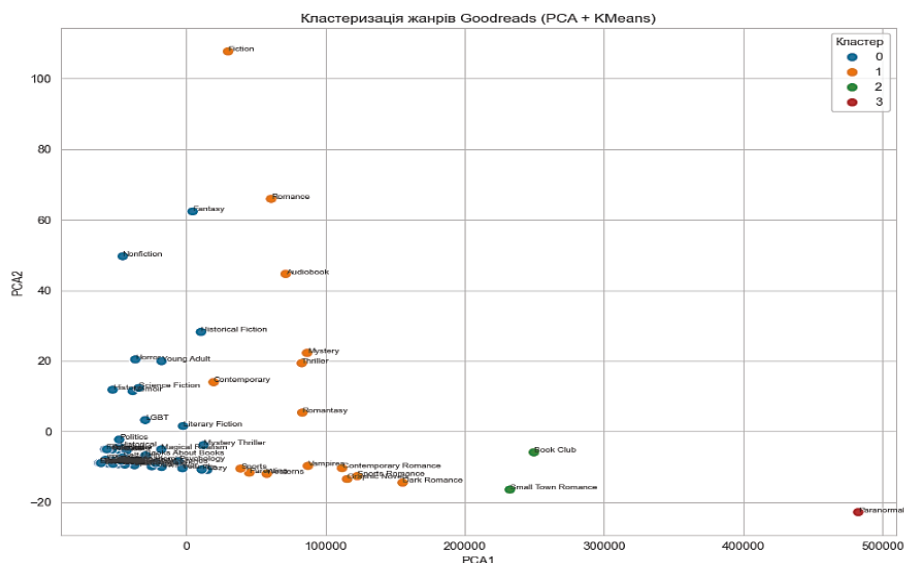


Рис. 5. Goodreads: кластеризація жанрів

Стабільність жанрів виявилася нерівномірною і це добре узгоджується з припущенням про різні механіки формування очікувань читачів у nonfictional і fictional напрямках. Найвищі значення спостерігалися у History зі значенням 0.64, у Nonfiction зі значенням 0.55. У цих сегментах автори спираються на верифіковані події й дати, архівні матеріали, статистику, наукові публікації та систему приміток і посилань. Фактчекінг і чіткі критерії точності роблять сприйняття більш передбачуваним, тому оцінки частіше концентруються у "зеленій зоні". Historical Fiction із 0.36 поводить як проміжна ланка між суворим Nonfiction і суто художньою прозою. Найнижчі значення були у Audiobook зі значенням 0.17 та у Fiction зі значенням 0.14 (рис. 6). Таке падіння очікуване з огляду на сильну різноманітність піджанрів і роль актора дубляжу у випадку аудіокниг, де інтерпретація тексту може як підняти, так і "згасити" сприйняття якості. Ризиковий профіль за нашою метрикою дає ще виразніший контраст. Mystery і Thriller мають показник 1.00. Це означає крайній розподіл з переважанням дуже високих або дуже низьких оцінок.

Для видавця це сигнал про "лотерейну" природу релізів у цих піджанрах. Невелика похибка у позиціюванні обкладинки, у темпі сюжету чи у фінальному твісті здатна різко зрушити сприйняття аудиторії в будь-який бік. Далі за рівнем ризику розміщуються Horror зі значенням 0.94, Romance зі значенням 0.89 і Fantasy зі значенням 0.89. Такий результат частково неочікуваний щодо Romance, який ми звикли вважати "масовим і безпечним", однак, зважаючи на "ефект завищених очікувань", абсолютно логічний і пояснюваний: коли читач іде на сильну емоційну обіцянку, розчарування карається різкіше, ніж у "спокійних" піджанрах. Це пояснює високий ризик за загальною позитивною тональністю ринку (рис. 7).

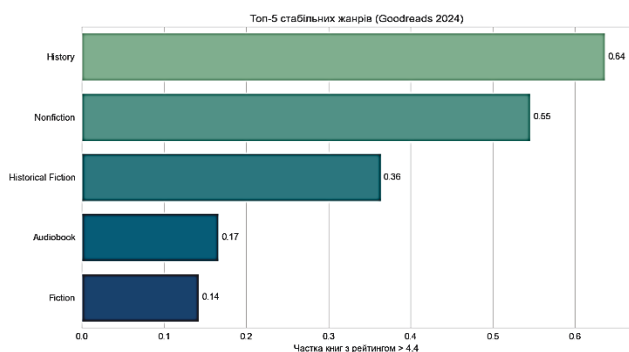


Рис. 6. Goodreads: п'ять лідерів стабільних жанрів

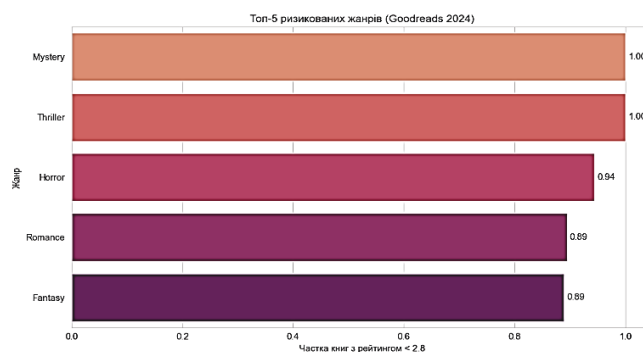


Рис. 7. Goodreads: п'ять лідерів ризикових жанрів

Для малих вибірок ми свідомо розділили на дві різні метрики нішевість і перспективність. Нішевий індекс N_j виділяє напрями з поєднанням добрих оцінок і відчутного, але не гігантського охоплення (жанри з менше ніж 15 книжками). Лідирують Paranormal зі значенням 1.8, Book Club зі значенням 1.4 і Small Town Romance зі значенням 1.3 (рис. 8). Це нішеві, але комерційно життєздатні сегменти з уже сформованим ядром аудиторії. Вони не масові, проте мають достатній попит, щоб підтримувати серійність і окупність, а за правильної стратегії можуть масштабуватися.

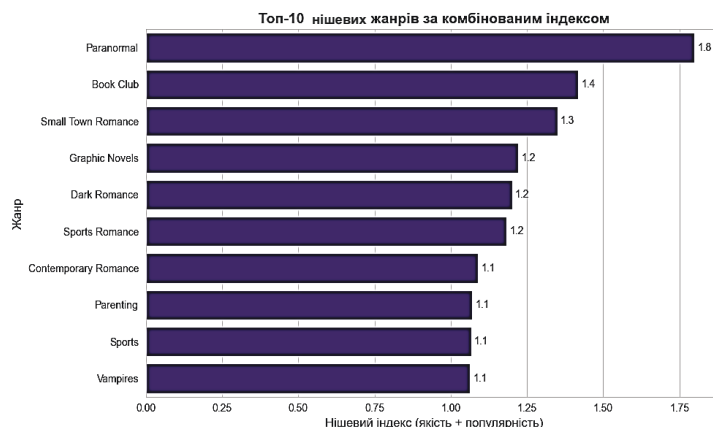


Рис. 8. Goodreads: десять лідерів нішевих жанрів за комбінованим індексом

Індекс перспективності P_j навмисно "підсвічує" те, що читачі вже високо оцінюють, але ще не встигли масово помітити. Тут у топі жанри World War II, Mental Health і Health. Вони всі вищі 4.0 за рейтингом і мають низьке охоплення – менше ніж 5 000 оцінок. Це класичні "приховані перлини", де додатковий маркетинговий імпульс і точне позиціонування можуть забезпечити тривале рівне споживання після стартового піку. Поділ є інструментальним для портфельних рішень. Нішеві, але життєздатні жанри слід масштабувати через серійність і розширення каналів. Жанри з високими оцінками та малим охопленням доцільно підтримувати фокусним промо і тестами позиціонування.

Окремо перевірено, чи можна спиратися на популярність як на проксі якості. Кореляція Пірсона, між середнім рейтингом і кількістю оцінок становить $r = 0.18$ при $p = 0.002$. Після логарифмування кількості оцінок тренд зберігається як слабкий додатний. Перевірка ранговою кореляцією також показує слабку залежність. Вилучення поодиноких точок з екстремально великим охопленням не змінює висновку про низьку силу зв'язку (рис. 9–10). Структура хмари точок підтверджує статистику. У діапазоні від 10^3 до 10^5 оцінок спостерігається найбільша концентрація творів із рейтингами приблизно 3.8–4.2. За дуже великих охоплень варіативність рейтингів зростає і більшість значень перебуває в інтервалі 4.0–4.3. Водночас присутні видання з рейтингами понад 4.4 у групах із невеликим і середнім охопленням. Є і протилежні випадки, коли твори з великою кількістю оцінок мають лише середні бали. Зв'язок нелінійний і гетероскедастичний. На лінійній шкалі видно вертикальну "стіну" спостережень для малих охоплень, а на лог-шкалі – слабкий нахил без вираженого зростання середніх значень. Окремі високовпливові точки на рівні близько мільйона оцінок не задають загального тренду. Межі шкали 1–5 створюють стелю близько 4.6, що додатково обмежує потенційний приріст рейтингу разом із ростом охоплення, Такий результат очікуваний, адже популярність іде хвилями і часто залежить від інформаційних приводів. На відміну від неї якість змінюється повільніше і може бути високою навіть у видань із невеликим охопленням. Тому вірусні сплески не замінюють системного розвитку лінійок і портфельного супроводу.

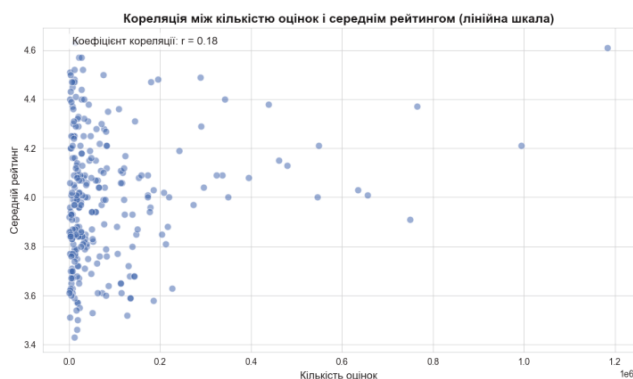


Рис. 9. Goodreads: кореляція між кількістю оцінок і середнім рейтингом (лінійна шкала)

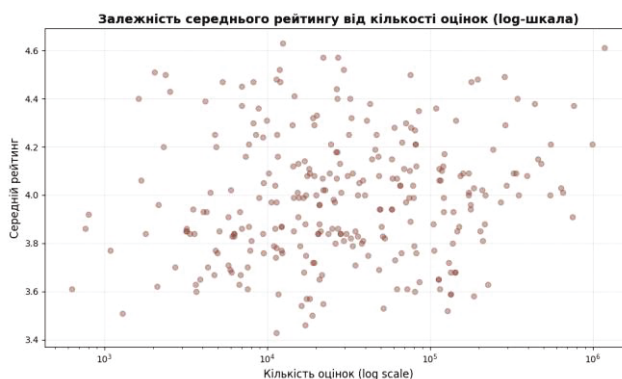


Рис. 10. Goodreads: залежність середнього рейтингу від кількості оцінок (логарифмічна шкала)

Комерційні дані Amazon підкреслили роль носіїв і цінових ніш. У вибірці домінують паперові формати із частками 36 % для Hardcover і 32 % для Paperback, тоді як Kindle має 10 %, а Audiobook 9 %. Це очікувано для підбірки "кращих книг року", де відбір зумовлює сильну присутність преміальних паперових релізів. Цифрові носії частіше трапляються у бізнесі та Nonfiction. Художня проза і біографії переважають на папері. Лише 8 % позицій виходять одразу у трьох і більше форматах. Найтипівша зв'язка – Hardcover разом із Paperback. Така структура підтверджує роль паперу як "якірного" каналу релізу і роль Kindle як каналу доступності. Узагальнені пропорції та перетини форматів згруповано на рис. 11–12.

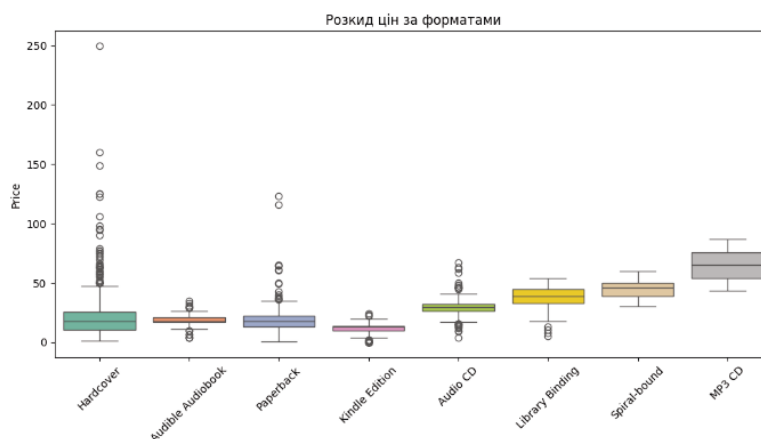


Рис. 11. Amazon: розкид цін за форматами

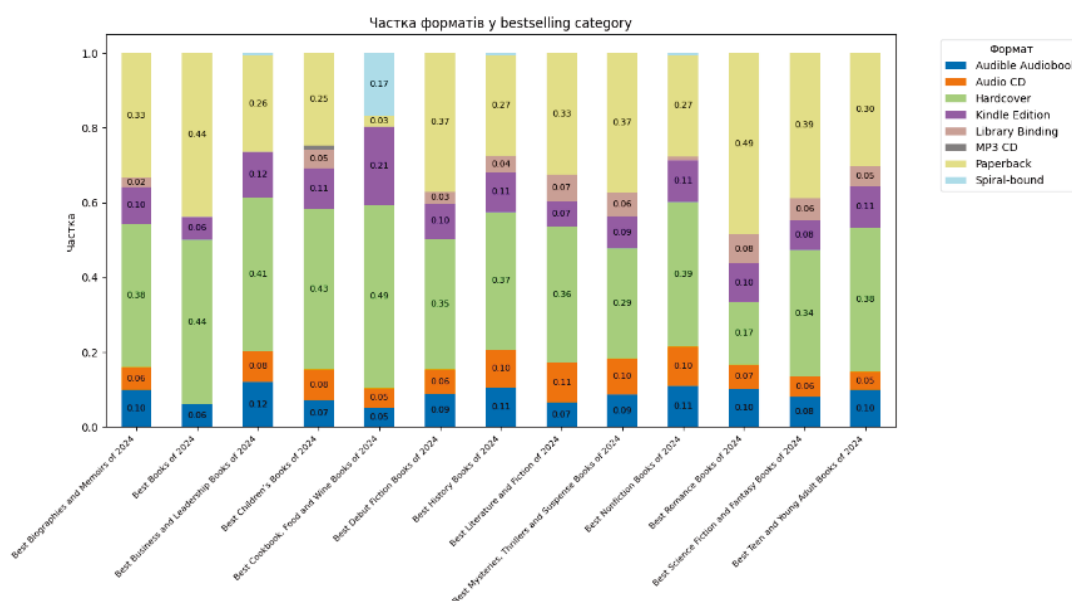


Рис. 12. Amazon: форматна насиченість книжкового каталогу

Розподіл цін "скошений праворуч" і має "довгий верхній хвіст". Медіана дорівнює 17.98 \$. Мінімум дорівнює 2.99 \$. Максимум дорівнює 249.52 \$. Середні значення формують "щаблі" між форматами: Kindle Edition ≈ 11.13 \$ < Paperback ≈ 18.01 \$ $\approx$ Audible Audiobook ≈ 18.34 \$ < Hardcover ≈ 21.49 \$ < Audio CD ≈ 29.91 \$ < Library Binding ≈ 37.19 \$ < Spiral-bound ≈ 44.52 \$ < MP3 CD ≈ 65.07 \$ (рис. 13–14). Асиметрія пояснюється тим, що у верхньому хвості зібрано колекційні та спеціальні видання, й окремі великі наклади з підвищеною собівартістю. Для Kindle діапазон цін вузький, що узгоджується з більш жорсткою ціновою межею електронних видань. Карту цін за форматами і типами продажу наведено у групі на рис. 13–14.

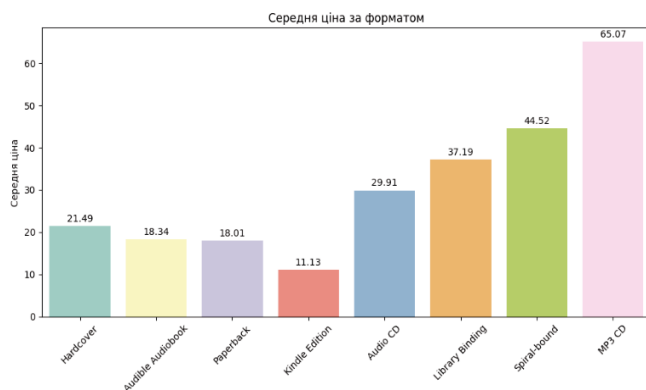


Рис. 13. Amazon: boxplot розкид цін за форматами

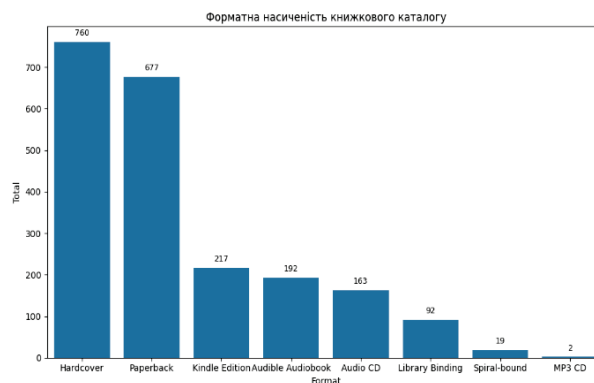


Рис. 14. Amazon: середня ціна за форматами



Профіль підкатегорій додатково розрізняє "дорогі" ніші та "бюджетне ядро". У межах bestselling-категорій найвищу середню ціну має Cookbook, Food & Wine Books – 26.07 \$. Далі йдуть блоки з вищим за середній рівнем: Nonfiction – 23.00 \$, Science Fiction and Fantasy – 22.59 \$, Literature and Fiction – 22.54 \$, History – 22.04 \$. Середній пояс формують Debut Fiction – 21.75 \$, Biographies and Memoirs – 20.95 \$, Business and Leadership – 20.73 \$, Mysteries, Thrillers and Suspense – 19.39 \$. Найнижчі значення спостерігаються у Romance – 18.20 \$ та Teen and YA – 18.43 \$, а Children's Books у цій вибірці опускається до 13.38 \$ (рис. 15). Отже, на рівні широких категорій дорожчими є напрями, близькі до Nonfiction і біографій, тоді як Romance, Teen / YA і частина дитячого сегменту тримають нижчі середні ціни. Це доповнює попередній зріз за підкатегоріями: там лідирують окремі преміальні ніші на кшталт Epic Fantasy та Literary Fiction LGBTQ+, а тут видно, що у грубішій агрегації найвищі середні утримують Biographies and Memoirs.

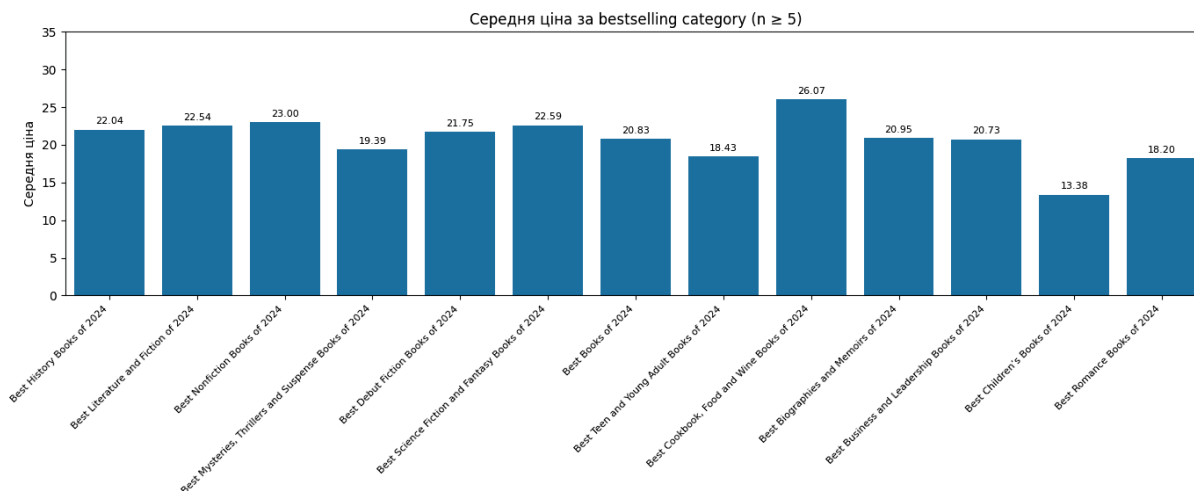


Рис. 15. Amazon: середня ціна за bestselling category

Перевірка зв'язків між основними показниками дає однозначну картину. Кореляція між ціною і рейтингом практично відсутня. Значення коефіцієнта дорівнює 0.06. Отже вищий прайс не супроводжується кращою середньою оцінкою. Кореляція між ціною і кількістю форматів слабка додатна. Значення коефіцієнта дорівнює 0.24. Це свідчить, що продукти із ширшою форматною лінійкою мають тенденцію до вищої середньої ціни. Кореляція між кількістю форматів і рейтингом слабка від'ємна. Значення коефіцієнта дорівнює мінус 0.14. Тобто розширення форматної палітри не пов'язане з підвищенням середнього рейтингу. Сумарні діаграми розкиду та коефіцієнти наведено на рис. 16–17.

Окремо оцінено співвідношення "ціна – оцінка" через індекс value for money. Для кожної книги обчислено відношення середнього рейтингу до середньої ціни. Найвище значення має Eyes That Kiss in the Corners у форматі Paperback. Значення дорівнює 9.06. Порівняння лідерів і "середняків" цього індексу показує, що окремі паперові позиції у популярних піджанрах формують найкращі відношення вартість проти оцінки.

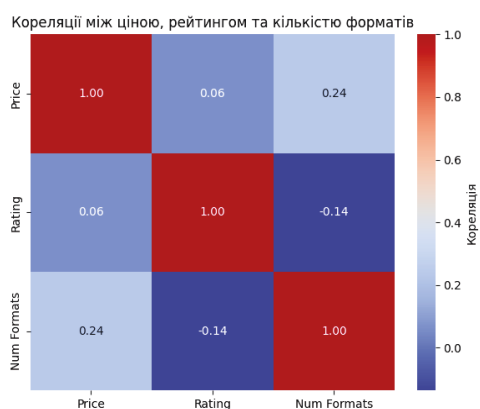


Рис. 16. Amazon: середня ціна за bestselling category

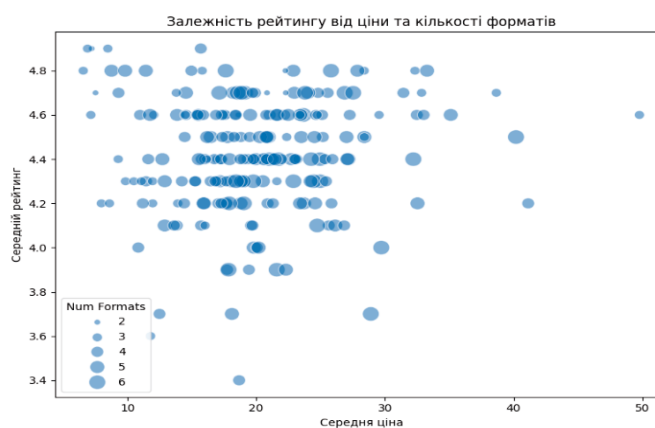


Рис. 17. Amazon: залежність рейтингу від ціни та кількості форматів

Кластеризація у просторі трьох ознак – середня ціна, середній рейтинг, кількість форматів – дає п'ять стійких груп (рис. 18). Перша група відповідає масовим бюджетним виданням, низькі ціни, один або два формати, рейтинги близько чотирьох. Друга група – масові з підвищеною ціною, два або три формати, рейтинги помірні. Третя група – збалансовані продукти середнього цінового рівня з трьома форматами. Четверта група – нішеві преміальні видання з високими цінами і невеликою кількістю форматів. П'ята група – "вітринні" позиції із широкою форматною палітрою



та цінами вище середніх, де рейтинг коливається в межах від 3.8 до 4.3. Таке розшарування узгоджується з очікуванням, що ціна і кількість форматів описують передусім маркетингову та виробничу стратегії.

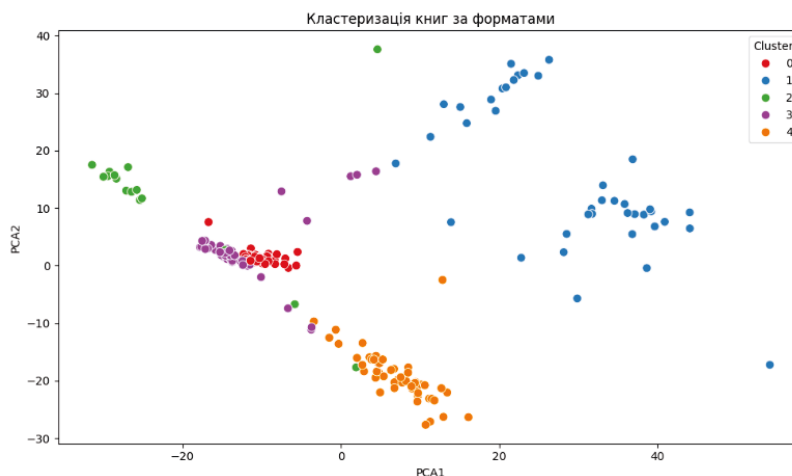


Рис. 18. Amazon: кластеризація за форматами

Медійна динаміка NYT віддзеркалила довготривалу видимість. Із 949 позицій 2024 р. у статусі long-lived залишилися 247 книжок. Найбільша частка припадає на Hardcover Nonfiction, Combined Print & E-Book Fiction та Hardcover Fiction (рис. 19). Це очікувано, бо саме ці три "опорні" таблиці NYT регулярно акумулюють найпотужніші релізи і мають стабільний попит у ритейлі та бібліотеках. Важливо, що різниця між ними невелика, отже тривала присутність у списку не є прерогативою лише одного сегменту і відображає загальний баланс між художньою і нехудожньою пропозиціями.

Порівняння середніх тривалостей і позицій для long-lived категорій демонструє різку асиметрію на користь серій для дітей і підлітків (рис. 20). Children's & Young Adult Series мають близько 307 тижнів середньої тривалості. Це узгоджується з логікою серійного споживання, коли кожна нова частина підтягує продажі попередніх, а також зі шкільними закупівлями і повторними перевиданнями. Такий контраст типовий для ринку, де "серійні" бренди забезпечують довге плато видимості, а одиночні романи частіше отримують хвилеподібну увагу. Для Hardcover Fiction і Combined Print & E-Book Fiction картина інша. Тут тривалість присутності коротша, а розкид позицій ширший. Більша частка релізів поводить як хвиля з яскравим стартом і швидким спадом. Hardcover Nonfiction тримається довше, ніж більшість художніх категорій, але коротше за дитячі серії. Це узгоджується з тим, що Nonfiction має триваліший життєвий цикл за рахунок тем із тривалою актуальністю. Перевидання у м'якій обкладинці та вихід оновлених видань продовжують обіг. У середніх позиціях видимі відтинки стабільності, які не характерні для більшості автономних романів. Розподіл усередині Children's Picture Books та Children's & Young Adult Series додає ще одну деталь. Дитячі книжки-картинки мають коротші хвилі окремих релізів, але часто повертаються до списку під святкові періоди і на початку навчального року. Серійні підліткові видання демонструють довші безперервні відрізки присутності. Це очікувано з огляду на відмінність каналів збуту та частоти перевидань. Додатково варто відмітити внесок бокс-сетів і омнібусів у дитячих і підліткових лінійках. Коли в серії виходить комплект або ювілейне перевидання, точка входу для нових читачів спрощується. Це підтримує довгі "поліці" присутності і підтягує середні позиції на кілька тижнів поспіль.

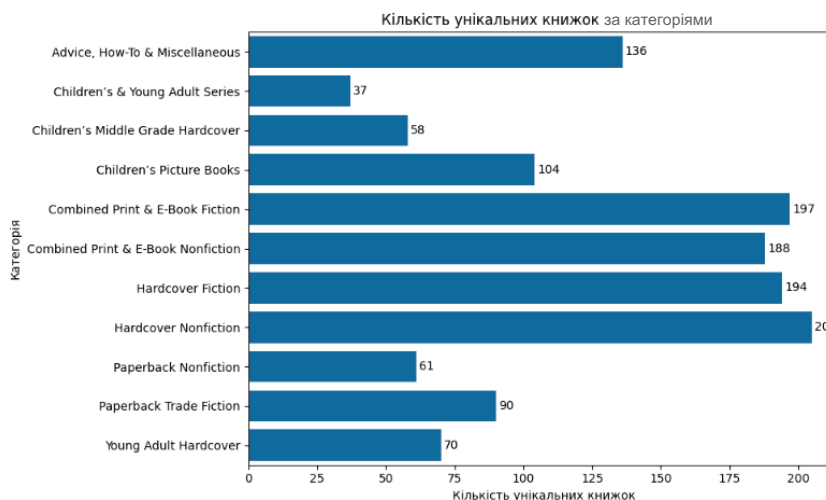


Рис. 19. NYT Bestseller: кількість унікальних книжок за категоріями

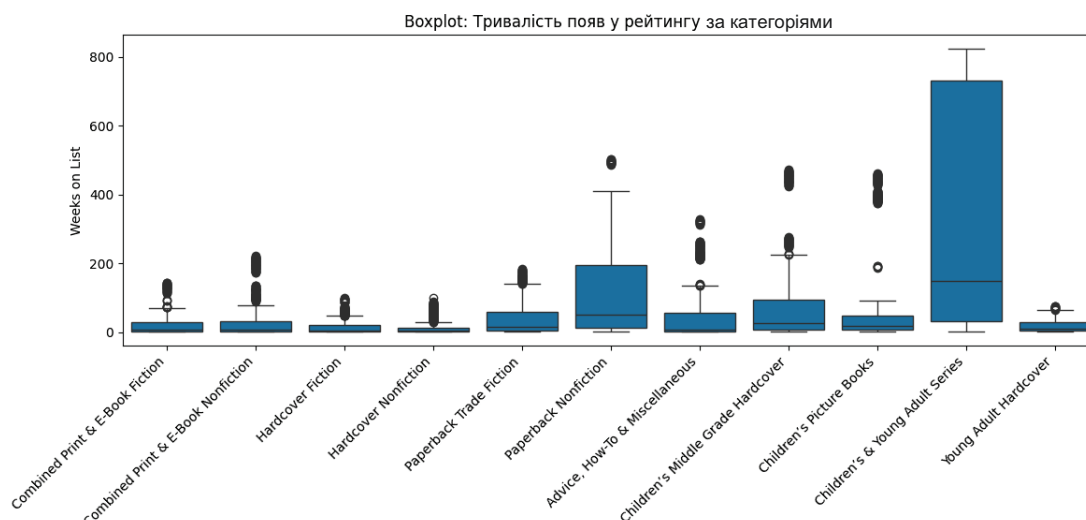


Рис. 20. NYT Bestseller: тривалість появ книжок у рейтингу за категоріями

Сезонність дебютів має два виразні піки. Найбільше успішних стартів у першому та у третьому кварталах. За тепловою картою найбільша кількість дебютів, що протрималися в рейтингах понад 20 тижнів, припала на перший квартал (рис. 21). У цей період різко зросли релізи в Children's Picture Books (16), Children's & Young Adult Series (14), Combined Print & E-Book Fiction (14) і Paperback Nonfiction (13). Другий підйом відбувся у третьому кварталі, де знову виділяється Children's Picture Books (9 дебютів) і помітно зросли показники Paperback Nonfiction (4) та кількох інших категорій. Четвертий квартал виявився спокійнішим, хоча поодинокі новинки з'являлися навіть у листопаді–грудні. Кумулятивні криві підтверджують таку динаміку (рис. 22). Children's Picture Books нарощували присутність найшвидше: від приблизно 10 позицій у січні до більш як 45 наприкінці року. Стійке, хоч і менш стрімке зростання демонстрували Paperback Nonfiction (приблизно із 17 до 28) та Combined Print & E-Book Nonfiction (з 13 до 20). Children's & Young Adult Series і Combined Print & E-Book Fiction досягли близько 20 "довгожителів", а їхні лінії на графіку помітно крутішають у серпні–вересні. Натомість Hardcover Fiction і Hardcover Nonfiction залишалися стабільно невеликими сегментами й завершили рік на рівні до 10 назв. Подібна структура випусків цілком очікувана: дитячі ілюстровані видання й серії традиційно виходять на початку року, коли плануються подарункові покупки, а комбіновані та підліткові формати активізуються ближче до осені, готуючись до святкового сезону. Nonfiction показав рівномірне нарощування впродовж усього року, що добре узгоджується з попередніми спостереженнями про стабільний попит на практичні та документальні книги.

Широта авторських портфелів різниться між блоками. Автори художньої та нехудожньої літератури частіше присутні у двох або трьох категоріях. Це очікувано для мультиформатних авторів, які виходять одночасно у твердій обкладинці, комбінованих (друк-цифра) списках і у форматах для різних аудиторій. Натомість у Children / Young Adult та "Other" домінує спеціалізація в одній категорії. Така спеціалізація природна для серій з усталеним позиціонуванням і цільовим віком.

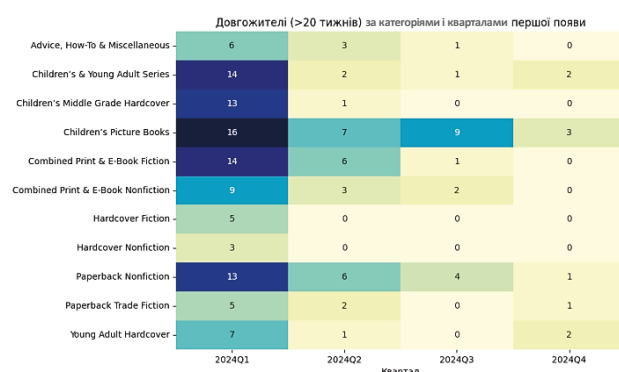


Рис. 21. NYT Bestseller: long-lived книжки за категоріями і кварталами першої появи в рейтингу

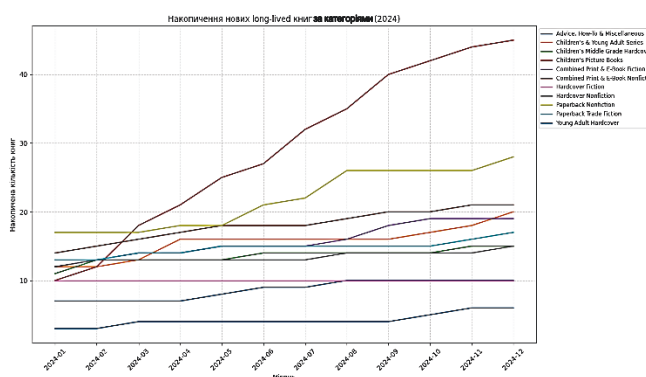


Рис. 22. NYT Bestseller: накопичення кількості long-lived книжок у ключових категоріях

Рейтинг long-lived авторів очолює Adam Wallace із шістьма релізами, далі Sarah J. Maas з п'ятьма і Colleen Hoover з чотирма (рис. 23). Це підтверджує, що стабільна видимість у NYT часто пов'язана або з активною серійністю, або з регулярними новими тайтлами в межах близького жанрового поля. Ідеться не про одиничний "постріл", а про повторюваний патерн присутності.

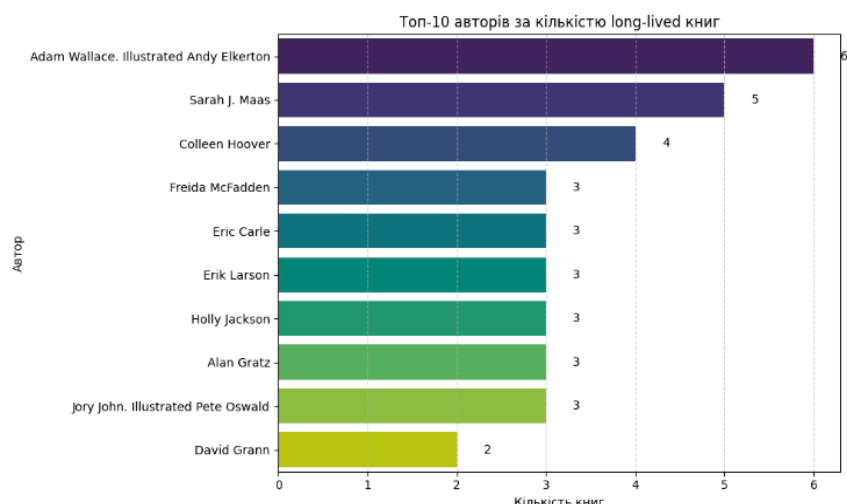


Рис. 23. NYT Bestseller: десять авторів – лідерів за кількістю long-lived книг 2024 року

Кластеризація за трійкою ознак виокремлює три типи категорій. Масові – з великою кількістю релізів і короткими періодами присутності. Збалансовані – з помірною кількістю позицій і стійкою середньою видимістю. Унікальні – з малою кількістю релізів, але з надзвичайно довгими періодами у списку. Остання група відповідає дитячим і Nonfiction -серіям, де працює ефект накопичення аудиторії. На рівні авторів виявлено вузькоспеціалізовані профілі, масових авторів із "короткоживучими" позиціями, збалансовані профілі та "лідерів довгої гри", у яких кожен реліз залишається надовго. На рівні окремих книжок виділяють чотири класи: новинки до 15 тижнів, стабільні Nonfiction -титли 15–50 тижнів, найстійкіші бестселери 100+ тижнів і сезонні явища з короткою видимістю після піку. Діаграми кластерів наведено на рис. 24–27.

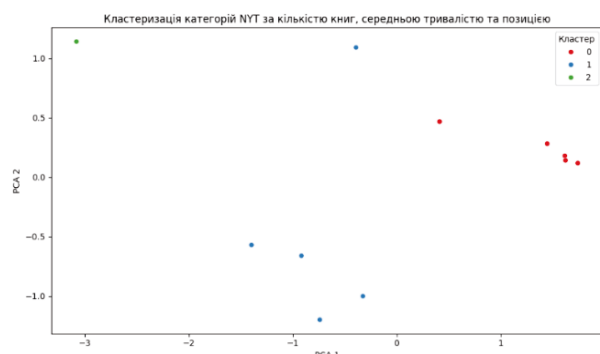


Рис. 24. NYT Bestseller: кластеризація категорій за кількістю книг, середньою тривалістю й позицією

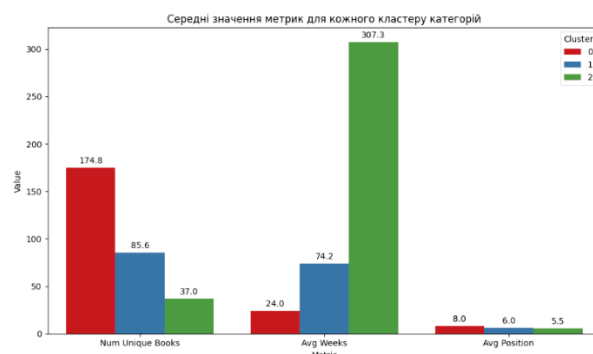


Рис. 25. NYT Bestseller: середні значення метрик для кожного кластеру категорій

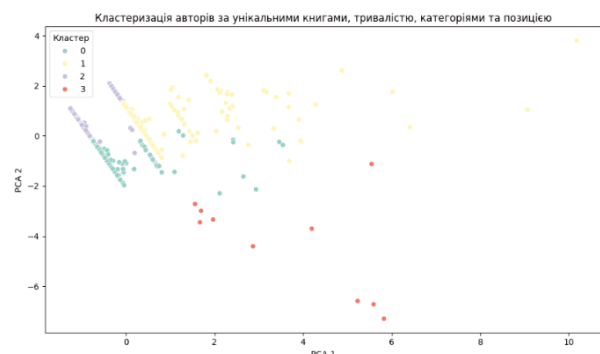


Рис. 26. NYT Bestseller: кластеризація авторів за унікальними книгами, тривалістю, категоріями й позицією



Рис. 27. NYT Bestseller: кластеризація книг за тривалістю, категоріями й позицією

Кількісний зріз long-lived дебютів показує виразну двогорбку сезонність: у першому та третьому кварталах зафіксовано по 142 релізи, тоді як другий квартал дав 118, а четвертий – 135 (рис. 28). Така різниця не випадкова, а лише підтвердила попередні отримані дані та добре вписується у типовий річний ритм книжкової індустрії, де пікові вікна



виходу припадають на початок року та період після літніх канікул, що узгоджується з відомими практиками планування великих прем'єр і маркетингових кампаній.

За сумарною тривалістю присутності лідирують довготривалі дитячі та підліткові серії, які формують "довгий хвіст" розподілу й забезпечують стаке оновлення попиту. Абсолютним рекордсменом є *Diary of a Wimpy Kid* із 822 тижнями у списку, далі *Harry Potter* із 821 тижнем і *Percy Jackson & The Olympians* із 755 тижнями (рис. 29). Такі приклади показують, що серійність і впізнаваний бренд здатні підтримувати інтерес читачів роками, створюючи ефект накопичення аудиторії й підтверджуючи стійкість франшиз як ключового драйвера довгострокової видимості в рейтингах NYT.

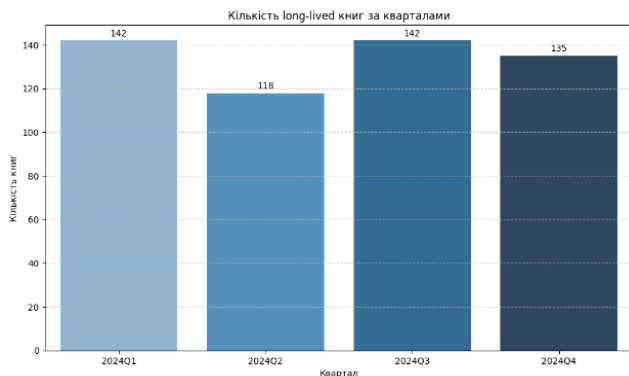


Рис. 28. NYT Bestseller: long-lived книжки за кварталами

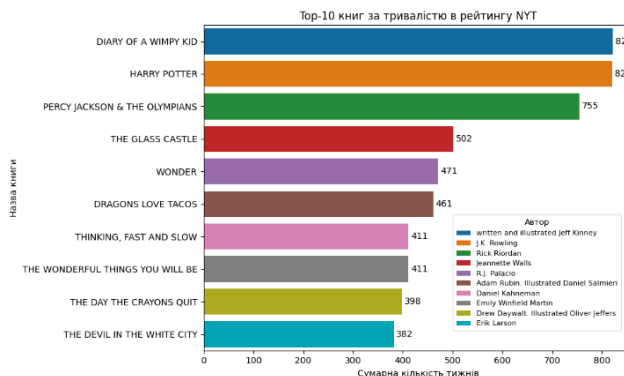


Рис. 29. NYT Bestseller: десять книг – лідерів за тривалістю в рейтингу

Інтегрований аналіз трьох джерел показав виразні відмінності у розподілах популярності та підтвердив очікувані ринкові закономірності. Для Goodreads основна маса нормалізованих значень сконцентрована у діапазоні 0.4–0.6, і лише 14 книжок перевищують 0.8 (рис. 30). Це добре узгоджується з попередніми спостереженнями щодо "довгого хвоста": навіть на платформі з активною читацькою аудиторією більшість видань утримуються на середньому рівні, а поодинокі релізи набирають виняткові оцінки. У каталозі Amazon максимум також припадає на діапазон 0.5–0.6, і лише вісім книжок мають значення вище 0.8 (рис. 31).

Така плоска форма розподілу збігається з дослідженнями великих комерційних маркетплейсів, де поєднується масова пропозиція та розпорошений попит. У рейтингах NYT концентрація ще вища: 64 книжки мають нормалізовану кількість тижнів, близьку до нуля, і лише шість перевищують 0.5 (рис. 32). Це очікувано й відображає жорсткий відбір і короткий життєвий цикл більшості бестселерів.

Дисперсійний аналіз для розрахунку вагових коефіцієнтів показав співвідношення 0.29 для Goodreads, 0.37 для Amazon і 0.34 для NYT. Такі пропорції логічні: комерційна компонента дещо переважає у підсумковій оцінці, що збігається з результатами досліджень про роль комерційних показників у загальному успіху видання. Підсумковий індекс Total Score має максимум близько 0.56 і виразну концентрацію у середньому сегменті (рис. 33). Це підтверджує, що досягнути одночасно високих значень за читацькими, ринковими та медійними метриками здатна лише невелика частина книжок, що цілком очікувано для змішаних ринкових даних і збігається з висновками міжнародних оглядів бестселерів.

Окремий аналіз показника Source Rank продемонстрував, що зі 126 книг, що представлені принаймні на двох платформах, – 78 представлених одразу на трьох, а решта – лише на двох. Для видань із Source Rank = 3 характерна цінова стабільність: міжквартильний розкид не перевищує 1.5, без екстремальних викидів (рис. 34). Це збігається з ринковими спостереженнями: багатоплатформні релізи зазвичай мають узгоджену політику ціноутворення. Частка таких релізів у кожному форматі становить близько 15–20 % (див. рис. 26), що підтверджує галузеві оцінки про обмежене коло "універсальних" бестселерів.



Рис. 30. Інтегрований аналіз: нормалізований рейтинг Goodreads

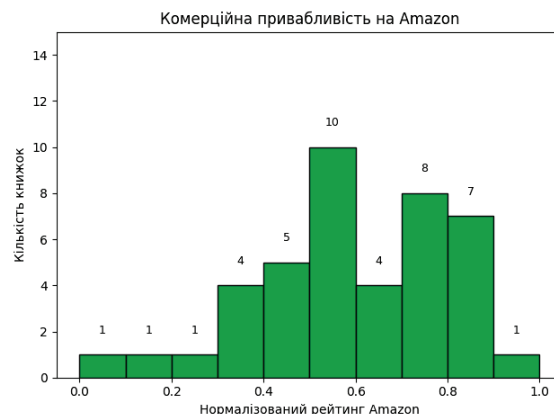


Рис. 31. Інтегрований аналіз: нормалізований рейтинг Amazon

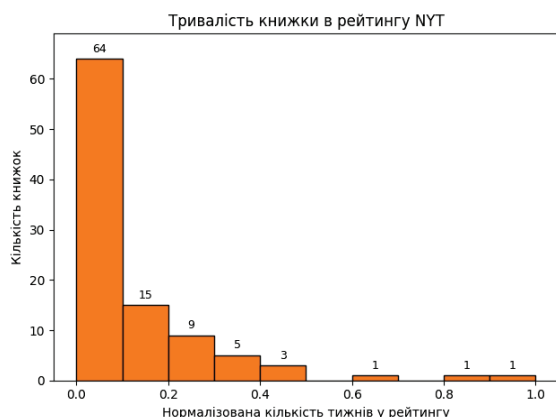


Рис. 32. Інтегрований аналіз: нормалізована кількість тижнів у рейтингу NYT Bestsellers

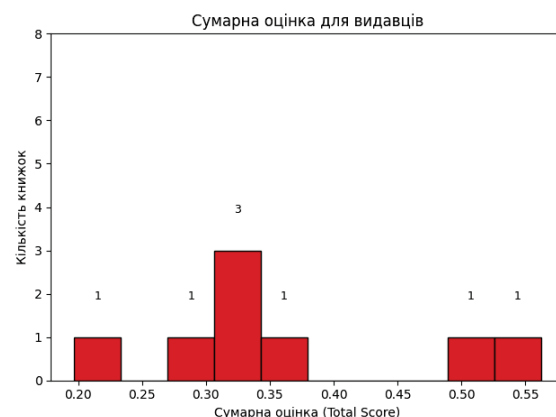


Рис. 33. Інтегрований аналіз: сумарна оцінка Total Score

Жанровий зріз виявив характерну структуру популярності. У групі з високим Source Rank найчастіше зустрічаються Romance і Fiction (по три позиції), а також Audiobook і Fantasy (по дві), що відображає стандартний набір для міжнародних бестселерів. Серед видань із середнім Source Rank домінують Romance і Fiction (по 38), далі йдуть Audiobook і Fantasy (по 28) та Nonfiction (27), що показано на рис. 34. Такі результати збігаються з іншими аналітичними роботами про стійкий глобальний попит на масові жанри. Профіль десяти книжок – лідерів (рис. 35), підтверджує цю картину: провідні позиції займають Fiction і Romance (по 7), далі Contemporary (5), а по 4 – Fantasy, Nonfiction і YA. Ця картина узгоджується з попередніми спостереженнями про гнучкість найуспішніших релізів, які поєднують масові жанри з елементами нішевих напрямів і завдяки цьому здатні довго утримувати увагу різних груп читачів.

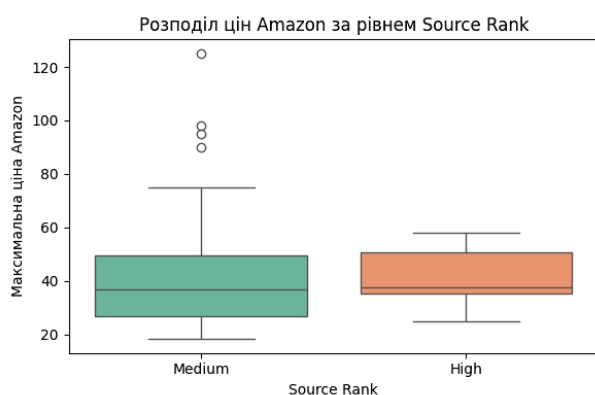


Рис. 34. Інтегрований аналіз: розподіл цін Amazon за рівнем Source Rank

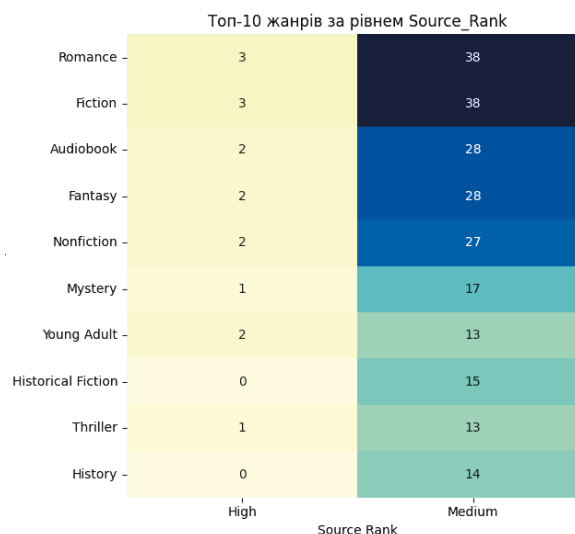


Рис. 35. Інтегрований аналіз: десяти книжок – лідерів жанрів за рівнем Source Rank

Фінальний список книжок із найвищим інтегральним потенціалом перекладу та просування охоплює *Fourth Wing*, *Yellowface*, *The Housemaid*, *Happy Place*, *Tomorrow*, and *Tomorrow*, *Remarkably Bright Creatures*, *The Wager* і *Demon Copperhead*. Усі вони демонструють переконливі результати на трьох платформах і поєднують сильні читачські оцінки, ринкову привабливість і медійну стійкість.

Дискусія і висновки

Отримані результати показують, що довготривала популярність книжки формується на перетині трьох взаємопов'язаних вимірів – читачького інтересу, ринкової привабливості та стабільної присутності в медіа. Поєднання даних Goodreads, Amazon і NYT Bestsellers доводить: успіх видання визначають не миттєві рейтингові сплески, а здатність утримувати увагу завдяки серійності, багатоформатній дистрибуції та продуманій ціновій стратегії. Ця динаміка добре корелює з моделями "довгого хвоста" та кумулятивного попиту, які пояснюють, як тривале споживання створює стійку додану вартість.

Порівняння з попередніми дослідженнями підтверджує і водночас розширює наявні уявлення. Висновки Kousha, Thelwall and Abdoli (2017) про зв'язок між кількістю рецензій і впливом книги доповнюються жанровою сегментацією та індексами нішевості й перспективності, показуючи, що високі рейтинги однаково можливі і в масових, і у вузьких сегментах. Дослідження Maity, Panigrahi and Mukherjee (2018) зосереджувалося на поведінкових сигналах із



Goodreads для прогнозу Amazon-бестселерів, тоді як запропонований підхід включає цінові та медійні параметри, доводячи, що мультиформатність і цінова стабільність є не менш впливовими факторами.

Спостереження Maloku and Maloku (2024) про значущість середнього рейтингу й кількості відгуків як ключових драйверів Amazon-бестселерів у представленій моделі отримали уточнення: фіксується відсутність кореляції між ціною та якістю, але водночас виявляється помірний позитивний зв'язок між кількістю форматів і комплексним індексом успіху. Аналіз сезонності (Pruett, 2020) стосувався лише друкованих видань, в той час як наші результати не лише відтворюють пікові періоди першого і третього кварталів, а й показують, як поява електронних та аудіоформатів змінила тривалість і характер присутності книжок у рейтингах після 2020 р. Нарешті, слабкі кореляції між відгуками Amazon і науковими цитуваннями, які виявили Kousha and Thelwall (2015), узгоджуються з визначенням у цій роботі показником слабкої залежності між кількістю оцінок і середнім рейтингом ($r \approx 0.18$), що ще раз підтверджує обмеженість популярності як показника якості.

Практичне значення представлених висновків виходить далеко за межі академічних дискусій. Видавці можуть використовувати інтегрований індекс і жанрові метрики для формування перекладацьких портфелів, вибору оптимального часу та формату релізів і визначення ніш, що здатні на довгу комерційну "життєздатність". Маркетингові команди та літературні агенти отримують інструменти для точнішого прогнозування попиту й сегментації аудиторій, а книжкові мережі та онлайн-платформи – аналітичну базу для керування асортиментом і персоналізації рекомендацій. Для дослідників культурних індустрій ці дані відкривають шлях до глибшого вивчення впливу серійності, мультиформатності й медійності на сталу популярність літератури.

Отже, мультиджерельний підхід не лише уточнює картину глобальних літературних трендів, а й надає практичну методологію для стратегічного планування – від добору авторів і жанрових напрямів до довгострокового управління видавничими інвестиціями та міжнародного просування.

Крім практичних висновків, дослідження відкриває і перспективи для подальших наукових пошуків. По-перше, це динамічний часовий вимір. Розширення вибірки на багаторічний період дасть змогу простежити еволюцію жанрових і цінових трендів у довгій перспективі й перевірити стійкість "довгого хвоста" в різні економічні та соціальні епохи, включно зі впливом глобальних криз чи різних технологічних зрушень.

По-друге, варто залучити текстову аналітику змісту. Інтеграція методів NLP для аналізу рецензій і самих текстів допоможе виявити стилістичні, тематичні або наративні фактори, що корелюють із довготривалою видимістю та високими рейтингами, розширюючи підхід від кількісних до якісних показників.

По-третє, перспективним є вивчення поведінкових сигналів читачів. Залучення даних із соціальних медіа, платформ на кшталт TikTok (BookTok), Wattpad чи Patreon дасть змогу дослідити, як фанатські спільноти та вірусні тренди впливають на комерційний успіх, серійність і життєвий цикл книжки.

Ще один напрям – моделювання цінових стратегій. Аналіз динамічного ціноутворення й акцій для друкованих, електронних та аудіоформатів дасть змогу точніше оцінити роль гнучкої цінової політики в підтримці стабільного попиту та довготривалої видимості.

Нарешті, перспективним кроком є розроблення прогнозних систем для видавців. Побудова моделей машинного навчання на основі інтегрованих метрик (жанр, формати, рецензії, медійність) може дати змогу прогнозувати ймовірність довготривалого потрапляння у списки бестселерів або визначати потенціал для міжнародного перекладу ще на етапі підготовки видання.

Зазначені напрями поглиблюють розуміння механізмів культурної та комерційної стійкості книжок і створюють підґрунтя для ще точніших інструментів стратегічного планування у видавничій сфері.

Внесок авторів: Володимир Дружинін – написання (перегляд і редагування); Євгеній Батрак – курування та валідація даних, написання (перегляд і редагування); Наталія Цьопа – концептуалізація, методологія, аналіз джерел, підготування огляду літератури; Ольга Ніколенко – формальний аналіз, збір емпіричних даних та їхня валідація, методологія, написання (оригінальна чернетка).

Джерела фінансування. Це дослідження не отримало жодного гранту від фінансової установи в державному, комерційному або некомерційному секторах.

Список використаних джерел

- Kousha, K., & Thelwall, M. (2016). Can Amazon.com reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*, 67(3), 566–581. <https://doi.org/10.1002/asi.23404>
- Kousha, K., Thelwall, M., & Abdoli, M. (2017). Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology*, 68(8), 2004–2016. <https://doi.org/10.1002/asi.23805>
- Maity, S. K., Panigrahi, A., & Mukherjee, A. (2018). *Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers* (No. arXiv:1809.07354). arXiv. <https://doi.org/10.48550/arXiv.1809.07354>
- Pruett, J. (2022). *New York Times Hardcover Fiction Bestsellers (1931–2020)*. Post45 Data Collective. <https://doi.org/10.18737/CNJV1733p4520220211>

References

- Kousha, K., & Thelwall, M. (2016). Can Amazon.com reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*, 67(3), 566–581. <https://doi.org/10.1002/asi.23404>
- Kousha, K., Thelwall, M., & Abdoli, M. (2017). Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology*, 68(8), 2004–2016. <https://doi.org/10.1002/asi.23805>
- Maity, S. K., Panigrahi, A., & Mukherjee, A. (2018). *Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers* (No. arXiv:1809.07354). arXiv. <https://doi.org/10.48550/arXiv.1809.07354>
- Pruett, J. (2022). *New York Times Hardcover Fiction Bestsellers (1931–2020)*. Post45 Data Collective. <https://doi.org/10.18737/CNJV1733p4520220211>

Отримано редакцією журналу / Received: 17.10.25

Прорецензовано / Revised: 20.10.25

Схвалено до друку / Accepted: 19.12.25



Volodymyr DRUZHYNIN¹, DSc (Engin.), Prof.
ORCID ID: 0009-0009-5049-0099
e-mail: volodymyr.druzhynin@knu.ua

Yevhenii BATRAK², PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0003-3027-4507
e-mail: batrak.yevhenii@ill.kpi.ua

Nataliia TSOPA², PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0002-3215-0711
e-mail: e-mail: tsopa.nataliia@ill.kpi.ua

Olha NIKOLENKO², Master's Student
ORCID ID: 0009-0003-2304-7821
e-mail: nikolenko.olha@ill.kpi.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

MULTISOURCE ALGORITHMIC ANALYSIS OF READER AND COMMERCIAL TRENDS IN THE BOOK MARKET

Background. The book market is undergoing rapid transformation under the influence of digital platforms, and timely identification of literary trends has become critically important for publishers working with translated literature. The aim of this study is to identify key reader and commercial trends of 2024 based on data from Goodreads, Amazon, and The New York Times Best Sellers and to assess their potential for shaping effective publishing strategies.

Methods. The dataset included 280 books from the Goodreads Choice Awards 2024, 225 titles from the Amazon Best Books 2024 list, and 949 books from the NYT Best Sellers ranking. Data were collected using Python scripts (Selenium, BeautifulSoup), then cleaned and normalized with pandas and NumPy. Statistical analysis, Principal Component Analysis (PCA), and K-Means clustering were applied to compare reader, commercial, and media indicators.

Results. The integrated analysis of the three sources revealed distinct patterns and shared tendencies. On the reader-preference level (Goodreads), Romance and Fiction attract the broadest audience, while History and Nonfiction consistently achieve high ratings. "Hidden gems" such as the World War II, Mental Health, and Health categories combine strong ratings with limited reach, offering opportunities for targeted promotion. In the commercial segment (Amazon), printed formats – Hardcover and Paperback – dominate, whereas the Kindle Edition remains the most affordable. Premium niches such as Epic Fantasy and regional literary fiction maintain the highest average prices, reflecting readers' willingness to invest in exceptional content. Correlation analysis found no direct link between higher price and better rating but showed a moderate relationship between the number of available formats and overall price. Media visibility (NYT Best Sellers) demonstrates that sustained public interest is driven by children's series and nonfiction, with some titles remaining on the list for more than 300 weeks. Debut seasonality shows two peaks – during the first and third quarters – raising the likelihood of long-term chart presence. Overall, only eight titles achieved strong performance simultaneously across reader, market, and media criteria, demonstrating balanced success and the greatest potential for translation and international promotion.

Conclusions. The integrated analysis shows that a book's commercial success is not determined by a single factor but requires a combination of sustained reader engagement, balanced pricing strategy, and the ability to remain visible in the public sphere. This comprehensive perspective enables publishers to plan portfolios not only by genre priorities but also by long-term economic feasibility. It is equally important to account for temporal demand fluctuations, as market cycles influence the effectiveness of marketing and translation campaigns. Aligning reader, commercial, and media criteria provides a foundation for more accurate forecasting and reduces the risk of one-time successes. Ultimately, the findings create a methodological basis for strategic decisions on translation, investment, and international project development.

Keywords: integrated data analysis, cross-platform comparison, open data, algorithmic analysis, clustering, reader-behavior modeling.

Дружинін Володимир є членом редакційної колегії цього видання, тому він не брав участі в процесі рецензування та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

Druzhynin Volodymyr is a member of the Editorial Board of this publication; therefore, they did not participate in the peer review process or the decision-making regarding the publication of this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.