МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА

БОЙКО УСТИМ ВІКТОРОВИЧ

Допускається до захисту:
завідувач кафедри
інформаційних технологій,
канд. техн. наук, доцент
_____ О. В. Зелінська
«_____» _____ 20__ р.

ВИЯВЛЕННЯ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА РЕЙТИНГ ТА КАСОВІ ЗБОРИ ФІЛЬМУ МЕТОДАМИ DATA SCIENCE

Спеціальність 122 «Комп’ютерні науки»

Кваліфікаційна (магістерська) робота

Науковий керівник:
Ю. С. Антонов, доцент кафедри
інформаційних технологій
к. ф.-м. наук, доцент
_____
(підпис)

Оцінка _____ / _____ / _____
(бали за шкалою ЄКТС/за національною шкалою)

Голова ЕК: _____
(підпис)

Вінниця 2024

АНОТАЦІЯ

Бойко У. В. Виявлення факторів, що впливають на рейтинг та касові збори фільму методами Data Science. Спеціальність 122 “Комп’ютерні науки”, Освітня програма “Комп’ютерні технології обробки даних”. Донецький національний університет імені Василя Стуса, Вінниця, 2023.

У кваліфікаційній роботі розглянуто роль кіноіндустрії в сучасному суспільстві. Досліджено історичний аспект та етапи формування рейтингу та касових зборів фільмів. Досліджено набір даних “tmdb_movies.csv”. Показано етапи аналізу даних, використані інструменти та методи аналізу. Виявлено, які фактори мають вплив на показники касових зборів та рейтингу фільмів в наборі даних.

Ключові слова: аналіз даних, R, датасет, фільм, кіноіндустрія, рейтинг, касові збори.

70 с., 26 рис., 50 джерел.

Boiko U. V. Factors affecting identification of the rating and box office of the film using Data Science methods. Specialty 122 "Computer science", Programme "Computer data processing technologies". Vasyl Stus Donetsk National University, Vinnytsia, 2023.

The qualification paper examines the role of the film industry in modern society. The historical aspect and stages of rating and box office formation of films are studied. Data set “tmdb_movies.csv” was studied. The stages of data analysis, used tools and methods of analysis are shown. It is revealed which factors have an influence on the box office performance and rating of the films in the data set.

Keywords: data analysis, R, dataset, film, film industry, rating, box office.

70 p., 26 fig., 50 sources.

ЗМІСТ

ВСТУП.....4
РОЗДІЛ 1.....6
ТЕОРЕТИЧНИЙ ОГЛЯД.....6
1.1 Роль кіноіндустрії у сучасному суспільстві.....6
1.2 Історичний аспект формування рейтингу фільмів.....9
1.3 Історичний аспект формування касових зборів фільмів.....11
1.4 Аналіз основних досліджень та публікацій.....13
Висновок до першого розділу.....20
РОЗДІЛ 2.....21
МЕТОДИКА ДОСЛІДЖЕННЯ ТА ОБРОБКА ДАНИХ.....21
2.1 Постановка задачі.....21
2.2 Опис обраних інструментів та технологій для аналізу та обробки даних.....21
2.2.1 Мова програмування R.....22
2.2.2 RStudio.....23
2.2.3 Tidyverse.....25
2.3 Вибір та опис набору даних.....27
2.4 Підготовка та очищення даних.....30
2.5 Вибір методів аналізу даних та їх обґрунтування.....38
Висновок до другого розділу.....41
РОЗДІЛ 3.....42
АНАЛІЗ ДАНИХ.....42
3.1 Описова статистика.....42
3.2 Кореляційний аналіз та візуалізація.....45
3.3 Регресійний аналіз.....53
3.4 Випадковий ліс.....58
Висновок до третього розділу.....63
ВИСНОВОК.....65
СПИСОК ЛІТЕРАТУРИ.....66

ВСТУП

Фільмова індустрія завжди привертала увагу суспільства через свій вплив на культуру, мистецтво та комерційний сектор. Фільми стали не тільки невід'ємною частиною нашого розважального життя, але і важливим джерелом прибутку для кінокомпаній та студій.

Сучасна кіноіндустрія стикається зі зростаючою конкуренцією та змінами в споживацьких звичках глядачів. Щоб привернути увагу глядачів і забезпечити прибуток, кінокомпанії повинні аналізувати та розуміти, які фактори впливають на успішність фільмів. Аналіз даних, що відносяться до кіно в сучасному світі стає все важливішим завданням для кінокомпаній, режисерів, продюсерів, маркетологів та дослідників.

Одним із ключових аспектів успішності фільму стали його рейтинг та касові збори. Розуміння того, які фактори на них впливають, може визначити успіх або невдачу кінострічки. Рейтинги, які встановлюються кінокритиками та глядачами, впливають на популярність фільму та його прийняття в суспільстві. Касові збори визначають комерційний успіх та прибуток, який отримує виробник або кінокомпанія. У світлі цього, виявлення та розуміння показників, що впливають на ці два основних аспекти, стає критичним завданням для кінематографічної індустрії.

Дана робота спрямована на вивчення важливих питань, пов'язаних із впливом різних факторів на рейтинги та касові збори фільмів. Використовуючи різноманітні методи аналізу та обробки даних, ми спробуємо відповісти на питання, що цікавлять кінокомпанії, кінокритиків та глядачів, і виявити ті ключові чинники, які визначають успіх фільмів.

Метою даної магістерської роботи є виявлення та аналіз факторів, які впливають на рейтинг та касові збори фільмів. Розглядаючи ці показники, ми прагнемо розкрити внутрішні закономірності та фактори, які детермінують комерційний успіх та сприяють популярності фільмів серед глядачів.

Об'єкт дослідження - це фільми, що входять до набору даних “tmdb_movies.csv” [1], який включає в себе інформацію про різні аспекти фільмів, такі як жанр, бюджет, режисер, акторський склад та інші.

Предмет дослідження цієї роботи - це аналіз факторів, які впливають на рейтинг та касові збори фільмів.

Завдання дослідження включають в себе:

- **Зібрання та підготовка даних:** Завантаження та очищення набору даних для подальшого аналізу.
- **Експлоративний аналіз даних (EDA):** Вивчення основних статистичних характеристик даних та виявлення взаємозв’язків між різними змінними.
- **Моделювання рейтингу та касових зборів фільмів:** Розробка моделей машинного навчання, які дозволять виявити вплив доступних факторів на рейтинг та касові збори фільмів.
- **Аналіз та інтерпретація результатів:** Визначення важливих факторів, які впливають на успішність фільмів та формулювання висновків з отриманих результатів.

Слід відзначити, що аналіз кіноіндустрії за допомогою різноманітних методів обробки та аналізу даних є актуальним завданням у світі, де кіно залишається важливим елементом культури та розваг. Ця магістерська робота спрямована на розкриття секретів успіху фільмів та надає можливість ділитися цими знаннями зі спільністю кінодослідників та працівників кіноіндустрії.

РОЗДІЛ 1
ТЕОРЕТИЧНИЙ ОГЛЯД

1.1 Роль кіноіндустрії у сучасному суспільстві

Кіно як форма мистецтва володіє особливим статусом у сучасному світі, відіграючи важливу роль у життєдіяльності кожної особи. Це не просто вид мистецтва, але й могутній соціокультурний феномен, який активно впливає на формування свідомості індивіда [2].

Кожен кінофільм занурює нас у конкретний культурний контекст. Ці витвори мистецтва є відображенням нашої сутності, демонструючи наші переконання та особливості спільного існування. Через фільми нам відкриваються наші страхи, погляди, слабкості та достоїнства, які, можливо, не завжди помітні у повсякденному житті.

Фільми не лише зображують дійсність, але й стають каталізаторами змін, заохочуючи нас переосмислювати і, можливо, коригувати власні світогляди. Аудіовізуальні адаптації роблять фільми доступними для глобальної аудиторії, що сприяє об'єднанню людей різних культур і сприйняттям світу у всій його багатоманітності. Вони є чудовою платформою для спілкування з досвідом інших людей, який дуже відрізняється від нашого.

Більше того, кіно активно формує наше бачення світу, впливаючи на формування ціннісних орієнтацій. Варто відмітити вплив кінематографу на моду та музичні тенденції — стиль одягу знаменитостей кіно часто стає взірцем для наслідування. Сучасне суспільство активно використовує мову образів, створену на великому екрані, переймаючи не тільки слова, але й манери, поведінку персонажів.

Також варто відзначити, що кіно може як закріплювати існуючі культурні стереотипи, так і сприяти їх демонтажу, пропонуючи аудиторії нові, іноді неортодоксальні погляди на ті чи інші явища та процеси. Так,

кінематограф стає потужним інструментом соціокультурної адаптації та впливу, допомагаючи суспільству розвиватися та адаптуватися до постійно змінюваних умов існування [3].

Кінематограф не лише розважальна галузь, а й могутній двигун економіки, що створює широкі можливості для працевлаштування. Різноманітні професії — від акторів до операторів та реквізиторів — забезпечують людей робочими місцями, а також стимулюють діяльність багатьох прилеглих секторів, включаючи громадське харчування, освітлення та виготовлення костюмів.

Крім того, кіноіндустрія сприяє формуванню додаткових джерел доходу через випуск супутньої продукції, зокрема фігурок персонажів фільмів. Ці предмети мають високу колекційну цінність та здатні привертати покупців величезними цінами, що, в свою чергу, генерує значущі прибутки для виробників та продавців.

До того ж, кінотеатри отримують значну частку своїх доходів від продажу закусок та напоїв під час сеансів. Статистика показує, що до 85% прибутку кінотеатру може надходити саме від таких продажів [4]. При цьому значна частина цих грошей повертається у державний бюджет через податки.

Важливо також враховувати вплив продажу квитків на фінансовий обіг в країні. Хоча основна частина коштів від продажу квитків надходить студіям, значна сума грошей також витрачається на покриття податків, що сприяє економічному зростанню.

У масштабі країни, особливо в США, де 0.1% населення зайнято в кіноіндустрії, ця сфера стає важливим сектором для створення робочих місць та економічного розвитку [5]. Цей відсоток, хоча й здається невеликим, виявляється досить значущим при порівнянні з іншими галузями.

Враховуючи вище сказане, кіноіндустрія відіграє ключову роль не

лише у культурному, а й економічному житті суспільства, сприяючи створенню нових робочих місць, залученню інвестицій та загалом підтримуючи стабільність та розвиток економіки.

Кіноіндустрія відкриває нові горизонти в освітньому процесі, стаючи не лише засобом навчання, але й інструментом для розширення освітніх можливостей. Навчальні фільми, які включають в себе відеолекції, дискусії та демонстрації експериментів, створюють революцію в сучасній освіті, дозволяючи студентам вчитися незалежно та ефективно, навіть без безпосереднього втручання вчителя або фасилітатора [6].

Цей підхід трансформує традиційну освітню парадигму, перетворюючи кіно на доповнення або навіть альтернативу книжкам як основному джерелу навчального матеріалу. Це підкреслює важливість використання візуальних та аудіальних матеріалів у навчанні, сприяючи більш глибокому та багатогранному осмисленню навчального процесу.

Більш того, фільми можуть слугувати засобом підкреслення значущості формальної освіти, включаючи акцент на важливість не тільки академічних предметів, але й шкільних активностей, таких як мистецтво та спорт. Через екран можна показати позитивний вплив цих сфер на розвиток особистості.

Крім того, фільми, що фокусуються на шкільному житті та освітньому процесі, можуть стати джерелом натхнення для педагогів та освітніх закладів у вдосконаленні навчальних методик та розробці інноваційних підходів до освіти [7].

Освітні фільми також можуть відігравати важливу роль у залученні громадськості до дискусій щодо освітньої політики та стратегій, допомагаючи формувати більш інформоване та освічене суспільство.

Фільми є ключовим елементом сучасного суспільства, вони відіграють багатогранну роль, впливаючи на культурні, освітні, економічні та інші аспекти людського життя.

1.2 Історичний аспект формування рейтингу фільмів

Досліджуючи витоки формування систем рейтингу фільмів, важливо звернути увагу на ті історичні контексти та чинники, які впливали на цей процес.

На початковому етапі розвитку кінематографії, основним джерелом формування рейтингу фільмів були відгуки критиків та реакції аудиторії. Звісно, перші фільми характеризувалися обмеженими бюджетами, і їхній комерційний успіх в більшій мірі був залежний від “сарафанного радіо”, цієї невидимої, але впливової мережі усного обговорення.

Але з розвитком кіноіндустрії та зростанням конкуренції між студіями, стали з'являтися більш структуровані та об'єктивні методи оцінки фільмів. Наприклад, було введено системи рейтингу, які допомагали оцінити фільми з урахуванням їх вікових обмежень.

У 1960-і роки в США була створена система MPAA(Motion Picture Association of America), яка давала фільмам певні рейтинги залежно від їх вмісту. Система рейтингів допомагала аудиторії визначити, які фільми підходять для певної вікової категорії, і тим самим формувала відношення глядачів до кінопродукції [8].

З часом в індустрії кіно з'явилися професійні асоціації критиків та кіножурналістів, які не тільки аналізували нові релізи, а й формували думку широкої аудиторії. Вони аналізували різні аспекти фільмів - від сценарію до акторської гри - і давали оцінки, які могли впливати на вибір глядачів. Крім того, популярність кіно як мистецтва зросла настільки, що стали проводитися численні кінофестивалі, на яких фільми конкурували за престижні нагороди та визнання експертів.

З приходом Інтернету та цифрових технологій, система рейтингу фільмів пережила ще одну революцію. З'явилися веб-сайти та мобільні додатки, які збирають відгуки та оцінки від звичайних глядачів. Це дало можливість широкій аудиторії активно впливати на рейтинг фільмів,

обговорювати їх та ділитися думками. Сервіси типу TMDb (The Movie Database), Rotten Tomatoes, Metacritic та багато інших стали важливими платформами для формування рейтингу фільмів, базуючись на численних рецензіях та оцінках користувачів і професійних критиків.

Сьогодні однією з ключових платформ у контексті формування рейтингів фільмів є IMDb (Internet Movie Database). Цей сервіс є джерелом важливої інформації, що включає не тільки деталі про кінопродукцію, але й рейтинги фільмів, які базуються на відгуках та оцінках звичайних користувачів та критиків [9].

IMDb дозволяє користувачам створювати особисті акаунти, де вони можуть залишати відгуки та оцінки, а також ділитися своїми думками з іншими користувачами. Система рейтингу на IMDb враховує різні фактори, включаючи кількість голосів, що були віддані за конкретний фільм, та середню оцінку кожного фільму, формуючи тим самим агрегований рейтинг.

Рейтинг IMDb, який вимірюється за десятибальною шкалою, став еталоном для визначення якості фільмів. Він впливає не тільки на вибір користувачів, але й часто використовується професійними критиками та аналітиками в якості додаткового інструменту для аналізу кінопродукції [10].

Таким чином, система рейтингу фільмів пройшла довгий шлях розвитку, від простих устних рекомендацій до комплексних механізмів, які включають в себе як думку професійних критиків, так і широкої аудиторії. Сучасна система рейтингу є результатом синтезу традиційних методів оцінки та новітніх цифрових технологій, дозволяючи створити більш об'єктивний та багатогранний погляд на кінопродукцію.

У світлі згаданого вище, можна зазначити, що система рейтингу фільмів є динамічною структурою, яка відображає не тільки якісні характеристики кінострічок, а й соціокультурні та економічні фактори

конкретного історичного періоду.

Однією з найважливіших змін у формуванні системи рейтингу стала можливість миттєвого обміну думками та відгуками через онлайн-платформи. Сьогодні кожен глядач має можливість не просто дізнатися думку експертів, а й висловити свою власну, створюючи тим самим колективну оцінку фільму.

Варто також зазначити роль соціальних мереж у формуванні рейтингу. Сьогодні популярні блогери та інфлюенсери можуть мати значний вплив на успіх або невдачу фільму, рекомендуючи стрічки своїм підписникам або ж критикуючи їх.

1.3 Історичний аспект формування касових зборів фільмів

В історії кіноіндустрії процес формування системи касових зборів фільмів пройшов декілька значущих етапів, відображаючи глибокі зміни у самій кіноіндустрії та суспільстві взагалі.

У перші десятиліття існування кінематографу (до середини 20-го століття) основним показником успіху фільму була кількість проданих квитків, яка дозволяла зробити висновки про популярність кінопродукції серед аудиторії.

Поява кінопрокатних компаній у 40-60 роках 20 століття знаменувала нову епоху в індустрії кіно, де акцент був перенесений на комерційну атрактивність фільмів. Паралельно з цим, почали активно формуватися системи обліку касових зборів, що використовували різноманітні методики та критерії для аналізу економічного успіху фільмових проектів.

Було впроваджено різноманітні інструменти для точного вимірювання касових зборів: від створення спеціалізованих агентств, що займаються статистичним аналізом і прогнозуванням, до використання складних алгоритмів для вивчення попиту та уподобань аудиторії.

Касові збори стали важливим показником, що відображає

популярність та комерційну успішність фільму, базуючись на його фінансових показниках. Цей показник допомагає індустрії кіно визначати, які жанри, режисери, актори та інші елементи кіно є найбільш привабливими для глядачів.

Згодом, у 70-90 роках 20 століття, кінокомпанії почали звертати особливу увагу на маркетингові стратегії, спрямовані на максимізацію касових зборів. Це означає не лише створення якісного контенту, а й розробка ефективних рекламних кампаній, а також стратегічне планування дати релізу, щоб уникнути конкуренції з іншими головними прем'єрами. Касові збори тепер враховували глобальні доходи, включаючи міжнародні продажі квитків [11].

Більше того, в сучасну епоху, поява нових технологій та цифрових платформ також значуще вплинула на систему касових зборів. Сьогодні, окрім традиційних кінотеатрів, значну частину прибутку приносять стрімінгові сервіси, які нерідко закуповують права на ексклюзивну демонстрацію фільмів. Касові збори тепер включають і доходи від платних підписок, продажу та оренди фільмів онлайн.

Таким чином, система касових зборів стала набагато більш складною і багатоаспектною, враховуючи множину каналів дистрибуції та взаємодію з глядачем у сучасному медіапросторі. У цьому контексті, аналіз фінансового успіху фільму вимагає глибокого дослідження ринку та аудиторії, а також уміння адаптувати стратегії під динамічно змінювані умови ринку.

Історичний аспект формування касових зборів демонструє постійний розвиток та адаптацію системи оцінювання комерційного успіху фільмів до змін у кіноіндустрії та суспільстві в цілому.

1.4 Аналіз основних досліджень та публікацій

Дослідження факторів, що впливають на рейтинг та касові збори фільмів, має довгу історію, яка нараховує понад 80 років. Вчені вже з

1940-х років займались вивченням цього питання, спершу фокусуючись на дослідницьких методиках. Перші дослідники, такі як Галлап та Гендель, систематично вивчали такі впливові фактори, як акторський склад, маркетинг, сюжет і оцінки, щоб прогнозувати касові збори [12].

Дослідники зазначали, що успіх фільму у кінопрокаті базується на трьох основних аспектах: характеристиках самого фільму, силі маркетингової стратегії та відгуках як критиків, так і аудиторії.

Дослідники також розглядали інші потенційні фактори, включаючи походження фільму, вартість його створення, графік показів, вплив режисера та нагород, вплив знаменитостей, жанр, відгуки, ступінь культурної близькості та фактори, що стосуються споживачів.

Першу модель багатофакторної регресії для прогнозування фінансового успіху фільму розробив Літман [13], у 1983 році, надихнувшись словами президента MPAA, Джека Валенті, про непередбачуваність успіху фільму на ринку.

Літман вказав на три ключові аспекти, які, на його думку, визначають успіх фільму:

- **Творча сфера**: історія, акторський склад, режисер, бюджет виробництва та рейтинг. Він зазначив, що важливість реалістичності та правдивості сюжету зменшилася, оскільки більшість з найбільш прибуткових фільмів належать до жанрів наукової фантастики, анімації чи супергеройських фільмів.
- **Розклад та стратегія випуску:** важливість обрання великого дистриб'ютора, щоб мати перевагу у переговорах, фінансових ресурсах, доступі до кінотеатрів тощо. Також було вказано на сезонність попиту на фільми, яка відображалася в періодах навколо Різдва, Нового року, літа та Великодня.
- **Маркетингові зусилля:** наголошено на важливості медійної кампанії та впливу відгуків після випуску фільму. Літман визнає

значущість ролі критиків та їх вплив на мотивацію аудиторії дивитися фільми.

Для тестування своїх теорій, Літман створив модель багатофакторної регресії із залежною змінною, як рентабельні платежі, які отримував дистриб'ютор. Він включив низку незалежних змінних, таких як жанр, рейтинг MPAA, наявність суперзірки у складі, бюджет виробництва, тип дистриб'ютора, період випуску, номінації та перемоги на премії "Оскар", а також відгуки критиків. Він використовував зірковий рейтинг, присвоєний щоденною газетою, як міру врахування відгуку критиків. Вищий рейтинг свідчить про вищу оцінку.

За результатами аналізу було виявлено, що велике значення мають виробничі витрати, які позитивно корелювали з успіхом в прокаті, відгуки критиків, деякі жанри (наукова фантастика та жахи), а також фільми, що розповсюджувалися великими дистриб’юторами. Період випуску на Різдво, номінації на премію “Оскар” та власне виграні нагороди теж впливали на касові збори.

Це дослідження Літмана стало фундаментом для багатьох наступних досліджень у наступні два десятиліття.

Дослідження проведене Артуром де Вані та Девідом Уолсом у 1999 році [14], базувалося на аналізі понад 2000 фільмів і висвітлювало різні аспекти впливу на доходи від прокату фільмів.

Вони визначили, що доходи від кінопрокату фільмів мають нескінченний діапазон варіацій і залежать від декількох блокбастерів, які приносять основну частину прибутку. Автори зауважили, що цей процес може бути описаний за допомогою розподілу Леві, де декілька фільмів приносять значний прибуток, в той час як більшість мають низькі доходи.

Щодо стратегій виробництва, вони зазначили, що, незважаючи на можливість впливу на успіх фільму за допомогою стратегічного вибору акторського складу, бюджету та кількість екранів для демонстрації,

фактичний успіх визначається аудиторією після прем'єри фільму. Тут важливою є динаміка взаємодії та обміну думками великої кількості людей, яка є складною та непередбачуваною, і яку не можна керувати навіть дорогими маркетинговими кампаніями.

Дослідники розглянули ряд факторів, які можуть впливати на прибутковість фільмів, включаючи наявність сиквелів, жанр, рейтинг, акторський склад, бюджет та кількість екранів на старті. Їх аналіз показав, що середній дохід від фільму в їхній вибірці становив 17 мільйонів доларів, що було значно вище за медіану в 6,9 мільйона доларів.

Автори дослідження також детально розглядали вплив зірок на успіх фільму, зокрема їхню здатність збільшувати початкові доходи та покращувати якість продукту. Однак вони підкреслили, що навіть наявність зірок у фільмі не гарантує його успіху, а лише збільшує шанси на позитивний результат, хоча й досить незначно.

Де Вані та Уоллс запропонували стратегію формування “портфеля фільмів”, яка передбачає одночасний вибір та інвестиції в декілька кінопроектів, замість того, щоб концентрувати всі ресурси на одному окремому фільмі.

Ця стратегія враховує високий рівень невизначеності та конкуренції у кіноіндустрії, де дуже складно передбачити, який фільм стане успішним. Інвестуючи в декілька проектів одночасно, студія може розподілити ризики, тобто навіть якщо деякі фільми не принесуть очікуваних доходів, інші проекти можуть виявитися успішними та компенсувати збитки.

Цей підхід допомагає зменшити вплив невдачі одного проекту на фінансовий стан студії, дозволяючи їй підтримувати стабільність та продовжувати роботу незважаючи на невизначеність ринку.

У 1999 році Абрахам Равід дослідив роль зірок та інших інформаційних сигналів у кіноіндустрії, пропонуючи дві альтернативні гіпотези щодо впливу зірок на успіх фільмів [15].

Перша гіпотеза підкреслює, що зірки швидко коригують свої гонорари, відображаючи свою ринкову вартість, та в основному зводжуючи до нуля вартість, яку вони додають фільму. Іншими словами, ті гроші які фільм має з касових зборів через залучення зірок кінематографу ідуть їм на зарплату. Як приклад, автор наводить Джона Траволта та Алісію Сільверстоун, чиї гонорари стрімко зросли після успіху їхніх фільмів.

Друга гіпотеза базується на думці, що виробничі керівники, які добре знають проект, можуть наймати відомих і дорогих зірок як сигнал іншим стейкхолдерам (наприклад, студіям або фінансистам) про високу якість проекту. Таке рішення є сміливим, оскільки кар'єра студійного керівника може бути безпосередньо пов'язана з успіхом або невдачею фільму.

Для тестування цих гіпотез Равід використовував вибірку з 200 фільмів, випущених між 1991 та початком 1993 року. Дослідження показало, що фільми із зірками у середньому приносили вищий дохід, причому найважливішою незалежною змінною був бюджет фільму. Також було відмічено, що існує позитивна кореляція між кількістю оглядів фільму та його доходами. Це може свідчити про те, що фільми, які обговорюються частіше, залучають більше глядачів, що в свою чергу призводить до збільшення доходів.

Однак, коли в модель було включено більше незалежних змінних, зв'язок між наявністю зірок у фільмі та його доходами зникав, а кофіцієнти ставали негативними.

Равід також звернув увагу на те, що великі бюджети не завжди корелювали з високою прибутковістю, іноді вони навіть спричиняли збитки. Це вказує на складність процесу визначення бюджету та необхідність обережного підходу до фінансування фільмів.

В дослідженні факторів, що впливають на прибутковість фільмів у кінотеатрах, було розглянуто різні аспекти, включаючи роль критиків,

рейтинги MPAA, жанр фільму, а також додаткові характеристики, такі як бюджет, кількість театрів, в яких йдуть покази, та чи є фільм продовженням іншого.

Щодо ролі критиків, то Джехошуа Еліашберг і Стівен Шуган в 1997 році з'ясували, що критики можуть діяти як впливові особи, що важливо для доходів від продажу квитків [16]. Однак Девід Рейнштейн і Крістофер Марк Снайдер в 2005 році вказали, що лише декілька критиків мали реальний вплив на споживчий попит [17].

Щодо жанру фільму, було проведено численні дослідження, де дослідники намагалися визначити, як жанр впливає на доходи від касових зборів. Результати виявились суперечливими: деякі з них показали позитивну кореляцію між певними жанрами і доходами, тоді як інші знайшли негативну залежність, особливо для драми. Проаналізувавши їх можна зазначити, що в різні періоди були популярні різні жанри і ситуація в цій сфері може змінюватись з часом через зміну смаків аудиторії.

Дослідження, де вперше використовувались глобальні касові збори як метрику успіху, на відміну від попередніх досліджень, які головним чином орієнтувались на ринок США, провели Пангаркер і Сміт у 2013 році. Мета їхнього дослідження полягала в тому, щоб виявити фактори, які сприяють успіху фільму в прокаті. Цього було досягнуто за допомогою багатофакторної регресійної моделі [18].

Основні висновки з цього дослідження включають:

- Вартість виробництва фільму є найважливішим чинником для його касових зборів. Фільми з більш високими виробничими бюджетами, як правило, мають вищі глобальні касові збори. Це відповідає результатам попередніх досліджень Літмана та Равіда. Проте були зауважені винятки, оскільки деякі фільми з низьким бюджетом також показали винятково високі результати.
- Вихід фільму великою студією позитивно впливає на доходи,

додаючи приблизно 11,3 мільйона доларів до зборів в порівнянні з фільмами, які не випускалися великими студіями.

- Номінації на нагороди також відіграють важливу роль у успіху фільму. Їхня модель оцінила, що фільми з номінаціями збільшують свої доходи приблизно на 39 мільйонів доларів.
- Модель виявила надзвичайно значний зв’язок між фільмами, які є сиквелами та касовими зборами. Результат для продовження означав, що сиквел може додати до касових зборів аж 169 мільйонів доларів.
- Вихід фільму під час свят не впливає значуще на його доходи, що суперечить днеяким попереднім дослідженням.
- Вплив критиків був ще одним несподіваним результатом. На відміну від попередніх досліджень, які виявили позитивні відносини між відгуками критиків та доходами, це дослідження не знайшло такого зв'язку. Це може бути пов'язано з тим, що оцінки були отримані від критиків з США, тоді як розглядалися глобальні доходи і при наявності оцінок іноземних критиків, результати могли б відрізнятися.

Дослідження статистичних закономірностей в рейтинговій поведінці фільмів провели Марлон Рамос, Анджело Кальвао та Селія Антенеодо у 2015 році [19]. Вони аналізували закономірність розподілу оцінок та кількість голосів для фільмів на основі даних з IMDB та Netflix. Основна увага приділялась розумінню того, як різні фактори впливають на рейтинг фільмів.

Автори використовують концепції статистичної фізики для аналізу соціальної динаміки, яка впливає на оцінювання фільмів. Досліджуючи розподіл голосів, вони виявили, що є позитивна кореляція між середнім рейтингом фільму та кількістю голосів, іншими словами чим більше оцінок фільму, тим кращий у нього рейтинг.

В результаті їхнього аналізу за різними критеріями, виявлено, що новіші фільми мають схильність отримувати більше голосів. Цікавим є той факт, що жанри фільмів не мали значного впливу на поведінку голосування, за винятком документальних та короткометражних фільмів. Ще однією важливою характеристикою є те, що фільми з вищим бюджетом отримують більше рейтингових оцінок, показуючи позитивний зв'язок між бюджетом фільму та кількістю голосів, що вказує на те, що такі фільми мають більшу популярність.

Автори вказують, що дивлячись на те, як голоси поширюються і накопичуються через науковий підхід, зокрема, використовуючи методи статистичної фізики, можна глибше зрозуміти те, як формується рейтинг фільмів.

Описуючи цей процес більш детально, вони говорять про механізм каскаду голосів, коли один відгук чи оцінка, залишена кимось, заохочує інших також залишити свій голос. Тобто, існує певний ланцюговий реакційний ефект, коли активність однієї особи викликає реакцію інших. Тому, важливо не лише дивитись на індивідуальні оцінки, але і розуміти цей груповий, майже «ланцюговий» процес оцінювання, який може істотно вплинути на загальний рейтинг фільму.

В цілому можна зробити висновок, що немає однозначної відповіді на питання про те, які фактори визначають рейтинг та касовий успіх фільму. Результати досліджень варіюються в залежності від використаних даних і методів, і це підкреслює складність прогнозування успіху фільмів на кіноекранах.

Висновок до першого розділу

В даному розділі було розглянуто роль кіноіндустрії в сучасному суспільстві, вплив фільмів на культурні, освітні, економічні та інші аспекти людського життя. Досліджено історичний аспект формування рейтингу фільмів, а саме тривалу еволюцію - від примітивних відгуків чи усних

рекомендацій до складних структур, що об’єднують думки як професійних аналітиків, так і масової аудиторії. Описано етапи формування касових зборів фільмів та їхнє неперервне вдосконалення та пристосування системи оцінки комерційного успіху кінопроектів до змін, що відбуваються в кіноіндустрії та суспільстві. Проведено аналіз основних досліджень і публікацій з обраної теми.

РОЗДІЛ 2
МЕТОДИКА ДОСЛІДЖЕННЯ ТА ОБРОБКА ДАНИХ

2.1 Постановка задачі

В рамках вказаної мети та завдань дослідження, поставлено наступні конкретні задачі:

- Вибір інструментів та технологій для аналізу “tmdb_movies.csv” датасету;
- Визначення та аналіз властивостей датасету “tmdb_movies.csv”, виявлення ключових параметрів та характеристик фільмів;
- Підготовка та очищення даних;
- Кореляційний та візуальний аналіз датасету, для вивчення факторів, які впливають на рейтинг та касові збори фільмів;
- Вибір та реалізація відповідних моделей машинного навчання для виявлення факторів впливу на рейтинг та касові збори фільмів;
- Оцінка точності та ефективності розроблених моделей;
- Визначення ключових факторів, що мають значущий вплив на рейтинг та касові збори фільмів;
- Подання аргументованих висновків та рекомендацій заснованих на отриманих результатів дослідження.

Виконання цих задач допоможе детально вивчити внутрішні та зовнішні фактори, які впливають на рейтинг та касові збори фільмів, і, отже, дозволить зрозуміти, які аспекти детермінують комерційний успіх фільмів на кіно ринку.

2.2 Опис обраних інструментів та технологій для аналізу та обробки даних

У сучасному світі аналіз даних стає все більш актуальним та важливим інструментом в дослідницькій роботі, а також у прийнятті

обґрунтованих рішень в бізнесі, науці та інших сферах. У кіноіндустрії, аналіз даних може вказати на тенденції та фактори, які визначають успіх фільму. Вибір правильних інструментів та технологій для аналізу та обробки даних є ключовим для отримання надійних результатів. Вони допоможуть виявити ключові показники, які впливають на комерційний та критичний успіх кінострічок.

2.2.1 Мова програмування R

Мова програмування R є однією з найкращих мов для статистичного аналізу та візуалізації даних. Станом на вересень 2023 року, за індексом PYPL, вона входила у сімку найпопулярніших мов програмування у всьому світі [20]. Вона спеціалізується на аналізі та обробці даних, створенні статистичних моделей, візуалізації даних через графіки та діаграми, розробці та використанні алгоритмів машинного навчання та є особливо корисною в наукових дослідженнях.

Мова R виникла як результат спільної розробки Росса Іхаки та Роберта Джентельмена з Університету Окленда в Новій Зеландії. Наразі її розвитком займається команда розробників R Development Core Team . R є реалізацією мови програмування S та включає в себе семантику лексичного огляду, яка була натхненна мовою Scheme. Проект був задуманий у 1992 році, перша версія була випущена у 1995 році, а стабільна бета-версія стала доступною у 2000 році [21].

Переваги мови програмування R:

- Відкритий код: R є вільним та відкритим програмним забезпеченням, дозволяючи використовувати, модифікувати та оптимізувати його в будь-який час та будь-де. Всі користувачі можуть вносити свій вклад у розвиток мови, пропонуючи нові пакети, виправляючи помилки та покращуючи код;
- Крос-платформеність: R не залежить від платформи та може працювати на всіх операційних системах, включаючи UNIX,

Windows та MacOS [22];

- Широкі можливості: R має великий набір вбудованих функцій для аналізу даних, а також можливість розширення функціональності за допомогою пакетів;
- Доступність пакетів: однією з найбільших переваг R є її система пакетів. Існують тисячі пакетів, розроблених для різноманітних завдань, від простої обробки даних до спеціалізованих статистичних методів і технік машинного навчання [23];
- Спільнота: за роки існування R з’явилася активна спільнота користувачів та розробників, яка підтримує новачків, створює нові пакети та відіграє свою роль у розвитку мови;
- Візуалізація даних: R відомий своїми можливостями у візуалізації даних за допомогою графіків та діаграм, що полегшує інтерпретацію та аналіз результатів;
- Інтеграція: R можна використовувати у поєднанні з іншими мовами програмування [24];
- Швидка адаптація нових технологій і концепцій: нові технології та концепції часто з'являються спочатку в R, що робить мову актуальною та на передовій технологічного прогресу.

Мова R особливо підходить для людей, які працюють з даними. Вона надає гнучкість програмування з багатими статистичними можливостями, роблячи її відмінним вибором для аналітиків, дослідників та науковців.

2.2.2 RStudio

RStudio є високофункціональним інтегрованим середовищем розробки (IDE) для мови програмування R. Воно розроблене в 2011 році компанією Posit, яка спеціалізується на науці про дані з відкритим вихідним кодом [25]. Широке визнання RStudio в спільноті аналітиків даних пояснюється його інтегрованим та простим підходом до аналізу

даних, візуалізації та статистичного моделювання.

RStudio визнається як інтегроване середовище розробки (IDE) для мови програмування R, але ефективніше його розглядати як комплекс засобів, спроектованих для підтримки аналітиків у керуванні, візуалізації, моделюванні даних та розгортанні моделей машинного навчання.

RStudio служить редактором коду з функціями підсвічування синтаксису, автозавершення коду та інструментами для налагодження. Він створений для безпосереднього написання коду на R, і маючи ці функції процес кодування стає плавнішим і ефективнішим. Крім того, RStudio надає інтерактивну консоль для виконання частин або цілих сценаріїв коду R та перегляду результатів у реальному часі [26].

На відміну від інших IDE, RStudio пропонує браузер робочого простору, який моніторить використані змінні, функції, списки та кадри даних у поточній сесії. Наявність візуального відображення об’єктів, з якими проводиться робота, є важливою можливістю. Також, RStudio включає вікно для побудови графіків, дозволяючи відображення графіків, які генеруються під час проведення дослідницького аналізу даних. Ці графіки можна редагувати та зберігати безпосередньо.

RStudio ефективно інтегрується з іншими засобами, такими як Git для контролю версій, дозволяючи відстежувати та керувати змінами в коді. Це особливо корисно при спільній роботі над проектами. RStudio також підтримує Shiny, дозволяючи створення веб-додатків та інтерактивних інформаційних панелей без знань веб-розробки. Використовуючи R Markdown і його інтерфейс блокнота, можна інтегрувати код, текст та графіки в один документ, що зручно для представлення аналітичних досліджень у вигляді інформативних звітів [27].

RStudio стає незамінним інструментом для розробників, аналітиків та науковців, які використовують R для аналізу даних, статистичних розрахунків, візуалізації даних, та машинного навчання. Це середовище не

тільки робить процес написання коду більш зручним та ефективним, але також надає обширні можливості для співпраці та обміну результатами роботи.

2.2.3 Tidyverse

Tidyverse - це набір пакетів для мови програмування R, призначений для обробки, аналізу та візуалізації даних. Він базується на ідеї "tidy data" (впорядкованих даних) [28], яка полягає в тому, що дані повинні бути легкими для обробки та аналізу. Цей набір пакетів зробив аналіз даних в R більш зручним та ефективним завдяки однорідному та логічному підходу до обробки даних.

Tidyverse включає в себе такі основні пакети:

- “ggplot2”: один з найпотужніших пакетів для візуалізації даних в R. Він дозволяє створювати графіки високої якості з великою гнучкістю та контролем, інтегрується з іншими інструментами аналізу даних і забезпечує зручний та ефективний підхід до візуалізації даних [29];
- “dplyr”: пакет, який надає потужні інструменти для обробки та маніпулювання даними. Він містить функції для фільтрації, групування, сортування, вибору та перетворення даних [30];
- “tidyr”: пакет для мови програмування R, який допомагає у перетворенні та організації даних у форматі, який називається “tidy data”. Основна ідея “tidy data” полягає в тому, що кожен стовпець таблиці представляє одну змінну, кожен рядок представляє одну спостережувальну одиницю, а кожна таблиця представляє одну таблицю даних [31]. Він допомагає зберігати дані в чистому і зрозумілому вигляді, що полегшує подальший аналіз та візуалізацію даних;
- “readr”: пакет, який розроблений для ефективного зчитування і завантаження даних з різних форматів, таких як CSV, TSV, Excel і

інших. Він спрощує процес читання даних та автоматично визначає типи даних, щоб забезпечити консистентність і чистоту даних [32];

- “purrr”: пакет, який надає інструменти для роботи з функціями та списками. Він спрощує і узагальнює роботу з ітераціями, векторизацією та обробкою даних, що розташовані в різних форматах, включаючи списки та фрейми даних [33]. Він став важливою складовою багатьох проектів у середовищі R, особливо в контексті аналізу даних та моделювання;
- “stringr”: пакет для мови програмування R, призначений для роботи з рядками та текстовими даними. Він допомагає аналізувати текстову інформацію більш ефективно та точно [34];
- “tibble”: пакет, який надає структуру даних, яка є більш сучасною та зручною для роботи, ніж звичайні data.frame. Це важливий інструмент при роботі з даними в R, особливо якщо потрібна зручна, читабельна та безпечна структура для зберігання та обробки інформації [35];
- “forcats”: пакет, який надає потужні інструменти для роботи з факторними змінними у мові R. Він дозволяє легко категоризувати, маніпулювати та обробляти факторні змінні, що є важливим кроком при аналізі та моделюванні категоріальних даних [36].

Ці пакети працюють разом, щоб забезпечити зручний та ефективний аналіз даних у мові R. Tidyverse став дуже популярним серед аналітиків даних та науковців через його чітку структуру та простоту використання.

2.3 Вибір та опис набору даних

Набір даних, обраний для цього дослідження отриманий з веб-сервісу TMDB. TMDB - це база даних, яка спеціалізується на

зберіганні інформації про фільми та телевізійні шоу [37]. Хоча вона має певні схожості з базою даних IMDB, між ними існують деякі основні відмінності:

- Історія та походження:
    - IMDB є однією з найдавніших та найвідоміших баз даних про кіно та телевізійні шоу. Вона була заснована у 1990 році.
    - TMDB є менш відомим, але все більш популярним джерелом інформації про фільми та телесеріали. Вона була створена спільнотою і запущена у 2008 році [38].
- Власність:
    - IMDB є власністю Amazon і є частиною інших послуг від цієї компанії, таких як Amazon Prime Video.
    - TMDB є відкритою та незалежною платформою, і вона розвивається завдяки внеску спільноти і розробників.
- Користувачі та спільнота:
    - IMDb має велику користувацьку базу і використовує рейтинги і відгуки користувачів та критиків для оцінки фільмів [39].
    - TMDb також має активну спільноту користувачів, які додають інформацію і відгуки про фільми.
- API:
    - IMDb має своє API, але воно обмежене та платне для деяких видів доступу.
    - TMDb має безкоштовне та доступне для розробників API, яке дозволяє створювати додатки, сайти і послуги, використовуючи їхні дані.
- Оцінки і рейтинги:
    - IMDb відомий своїми числовими рейтингами та оцінками,

які надаються фільмам.

- TMDb також має рейтинги, але вони можуть бути менш формалізованими і відрізнятися від IMDb.

Обраний для дослідження набір зберігається у файлі “tmbd_movies.csv”. Цей датасет містить широкий спектр інформації про різні фільми, що дозволяє провести глибокий аналіз та визначити показники, які впливають на рейтинг та касові збори фільмів.

Для початку роботи з набором даних потрібно його завантажити до середовища аналізу.

Лістинг коду 2.1:

```
# Завантаження пакетів та набору даних
library(tidyverse)
TMDB <- read_csv("U:/Master's_work_Boiko/tmdb_movies.csv")
```

У датасеті зібрані дані про фільми з 1970 року по 2015 рік. Набір даних містить 10866 спостережень за наступною 21 змінною (рис. 2.1):

1. **id:** унікальний ідентифікатор фільму в наборі даних;
2. **imdb_id:** ідентифікатор фільму на IMDB;
3. **popularity:** показник, який вказує на ступінь популярності фільму серед глядачів;
4. **budget:** бюджет фільму, тобто вартість його виробництва;
5. **revenue:** касові збори фільму;
6. **original_title:** оригінальна назва фільму;
7. **cast:** список акторів, які зіграли головні ролі у фільмі;
8. **homepage:** веб-сайт фільму;
9. **director:** режисер фільму;
10. **tagline:** слоган або заголовок, пов’язаний з фільмом;
11. **keywords:** ключові слова або терміни, які дають характеристику фільму;
12. **overview:** короткий опис сюжету фільму;

13. **runtime:** тривалість фільму у хвилинах;
14. **genres:** жанр або жанри фільму;
15. **productions_companies:** назви студій або компаній, які виробляли фільм;
16. **release_date:** дата виходу фільму;
17. **vote_count:** кількість голосів, які фільм отримав від глядачів на сервісі TMDB;
18. **vote_average:** середній рейтинг фільму, який обчислений на основі голосів глядачів;
19. **reliase_year:** рік випуску фільму;
20. **budget_adj:** бюджет фільму, скоригований з урахуванням інфляції;
21. **revenue_adj:** Касові збори фільму, скориговані з урахуванням інфляції.

```
 [1] "id"           "imdb_id"         "popularity"
 [4] "budget"       "revenue"         "original_title"
 [7] "cast"         "homepage"        "director"
[10] "tagline"      "keywords"        "overview"
[13] "runtime"      "genres"          "production_companies"
[16] "release_date" "vote_count"      "vote_average"
[19] "release_year" "budget_adj"      "revenue_adj"
```

Рис. 2.1 - Результат виконання функції “names()”

Цей набір даних дозволяє проаналізувати взаємозв'язки між різними показниками та визначити, які з них найбільше впливають на успіх фільму в комерційному та критичному відношенні. Варто зазначити, що він містить як кількісні, так і категоріальні дані, тому для аналізу може бути використано різні методи статистичної обробки та машинного навчання.

2.4 Підготовка та очищення даних

Перед початком будь-якого аналізу даних критично важливо підготувати та очистити набір даних, з яким буде проводитись дослідження. Це важливий крок, оскільки якість даних напряму впливає на

достовірність висновків, отриманих від аналітичних моделей.

Очищення даних – це процес коригування, вдосконалення та організації інформації у наборі даних з метою досягнення єдності та готовності для подальшого аналізу [40]. Цей процес обумовлює вилучення пошкоджених, некоректних або неістотних даних та їхнє перетворення на структуру, доступну для обробки, щоб забезпечити ефективний аналіз.

Для забезпечення якості даних та їхньої придатності для використання, слід виконати наступні етапи процесу очищення даних:

- **Вилучення нерелевантних даних**. Слід видалити всі дані, які не мають відношення до аналітичних потреб проекту;
- **Вилучення дублікатів.** Пошук та видалення однакових записів, якщо такі існують, для уникнення спотворених результатів аналізу;
- **Виправлення структурних вад**. Структурні вади охоплюють аспекти, такі як орфографічні вади, невірні найменування, неправильне застосування великих літер, невірний вибір слів та інше. Ці помилки можуть відігравати роль у аналізі даних, адже навіть якщо для людини вони є очевидними, більшість систем машинного навчання не в змозі ідентифікувати ці помилки, що може викривити результати аналізу;
- **Усунення пропущених значень.** Слід визначити відсутні дані, після чого їх заповнити або видалити;
- **Фільтрація викидів.** Викиди - це точки даних, які значно відрізняються від інших, виходячи за межі норми та можуть сильно спотворити аналіз [41]. Слід їх ідентифікувати та відфільтрувати;
- **Перевірка даних.** Після очищення набору даних, останнім кроком є його фінальна перевірка для забезпечення точності та якості даних, перед їх аналізом.

Отже, спочатку потрібно видалити не потрібні для дослідження змінні. Вибір змінних для видалення базується на їхній релевантності та потенційному впливі на досліджувану залежність.

Обрані для видалення змінні:

- **id** та **imdb_id:** унікальні ідентифікатори фільмів, хоча й корисні для зберігання та з'єднання даних між різними базами даних, але вони не мають практичного впливу на рейтинг або касові збори фільму;
- **popularity:** вона розраховується для внутрішніх потреб TMDB і враховує ряд внутрішніх чинників, призначених для покращення результатів пошуку на платформі. Це динамічний показник, що змінюється щодня, маючи плаваюче значення без верхньої межі, що робить його складним для інтерпретації;
- **cast**: імена акторів представлені як текстові рядки. Для їхнього аналізу потрібно перетворити ці імена на числові ідентифікатори, створивши велику кількість додаткових змінних. Це призводить до дуже великої кількості унікальних факторів, що збільшує складність даних та може призвести до проблеми “прокляття розмірності” [42]. Також не всі актори в однаковій мірі впливають на успіх фільму. Головна роль може мати значний вплив, тоді як менш відомі актори можуть не мати такого великого значення. Тому агрегування всіх акторів у фільмі може не відображати реальний їхній вплив на успіх фільму;
- **director:** режисери, безумовно, мають значний вплив на якість та успіх фільму. Однак, як і з акторами, існує проблема великої кількості унікальних режисерів, щоб їх аналізувати потрібно перетворити їхні імена на числові ідентифікатори, що також створює додаткові змінні;
- **homepage:** веб-сторінка фільму не має прямого впливу на його

обрані для дослідження показники;

- **tagline:** слоган фільму також, скоріш за все, не вплине на аналіз;
- **keywords:** ключові слова можуть бути цікавими для аналізу контенту, але для нашого дослідження вони менш інформативні;
- **overview:** короткий огляд фільму не має статистичного значення для аналізу;
- **production_companies**: компанії, які створили фільм, можуть бути важливими, але натомість можуть ускладнити аналіз через велику варіативність;
- **release_date:** дата випуску вже представлена змінною “release_year”;
- **budget_adj** і **revenue_adj:** ці змінні показують бюджет та касові збори, але з урахуванням інфляції. Вони видаляються для спрощення аналізу, так як основні змінні budget та revenue вже присутні в наборі даних.

З урахуванням вищезазначеного можна видалити ці змінні з набору даних, щоб спростити аналіз та зосередитись на ключових факторах, які найбільше впливають на рейтинг та касові збори фільму.

Лістинг коду 2.2:

```
# Видалення нерелевантних даних
TMDB_cleaned <- TMDB %>% select(-c(id, imdb_id, popularity, cast, homepage, director, tagline, keywords, overview, production_companies, release_date, budget_adj, revenue_adj ))
```

У результаті цього коду буде створено новий DataFrame “TMDB_cleaned”, який не буде містити вказані вище змінні, залишивши тільки релевантні для аналізу атрибути.

Перевірка та обробка відсутніх значень є ще одним критично важливим етапом підготовки даних. Відсутні значення можуть впливати на результати аналізу та на вивчення моделі, тому перед проведенням аналізу

важливо визначити та вирішити, як обробити ці значення, наприклад, видалити або замінити їх.

Для початку, потрібно дізнатися, у яких змінних існують відсутні значення, та яка їхня кількість (рис. 2.2).

Лістинг коду 2.3:

```
# Перевірка на кількість NA значень у кожній змінній
na_count <- sapply(TMDB_cleaned, function(y) sum(length(which(is.na(y)))))
print(na_count)
```

```
      budget      revenue original_title      runtime       genres
           0            0              0            0           23
  vote_count vote_average   release_year
           0            0              0
```

Рис. 2.2 - Кількість відсутніх значень у кожній змінній

Відсутні значення містяться у змінній “genres”. Існує кілька способів обробки відсутніх даних, але в даному контексті, найкраще буде просто видалити рядки, які містять пропуски.

Деякі змінні у нашому наборі можуть містити значення “0” (рис 2.3). Це означає, що дані можуть бути фактично відсутніми або неправильними, особливо у таких змінних як “budget”, “revenue” та “runtime”, де “0” не є реалістичним значенням. У таких випадках аналогічно до відсутніх значень, значення “0” можна видалити.

Лістинг коду 2.4:

```
# Перевірка на кількість 0 значень у кожній змінній
zero_count <- sapply(TMDB_cleaned, function(y) sum(y == 0, na.rm = TRUE))
print(zero_count)
```

```
      budget      revenue original_title      runtime       genres
        5674         5993              0           30            0
  vote_count vote_average   release_year
           0            0              0
```

Рис 2.3 - Кількість нульових значень у кожній змінній

Нульові значення у “budget” та “revenue” вказують на відсутність фінансової інформації для деяких фільмів, і можуть спотворити результати

аналізу. До прикладу, при спробі знайти кореляцію між бюджетом та касовими зборами, їхні нульові рядки будуть сильно впливати на точність наших результатів.

Значення нуль у змінній “runtime” вказує на відсутність інформації про тривалість фільму, що також є критичним параметром, який повинен бути взятий до уваги.

Видалення рядків із нульовими значеннями може бути обгрунтованим, якщо ці значення вважаються недійсними або якщо їх заміна може призвести до спотворення даних. У нашому випадку, оскільки ми не можемо точно замінити нульові значення на якісь інші значення (наприклад, середнє або медіанне), видалення цих рядків є розумним рішенням (рис. 2.4).

Лістинг коду 2.5:

```
# Видалення рядків, де budget, revenue та runtime дорівнює 0
TMDB_cleaned <- TMDB_cleaned %>%
  filter(budget > 0 & revenue > 0 & runtime > 0)
# Визначення мінімального значення для змінних budget, revenue та runtime
min_data <- sapply(TMDB_cleaned[, c("runtime", "budget", "revenue")], min, na.rm
= TRUE)
```

```
runtime  budget revenue
     15       1       2
```

Рис 2.4 - мінімальні значення змінних “budget”, “revenue” та “runtime”

Важливим етапом очищення даних є видалення дублікатів, якщо вони існують. Дублікати рядків у наборі даних можуть вплинути на результати аналізу, створюючи зміщення або невірні висновки, тому їх потрібно виявити. Вони можуть виникнути через помилки під час збору даних або злиття різних джерел даних.

Перевіримо наш DataFrame на наявність дублікатів рядків. Якщо вони будуть знайдені, то видалимо їх за допомогою функції “distrinct()” від “dplyr”.

Лістинг коду 2.6:

```
# Перевірка на наявність дублікатів
original_nrow <- nrow(TMDB_cleaned)
duplicated_rows <- TMDB_cleaned[duplicated(TMDB_cleaned),]
# Видалення дублікатів
TMDB_cleaned <- distinct(TMDB_cleaned)
# Перевірка результату
new_nrow <- nrow(TMDB_cleaned)
print(paste("Кількість рядків до видалення дублікатів: ", original_nrow))
print(paste("Кількість рядків після видалення дублікатів: ", new_nrow))
```

В результаті перевірки було виявлено та видалено один дублікат даних (рис. 2.5).

```
[1] "Кількість рядків до видалення дублікатів:  3855"
[1] "Кількість рядків після видалення дублікатів:  3854"
```

Рис 2.5 - Результат видалення дублікатів

Ще одним етапом обробки даних є фільтрація або коригування викидів. Проте у нашому випадку вони можуть представляти реальні дані, які важливі для дослідження.

Важливо розрізняти викиди, які є аномаліями, від викидів, які є дійсними значущими даними. Успішні фільми з високими касовими зборами чи бюджетом слід зберегти в датасеті, оскільки вони мають велике значення для аналізу. З іншого боку, фільми з дуже низькими касовими зборами та бюджетом, можуть бути вказівниками на помилки в даних. Наприклад, на рисунку 2.4 показано мінімальне значення бюджету фільму, яке дорівнює одиниці. Встановлення порогових значень для змінних “budget” та “revenue”, допоможе відфільтрувати фільми, які мають надзвичайно низькі касові збори або бюджет. Такий підхід зазначить, що наш аналіз базується на даних фільмів, які мали певний мінімальний рівень комерційного успіху або інвестицій (рис. 2.6).

Лістинг коду 2.7:

```
# Встановлення порогового значення
revenue_threshold <- 11000
budget_threshold <- 1000
# Відфільтровування фільмів з касовими зборами та бюджетом нижче порогового значення
TMDB_cleaned <- TMDB_cleaned %>%
  filter(revenue > revenue_threshold & budget > budget_threshold)
# Перевірка результату
summary(TMDB_cleaned$budget )
summary(TMDB_cleaned$revenue )
```

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   7000  10000000  25000000  37609526  50000000 425000000
    Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
   13308   14460000   46236000  109155433  126216940 2781505847
```

Рис 2.6 - Кількісні зведення по “budget” та “revenue” після фільтрації

Після всіх етапів очищення даних, слід здійснити фінальну перевірку для впевненості, що дані готові для аналізу. Вона може включати перевірку типів даних, розміру DataFrame, а також загальний огляд даних.

Перевірка типів даних критично важлива, оскільки неправильні типи даних можуть призвести до проблем під час аналізу. Визначення правильності типів даних змінних є першочерговим завданням, щоб уникнути конфліктів та помилок у майбутньому. Функція “str()” в R дозволяє швидко переглянути імена, типи даних, а також перші кілька значень усіх змінних у нашому наборі (рис. 2.7).

```
tibble [3,801 × 8] (S3: tbl_df/tbl/data.frame)
 $ budget        : num [1:3801] 150000000 150000000 110000000 200000000 190000000
135000000 155000000 108000000 74000000 175000000 ...
 $ revenue       : num [1:3801] 1513528810 378436354 295238201 2068178225 1506249360
...
 $ original_title: chr [1:3801] "Jurassic World" "Mad Max: Fury Road" "Insurgent" "Star
Wars: The Force Awakens" ...
 $ runtime       : num [1:3801] 124 120 119 136 137 156 125 141 91 94 ...
 $ genres        : chr [1:3801] "Action|Adventure|Science Fiction|Thriller"
"Action|Adventure|Science Fiction|Thriller" "Adventure|Science Fiction|Thriller"
"Action|Adventure|Science Fiction|Fantasy" ...
 $ vote_count    : num [1:3801] 5562 6185 2480 5292 2947 ...
 $ vote_average  : num [1:3801] 6.5 7.1 6.3 7.5 7.3 7.2 5.8 7.6 6.5 8 ...
 $ release_year  : num [1:3801] 2015 2015 2015 2015 2015 ...
 - attr(*, "na.action")= 'omit' Named int [1:23] 425 621 998 1713 1898 2371 2377 2854
3280 4548 ...
  ..- attr(*, "names")= chr [1:23] "425" "621" "998" "1713" ...
```

Рис 2.7 - Результат використання функції “str()” для виведення структури набору даних

З наведеного рисунку можемо побачити, що всі змінні в наборі даних мають відповідні їм типи.

Наступним кроком буде перевірка кількості рядків та стовпців у нашому DataFrame, щоб зрозуміти, скільки інформації доступно для аналізу, і як багато даних було втрачено під час попередніх етапів обробки (рис 2.8).

```
[1] "Кількість рядків:  3801"
[1] "Кількість стовбців:  8"
```

Рис 2.8 - Кількість рядків і стовпців, що залишились після обробки

Під час обробки була видалена досить значна частина даних. Проте не слід забувати, що дані які було видалено містили пропущені, нульові, повторювані або помилкові значення, які завадили б нам отримати достовірні висновки при аналізі факторів, які впливають на рейтинг та касові збори фільмів.

Останнім кроком в обробці даних, буде їх візуалізація за допомогою функції “fix()” [43] (рис. 2.9). Використовуючи дану функцію, можна візуально переглядати свій датасет, що допомагає у виявленні аномалій, помилок чи інших неправильностей у даних. За допомогою цього інструменту, можна також легко змінювати значення прямо в DataFrame, якщо це потрібно, що робить функцію універсальним інструментом для огляду та корекції даних перед аналізом.

Лістинг коду 2.8:

```
# Використання функції fix() для візуалізації даних
fix(TMDB_cleaned)
```

	budget	revenue	original_ti>	runtime	genres	vote_count	vote_av>	relea>
1	150000000	1513528810	Jurassic Wo>	124	Action>	5562	6.5	2015
2	150000000	378436354	Mad Max: Fu>	120	Action>	6185	7.1	2015
3	110000000	295238201	Insurgent	119	Advent>	2480	6.3	2015
4	200000000	2068178225	Star Wars: >	136	Action>	5292	7.5	2015
5	190000000	1506249360	Furious 7	137	Action>	2947	7.3	2015
6	135000000	532950503	The Revenant	156	Wester>	3929	7.2	2015
7	155000000	440603537	Terminator >	125	Scienc>	2598	5.8	2015
8	108000000	595380321	The Martian	141	Drama|>	4572	7.6	2015
9	74000000	1156730962	Minions	91	Family>	2893	6.5	2015
10	175000000	853708609	Inside Out	94	Comedy>	3935	8	2015
11	245000000	880674609	Spectre	148	Action>	3254	6.2	2015
12	176000003	183987723	Jupiter Asc>	124	Scienc>	1937	5.2	2015
13	15000000	36869414	Ex Machina	108	Drama|>	2854	7.6	2015

Рис 2.9 - Візуалізація даних за допомогою функції “fix()”

Уважний підхід до перевірки даних на кожному етапі процесу допомагає забезпечити вірність і достовірність дослідження та аналізу, мінімізуючи ризик впливу аномалій, помилок та неправильної інтерпретації даних. Після ретельної обробки ми маємо повністю підготовлений і перевірений набір даних, який готовий до подальшого аналізу та моделювання.

2.5 Вибір методів аналізу даних та їх обґрунтування

У сучасному світі аналіз даних стає все більш важливим інструментом при прийнятті рішень. Існує безліч методів, які можна застосовувати для аналізу даних, і вибір конкретного методу вимагає розуміння специфіки датасету і цілей дослідження.

Дескриптивна статистика відіграє ключову роль, коли ми хочемо отримати базове розуміння датасету. Це перший крок, який дає можливість вивчити основні характеристики даних, такі як середні значення, медіана та стандартне відхилення. Ця інформація надає нам глобальне розуміння розподілу даних, виявляє тенденції, відхилення та можливі аномалії [44].

Після того, як ми отримаємо загальний огляд даних, важливо дослідити можливі взаємозв'язки між різними змінними. Це можна зробити за допомогою кореляційного аналізу [45]. Він дозволяє визначити, чи

існують значущі взаємозв'язки між показниками, наприклад, між бюджетом фільму і його касовими зборами. Вивчення цих залежностей є ключовим для розуміння, які фактори найбільше впливають на рейтинг фільму або його касовий успіх.

Часто, при роботі з числовими даними, можна зосередитися виключно на цифрах та статистичних показниках. Однак числовий аналіз, не зважаючи на його важливість, відображає лише частину інформації, яка може бути отримана з датасету.

Візуальний аналіз даних, який включає в себе створення графіків, діаграм і інших візуалізацій, дозволяє додати новий вимір до сприйняття даних. Графічне представлення інформації може "оживити" числа, перетворивши їх на зображення, які легко інтерпретувати та зрозуміти на інтуїтивному рівні.

Діаграми та графіки відкривають можливість спостерігати за динамікою, трендами та взаємозв'язками між різними змінними. Вони можуть відобразити, як одна змінна змінюється у відповідь до іншої, або ж виявити аномалії в даних, які можуть бути приховані при простому числовому аналізі.

Комбінація числового та візуального аналізів даних стає ключовим елементом глибокого розуміння і інтерпретації інформації, яка міститься у наборі даних.

Коли дослідники або аналітики стикаються з завданням прогнозування майбутніх подій або визначенням того, як одна змінна може впливати на іншу, вони шукають засоби, які дозволили б зрозуміти ці складні взаємодії. У таких ситуаціях регресійний аналіз виступає як незамінний інструмент, що надає глибокий аналітичний вигляд на досліджувані процеси. Він дає змогу не просто визначити наявність зв'язку між змінними, але й кількісно оцінити цей зв'язок [46]. Це означає, що за допомогою регресії можна з'ясувати, наскільки сильно змінна X впливає на

змінну Y.

Також важливим аспектом регресійного аналізу є можливість визначення статистичної значущості показників. Це дозволяє відрізнити дійсні, суттєві взаємодії від тих, які можуть бути просто випадковими або незначущими.

Крім того, регресійний аналіз надає можливість створювати прогнозні моделі на основі існуючих даних. Ці моделі можуть бути використані для вивірених прогнозів майбутніх подій, базуючись на інформації, яка вже відома.

Таким чином, регресійний аналіз надає дослідникам інструменти для детального вивчення взаємодій між змінними та створення обґрунтованих прогнозів, виходячи з цих знань.

Також, для дослідження того, які чинники впливають на успіх фільму, було обрано метод випадкового лісу, який є надійним інструментом у сфері аналізу даних [47]. Випадковий ліс, або “Random Forest”, це метод в машинному навчанні, який використовує багато дерев рішень для вирішення проблеми і потім об'єднує їх результати для отримання більш точного та стабільного прогнозу.

Він може знаходити складні нелінійні взаємозв'язки між змінними, що часто зустрічаються в реальних датасетах.

Випадковий ліс надає корисну інформацію про важливість кожної змінної для прогнозування, що дозволяє зрозуміти, які характеристики мають найбільший вплив на рейтинг і доходи.

Він широко використовується у наукових дослідженнях і підтримується багатьма статистичними пакетами в мові R [48].

Таким чином, випадковий ліс — це потужний і гнучкий інструмент, який може дати нам більш точний та надійний прогноз або класифікацію, використовуючи колективну мудрість багатьох дерев рішень.

Висновок до другого розділу

В даному розділі було докладно визначено завдання дослідження та обрано інструменти для його виконання. Зокрема, було розглянуто особливості мови програмування R, яка вибрана для аналізу даних, а також середовища RStudio та пакет Tidyverse, що дозволить оптимізувати робочий процес та забезпечити гнучкість аналізу.

Важливим етапом дослідження стала вибірка і опис набору даних, а також їх підготовка і очищення, що забезпечує якісний аналіз без спотворень або невідповідностей в даних.

Детально обговорено методику аналізу даних, вибрані методи, їх особливості та причини вибору для цього конкретного дослідження. Таким чином, у цьому розділі було закладено міцний фундамент для подальшого аналізу даних та отримання об'єктивних результатів дослідження.

РОЗДІЛ 3
АНАЛІЗ ДАНИХ

3.1 Описова статистика

Після підготовчої роботи, здійсненої у другому розділі, наступний етап дослідження зосереджується на аналізі даних.

Передусім, слід провести описову статистику. Це дозволить отримати загальне уявлення про розподіл та тенденції даних, які включають в себе різноманітні кількісні показники. Описова статистика є критично важливою, оскільки вона закладає фундамент для подальшого аналізу.

Спочатку дізнаємось та проаналізуємо кількісні зведення по кожній змінній в наборі даних за допомогою функції “summary()” (рис. 3.1).

```
     budget             revenue          original_title        runtime
 Min.   :     7000   Min.   :     13308   Length:3801        Min.   : 26.0
 1st Qu.: 10000000   1st Qu.:  14460000   Class :character   1st Qu.: 96.0
 Median : 25000000   Median :  46236000   Mode  :character   Median :106.0
 Mean   : 37609526   Mean   : 109155433                      Mean   :109.4
 3rd Qu.: 50000000   3rd Qu.: 126216940                      3rd Qu.:119.0
 Max.   :425000000   Max.   :2781505847                      Max.   :338.0
    genres            vote_count      vote_average    release_year
 Length:3801        Min.   :  10.0   Min.   :2.200   Min.   :1960
 Class :character   1st Qu.:  73.0   1st Qu.:5.700   1st Qu.:1995
 Mode  :character   Median : 207.0   Median :6.200   Median :2004
                    Mean   : 533.8   Mean   :6.174   Mean   :2001
                    3rd Qu.: 584.0   3rd Qu.:6.700   3rd Qu.:2010
                    Max.   :9767.0   Max.   :8.400   Max.   :2015
```

Рис 3.1 - Кількісні зведення по кожній змінній

Для кількісних змінних функція summary() видає мінімальне та максимальне значення, середнє арифметичне, медіану, нижній (1st Qu.) та верхній (3rd Qu.) квартиль. Нижній квартиль - значення, яке 25% значень у вибірці не перевищують, а верхній квартиль - значення, яке не перевищують 75% значень.

Основні висновки по кількісним зведенням:

- budget: середній бюджет фільмів становить приблизно 36.7

мільйонів доларів, з мінімальним значенням 7 тисяч доларів та максимальним - 425 мільйонів доларів;

- revenue: Середні касові збори становлять близько 109.2 мільйонів доларів. Максимальні збори - 2.78 мільярда доларів, що вказує на велику різноманітність у фінансовому успіху фільмів;
- runtime: Середня тривалість фільму в наборі даних - близько 109 хвилин;
- vote_count: В середньому, на кожен фільм припадає близько 534 голосів;
- vote_average: Середня оцінка фільмів становить 6.17 бали з 10.

Для більш наглядного представлення слід побудувати графіки розподілу основних змінних (рис. 3.2, рис. 3.3, рис. 3.4, рис. 3.5).

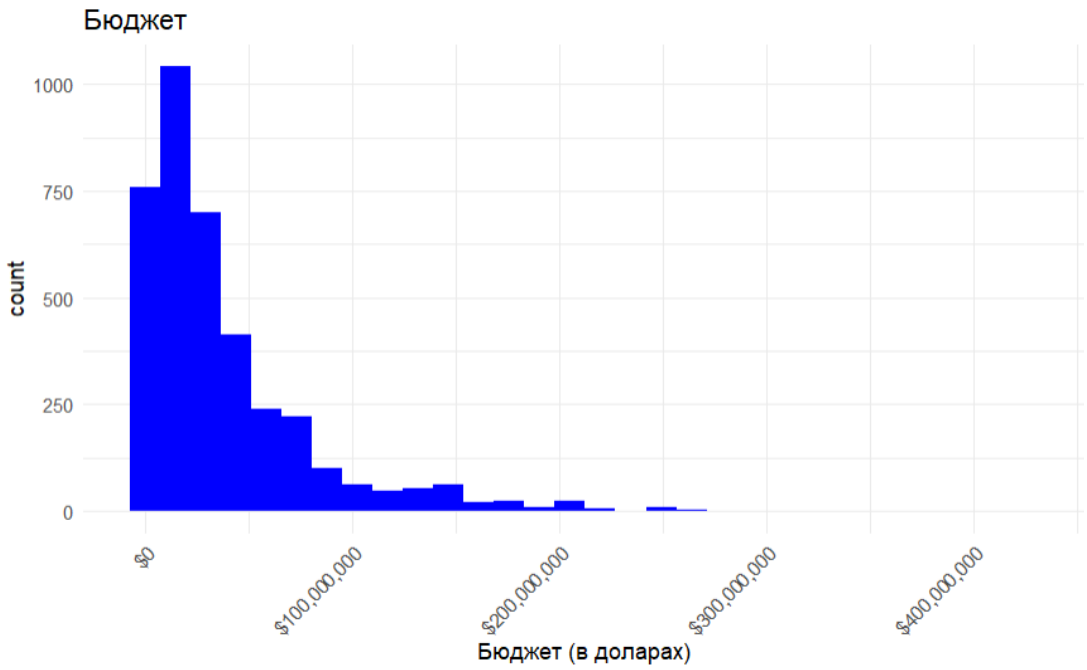

Рис 3.2 - Графік розподілу бюджету фільмів

На графіку розподілу бюджету видно, що переважна більшість фільмів мають відносно низький бюджет, в той час як існує певна кількість фільмів з значно вищим бюджетом. Це свідчить про різноманітність фінансових інвестицій в даному наборі даних.

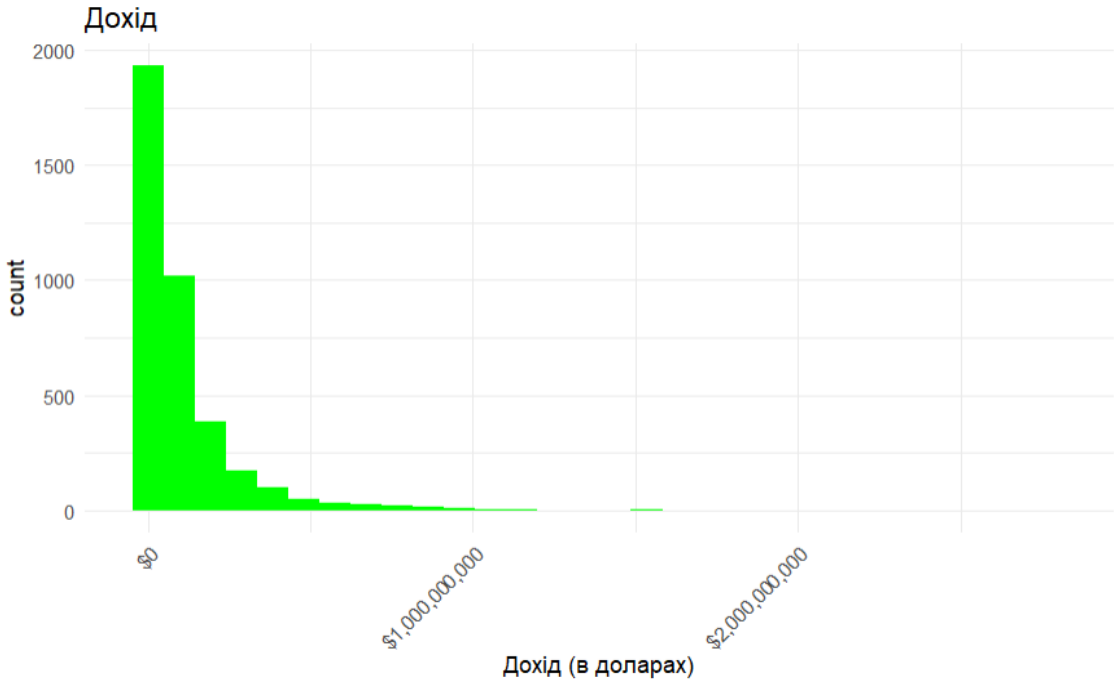

Рис 3.3 - Графік розподілу касових зборів фільмів

Графік розподілу касових зборів демонструє аналогічну картину, де більшість фільмів зібрали відносно мало коштів, але деякі з них досягли високих зборів. Це може вказувати на високу конкуренцію у кіноіндустрії та не та, що мала кількість фільмів здатна залучити велику аудиторію.

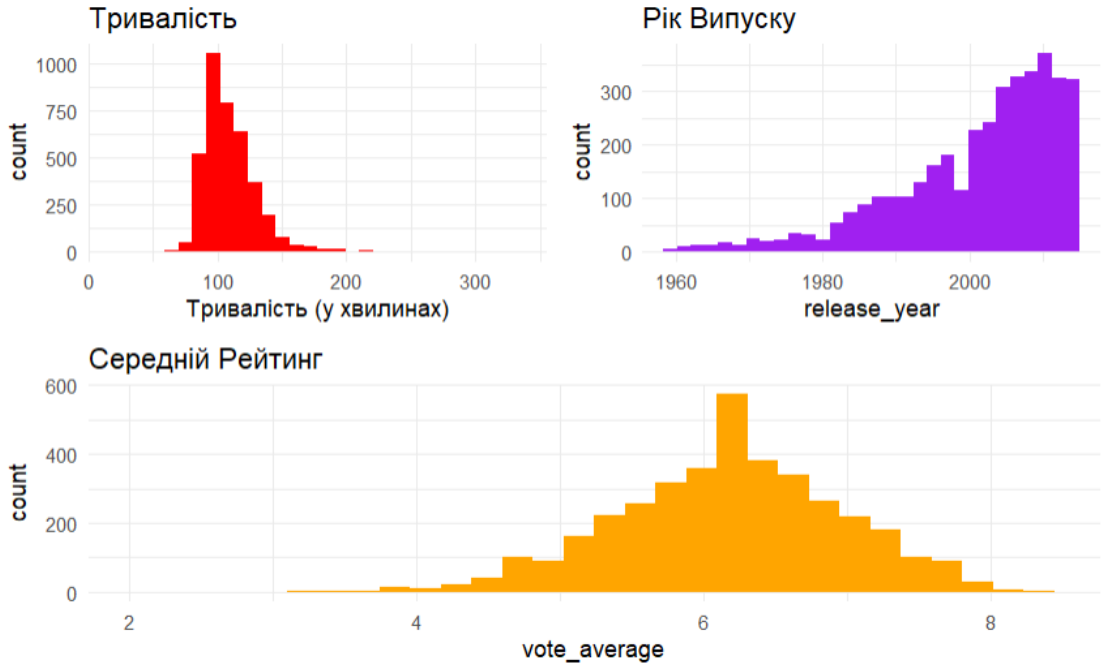

Рис 3.4 - Графіки розподілу тривалості, середнього рейтингу та року випуску фільмів

Середній рейтинг фільмів, в основному зосереджений в діапазоні між 5 і 8 балами, що свідчить про тенденцію отримання фільмами трохи вищих за середню оцінок.

Тривалість основної частини фільмів варіюється від 90 до 120 хвилин, що є типовим для стандартних художніх фільмів.

Розподіл року випуску показує, що фільми у наборі даних походять з різних років, з досить рівномірним представленням кожного року.

Розподіл кількості голосів показує, що більша частина фільмів отримала відносно невелику кількість голосів, що може свідчити про обмежену популярність більшості фільмів у датасеті.

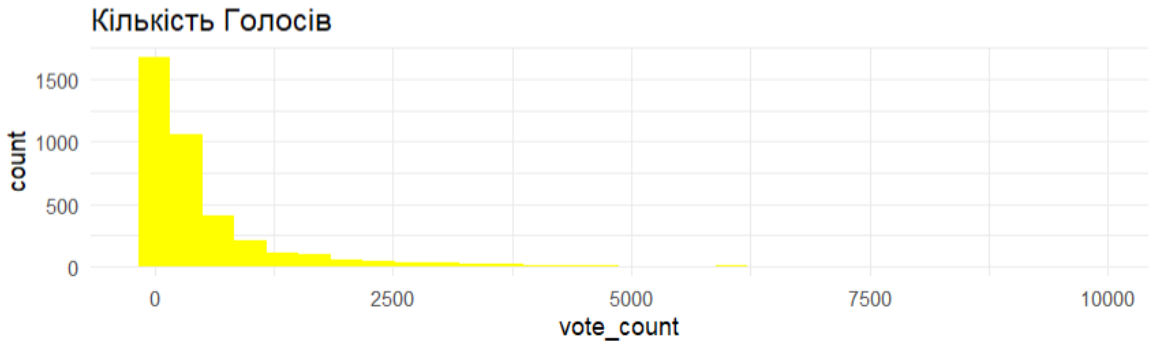

Рис 3.5 - Графік розподілу кількості голосів

Ці графіки розподілу несуть корисну інформацію для подальшого аналізу та визначення ключових факторів, які можуть впливати на рейтинг та касові збори фільмів.

3.2 Кореляційний аналіз та візуалізація

Слід провести кореляційний аналіз та візуалізація даних, з метою виявлення взаємозв’язків між різними змінними та їх впливом на касові збори та рейтинг фільмів.

Зосередимося на вивченні зв'язків між основними змінними, такими як бюджет, касові збори, тривалість фільму, кількість голосів, середня оцінка та рік випуску. Зобразимо їх за допомогою кореляційної матриці (рис. 3.6). На ній представлення значення кореляції, що варіюються від -1

до 1, де 1 означає сильний позитивний зв'язок між змінними, -1 означає сильний негативний зв'язок, а значення близьке до 0, вказує на слабкий або відсутній зв'язок.

Лістинг коду 3.1:

```
# Кореляційний аналіз
data_numbers <- TMDB_cleaned %>% select(-c(original_title, genres))
correlation_matrix <- cor(data_numbers)
# Візуалізація кореляційної матриці
ggcorrplot(correlation_matrix, lab = TRUE,title ='Correlation Matrix of Movie Variables', lab_size = 3)
```

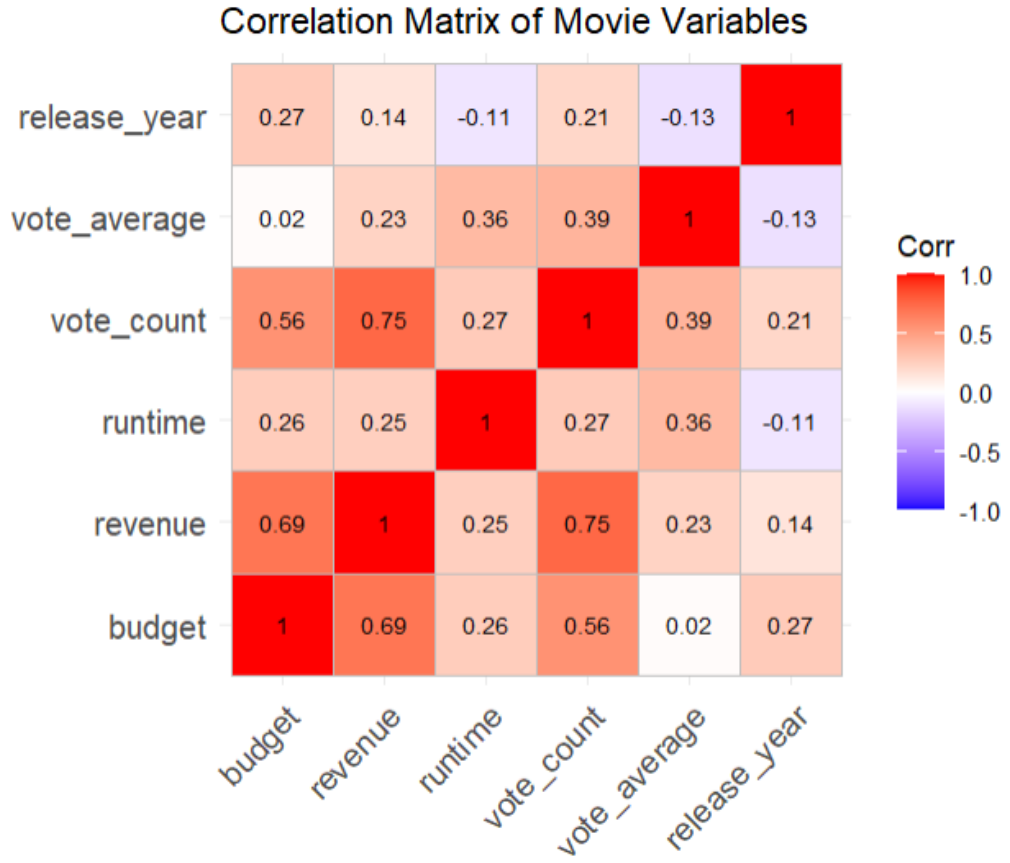

Рис 3.6 - Кореляційна матриця

Існує сильний позитивний зв'язок між бюджетом та касовими зборами фільмів. Більш детально цей зв'язок можна побачити побудувавши окремий графік розсіювання (рис 3.7).

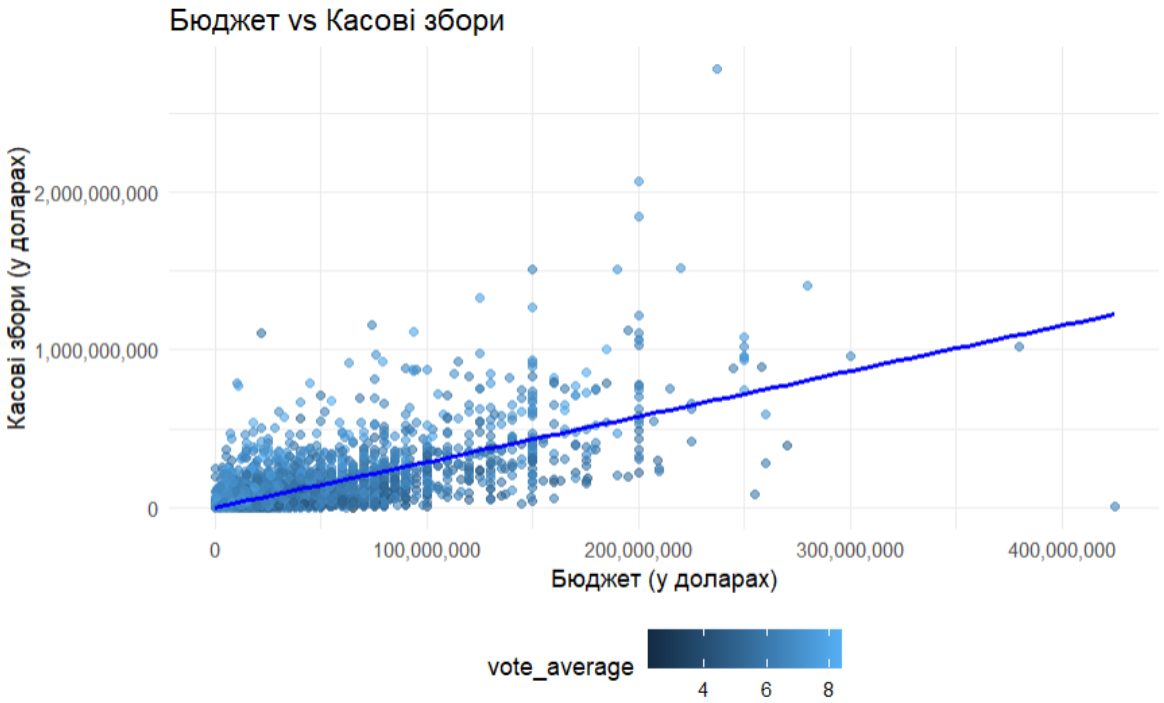

Рис 3.7 - Графік розсіювання для змінних “budget” та “revenue”

Висока позитивна кореляція змінних “budget” та “revenue” свідчить про те, що фільми з більшим бюджетом зазвичай мають вищі касові збори, але існують винятки, коли фільми з високим бюджетом провалювались в прокаті. Це можна пов’язати з тим, що велики витрати на виробництво дозволяють створювати більш високоякісні фільми з відомими акторами та розширеними рекламними кампаніями. Цим також можна пояснити помірно сильний позитивний зв’язок між цими змінними та кількістю голосів, оскільки фільми з більшим бюджетом чи касовими зборами привертають більше уваги й отримують більше оцінок від глядачів.

Кореляція між кількістю голосів (vote_count) та середнім рейтингом (vote_average) фільму становить приблизно 0.39. Це вказує на помірний позитивний зв'язок між цими двома змінними. Іншими словами, фільми, які отримали більше голосів, зазвичай мають трохи вищий середній рейтинг.

Цей результат може свідчити про те, що популярніші фільми, які привертають більше уваги глядачів та отримують більше голосів, також

мають тенденцію до отримання вищих рейтингів (рис 3.8). Однак цей зв'язок не є дуже сильним, що означає, що існують інші фактори, які також впливають на середню оцінку фільму.

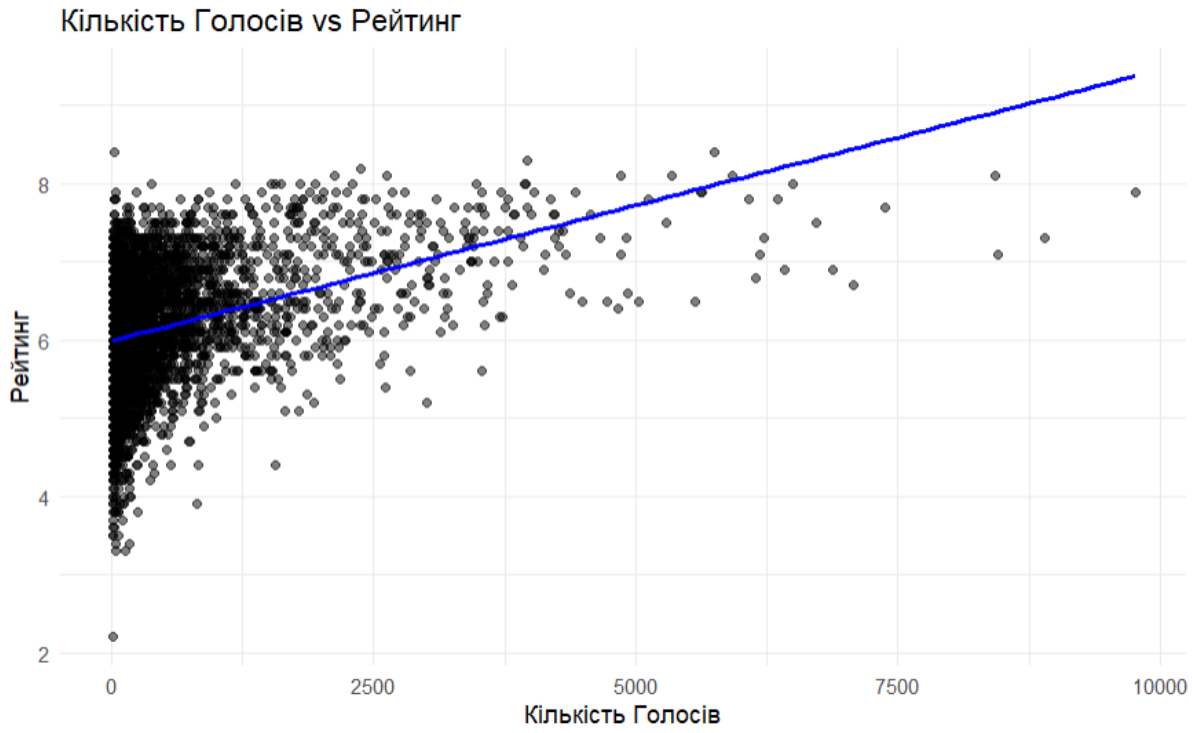

Рис 3.8 - Графік розсіювання для змінних “vote_count” та “vote_average”

Слабкий позитивний зв’язок між бюджетом і доходами фільму та його середньою оцінкою означає, що збільшення бюджету та комерційний успіх не обов’язково гарантують вищий рейтинг.

Невеликий позитивний зв’язок між тривалістю фільму та такими змінними як бюджет, касові збори та середній рейтинг може вказувати, що довші фільми мають схильність заробляти більше грошей та мають трохи вищі середні оцінки.

Кореляції між роком випуску та іншими змінними не дуже виражена, що може вказувати на те, що часовий фактор не має великого впливу на касові збори, бюджет або популярність фільму.

Важливо пам’ятати, що висока кореляція між двома змінними не обов’язково означає, що одна прямо впливає на іншу. Інші змінні, які не включенні в аналіз можуть також впливати на результати. Також кореляції

можуть бути викривлені через нетипові значення або розподіли в даних. Для глибшого розуміння зв’язків слід провести детальніший аналіз.

Також слід проаналізувати тенденції в касових зборах та рейтингах за різними роками (рис. 3.9). Це допоможе зрозуміти, як ці показники змінюються з часом. Проведемо візуалізацію цих тенденцій.

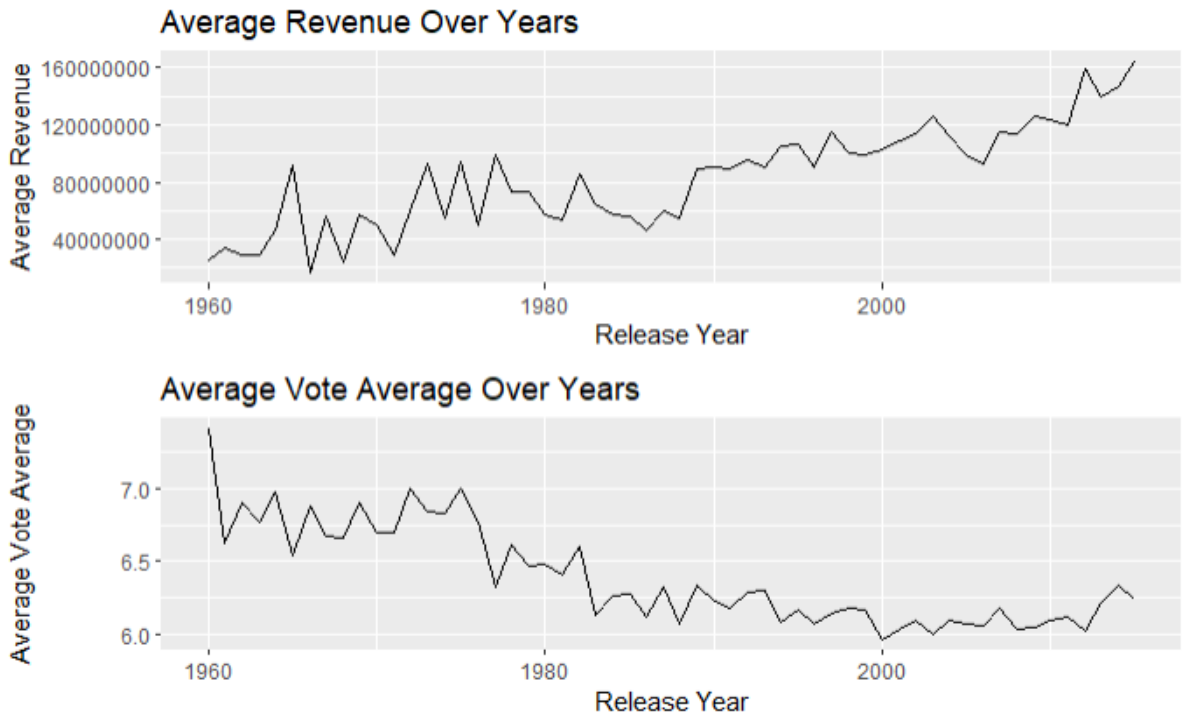

Рис 3.9 - Графіки тенденцій в касових зборах та рейтингах за роками

На графіках представлені динаміки змін касових зборів та середнього рейтингу фільмів протягом років.

З першого графіку видно, що середні касові збори фільмів коливаються з року в рік. Однак, є чітка тенденція до зростання касових зборів протягом досліджуваного періоду. Це може бути пов’язано з різними факторами, такими як зростання популярності кіно, підвищення цін на квитки, розвиток технологій тощо.

На другому графіку видно, що середній рейтинг фільмів, в досліджуваному наборі даних, з часом знижується. Починаючи з початку графіку, де середній рейтинг був близько 7 балів, він поступово зменшується до приблизно 6 у 2000-х роках. Є певні коливання в рейтингу

з року до року, але загальний тренд зниження є домінуючим. Це може бути зумовлено різними факторами, такими як зміни в індустрії кіно, в смаках аудиторії, збільшення кількості випускаємих фільмів, які можуть мати менший рейтинг.

Далі можна дослідити вплив різних жанрів на рейтинг та касові збори фільмів. Для цього слід розділити їх на окремі бінарні змінні. Такі змінні приймають лише два значення, наприклад 0 або 1. Жанри фільмів часто перетинаються, тому розділення їх на бінарні змінні дозволяє точніше визначити, який вплив має кожен з них на успіх фільмів. Інтерпретувати вплив жанрів стає простіше, коли кожен з них представлений окремою змінною. Бінарні змінні спростять аналіз, дозволяючи легко виявити взаємозв’язки між жанрами та іншими факторами.

Лістинг коду 3.2:

```
# Розділяємо жанри на різні рядки
data_expanded <- TMDB_cleaned %>%
  separate_rows(genres, sep = "\\|")
# Перетворення жанрів на бінарні змінні (dummy variables)
data_dummies <- data_expanded %>%
  pivot_wider(names_from = genres, values_from = genres, values_fill = list(genres = "Absent")) %>%
  mutate(across(-c(original_title, runtime, vote_count, vote_average, release_year, budget, revenue), ~ifelse(. == "Absent", 0, 1)))
```

Після створення бінарних змінних жанрів, для дослідження їх впливу, можемо використати графіки середніх значень.

На першому графіку (рис. 3.10) можна побачити як різні жанри фільмів впливають на їхню середню оцінку. Це дозволяє виявити, які з них є найбільш високо оцінені глядачами.

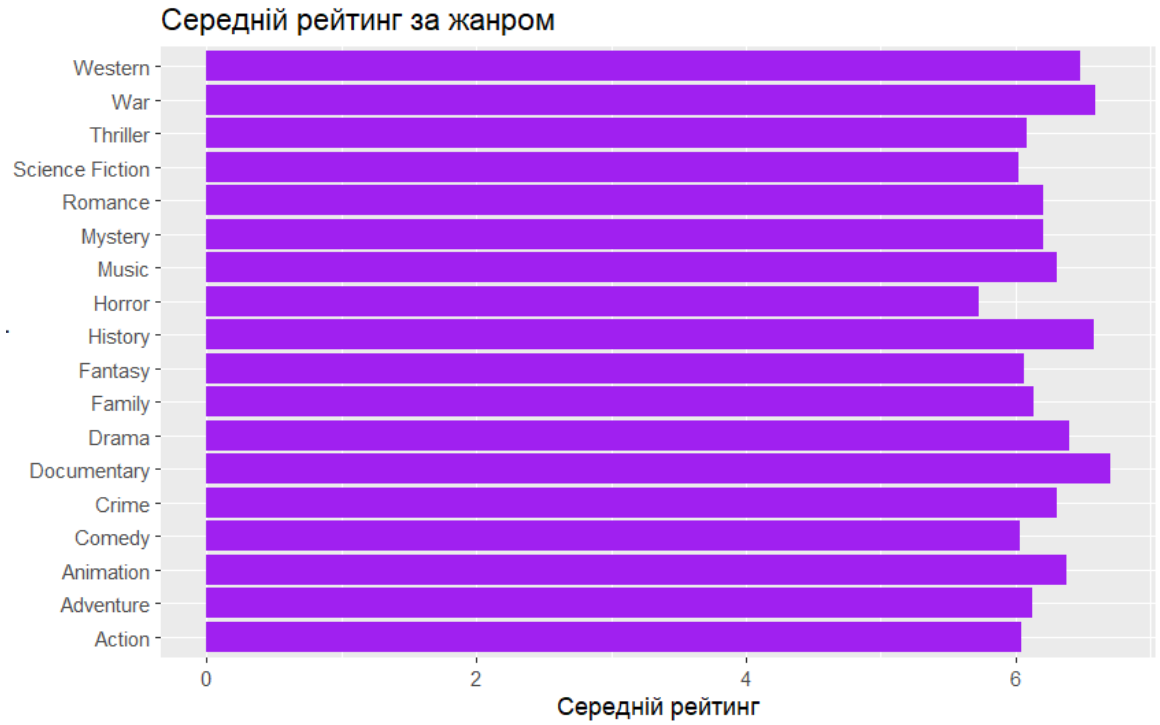

Рис 3.10 - Графік середнього рейтингу за жанром

Слід звернути увагу на кілька ключових аспектів:

- Жанри, які, як правило, отримують високі середні оцінки, зазвичай включають драму, історичні, військові та біографічні фільми. Це може вказувати на те, що глядачі високо оцінюють фільми з глибокими сюжетами та розвиненими персонажами. Такі фільми часто мають сильний емоційний вплив та змушують задуматися;
- Жанри, що займають середні позиції у рейтингу, можуть включати різноманітні типи фільмів, від романтичних до пригодницьких та кримінальних;
- Такі жанри як комедія, бойовики, фантастика та горор, часто мають нижчі оцінки. Це може бути пов’язано з тим, що вони зазвичай не містять складних сюжетів або глибоких персонажів.

Загалом, ці спостереження показують, що глядачі віддають перевагу в високій оцінці фільмів, які багаті на емоції та мають інтелектуально

насичені сюжети, що часто знаходяться в таких жанрах, як драми та історичні фільми. З іншого боку, більш легкі та розважальні жанри, хоча й популярні, можуть не досягати таких високих середніх оцінок.

На другому графіку (рис. 3.11) можна побачити, які жанри зазвичай приносять більший комерційний успіх. Кожен рядок представляє окремий жанр, а ширина стовпця відображає середній чистий дохід для нього. Цей графік надає уявлення про те, які жанри є найбільш прибутковими після врахування витрат на бюджет.

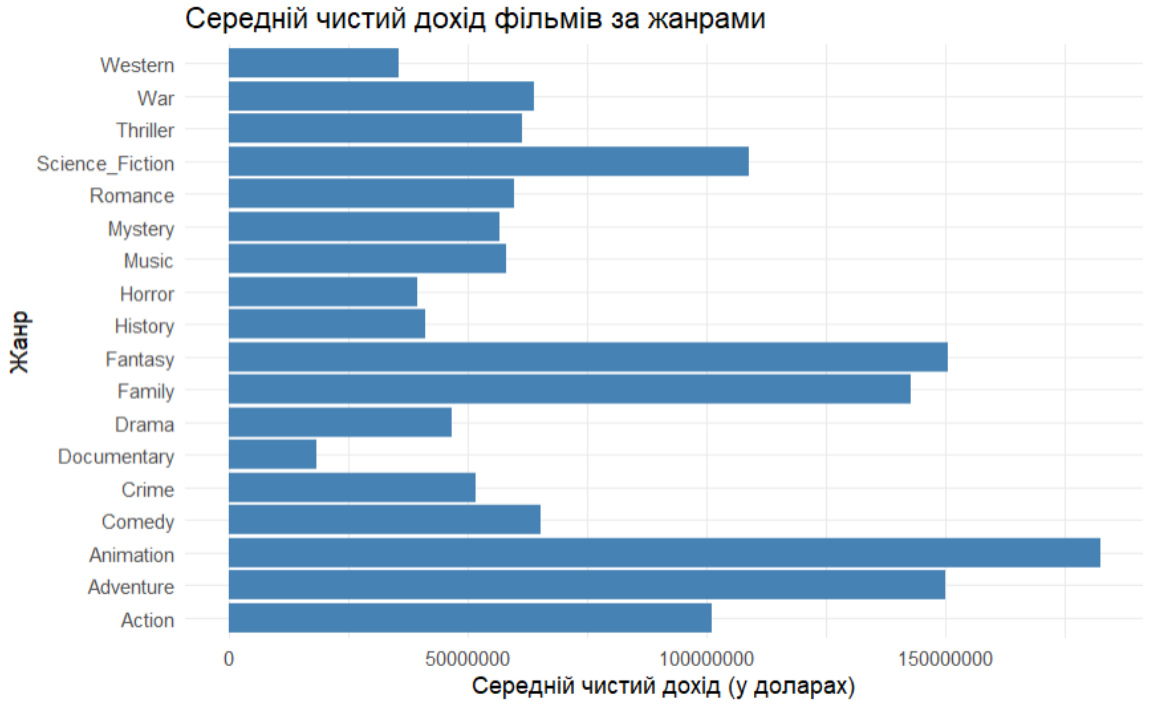

Рис 3.11 - Графік середнього чистого доходу фільмів за жанрами

Аналіз графіка середніх касових зборів по жанрах дозволяє зробити наступні висновки:

- Жанри з найбільшими касовими зборами: Це мультфільми, бойовики, пригодницькі фільми, фентезі та науково-фантастичні фільми. Їхній комерційний успіх може бути обумовлений тим, що ці жанри часто мають великі бюджети, високий рівень спецефектів та широку маркетингову кампанію;
- Середні касові збори: Жанри з середніми касовими зборами, такі

як драми, трилери чи комедії, можуть не мати такого ж великого комерційного успіху, як попередні жанри, але все ж здатні залучити значну аудиторію;

- Нижчі касові збори: Жанри з нижчими середніми касовими зборами, такі як документальні або артхаусні фільми, можуть мати обмеженішу аудиторію через специфічність сюжету, стилістичні особливості або менш широку дистрибуцію. Ці фільми часто звертаються до певної ніші глядачів та можуть мати вищу культурну або художню цінність, але не завжди високі касові збори.

Ці графіки надають важливу інформацію про оцінку та комерційну успішність різних жанрів у кіноіндустрії. Це може бути корисною інформацією для розробників фільмів, маркетологів, а також для подальших досліджень у сфері кіноіндустрії.

3.3 Регресійний аналіз

Далі слід зосередитись на регресійному аналізі, інструменті для розуміння взаємозв'язків між різними змінними у нашому дослідженні касових зборів та рейтингів фільмів. Регресійний аналіз дозволяє нам не тільки виявити важливі фактори, які впливають на успіх фільму, але й оцінити силу та напрямок цього впливу.

Використаємо рейтинг та касові збори як залежні змінні, щоб оцінити, як різні предиктори, такі як бюджет, кількість голосів, екранний час та рік випуску, впливають на них.

Спочатку побудуємо регресійну модель для прогнозування касових зборів. Використаємо змінну “revenue” в якості відгуку, та “budget”, “runtime”, “vote_count”, “release_year” в якості предикторів (рис. 3.12).

```
Call:
lm(formula = y_train_rev ~ ., data = X_train_rev)

Residuals:
       Min         1Q     Median         3Q        Max
-685746685  -34163378   -7313334   19570384 1220676130

Coefficients:
                     Estimate       Std. Error t value             Pr(>|t|)
(Intercept)  2532800569.07497  362102218.37767   6.995     0.00000000000335 ***
budget               1.66927          0.05524  30.218 < 0.0000000000000002 ***
runtime          -134044.05519     102527.27534  -1.307                0.191
vote_count        102988.98671       2602.64007  39.571 < 0.0000000000000002 ***
release_year    -1263493.84561     179951.31010  -7.021     0.00000000000278 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98230000 on 2655 degrees of freedom
Multiple R-squared:  0.6723,    Adjusted R-squared:  0.6718
F-statistic:  1362 on 4 and 2655 DF,  p-value: < 0.00000000000000022

MSE for Revenue Model:  11286474709387372
R2 for Revenue Model:  0.672293
```

Рис 3.12 - Результат підгонки множинної лінійної регресії для прогнозування касових зборів

Ця модель має R-квадрат 0.672, це означає, що приблизно 67% варіації у касових зборах фільмів можна пояснити за допомогою обраних предикторів. Скоригований R-квадрат трохи менший, а саме 0.671, що також є досить високим показником враховуючи кількість змінних у моделі. Скоригований R-квадрат є варіацією звичайного R-квадрату, який використовується в статистичному аналізі регресії для вказівки частки варіації залежної змінної, яка пояснюється моделлю регресії, але з урахуванням кількості предикторів у моделі [49]. Звичайний R-квадрат має тенденцію зростати з додаванням нових незалежних змінних до моделі, навіть якщо ці додаткові змінні не мають статистично значимого впливу на залежну змінну. Це може призвести до перенавчання моделі, коли вона добре пояснює вибірку даних, на якій була побудована, але погано працює на нових даних. Скоригований R-квадрат виправляє це шляхом штрафування за додавання не значущих змінних. Він враховує не тільки кількість змінних, але і розмір вибірки, і може зменшуватися, якщо додати

до моделі незначущі змінні. Тому скоригований R-квадрат дає більш надійну оцінку якості моделі.

Коефіцієнти моделі прогнозування касових зборів:

- Перехідний член (intercept) є значимим з великим коефіцієнтом, що свідчить про значні базові касові збори навіть без впливу предикторів;
- Бюджет має позитивний і значимий ефект на касові збори. З кожним додатковим доларом бюджету очікувані касові збори також збільшуються. Це один з найсильніших предикторів у моделі. Це підтверджується його коефіцієнтом та значенням p-value (Pr(>|t|)), де значення менше 0.05 (або іншого передбаченого порогового) вказує на статистично значущий вплив предиктора на відгук;
- Кількість голосів також є значимим предиктором з позитивним коефіцієнтами, що вказує на те, що фільми з більшою кількістю голосів, як правило, мають вищі касові збори;

Показники значущості, а саме зірочки біля коефіцієнтів, які вказують на рівень p-значення, підкреслюють надійність впливу цих змінних на касові збори. Більша кількість зірок означає вищу впевненість у впливі предиктора на змінну відгуку.

F-статистика є великою з p-значенням менше 0.0001, що свідчить про статистичну значимість моделі. F-статистика це статистичний тест, який використовується для оцінки значущості моделі регресії в цілому або для порівняння її з іншою моделлю. Вона вимірює те, наскільки добре модель пасує до спостережуваних даних порівняно з моделлю, яка не має інших предикторів, крім константи (інтерцепта) [50]. Висока F-статистика (і відповідно низьке p-значення) свідчить про те, що модель регресії в цілому має статистичну значущість і що принаймні один з предикторів моделі має ненульовий коефіцієнт. Іншими словами, це показує, що модель

краще передбачає відгук, ніж модель без предикторів (за винятком константи).

Резидуали мають широкий діапазон, що вказує на те, що хоча модель добре підходить для багатьох спостережень, є викиди або екстремальні значення, які модель не враховує.

Отже, створена модель множинної лінійної регресії є міцною у поясненні варіативності касових зборів фільмів. Ключові фактори, такі як бюджет та кількість голосів, мають міцний позитивний зв'язок з касовими зборами. Ймовірно, є викиди або виняткові випадки, де передбачення моделі не є точним, що може бути пов'язано з екстремальними значеннями бюджету або іншими факторами, які не включені до моделі.

Після створення моделі регресії для касових зборів, слід створити таку модель для середнього рейтингу фільмів. Використовуючи змінну "vote_average" як залежну і інші змінні, включаючи "budget", "runtime", "vote_count" та "release_year" як незалежні (рис. 3.13).

```
Call:
lm(formula = y_train_vote ~ ., data = X_train_vote)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2728 -0.4093  0.0443  0.4608  2.6066

Coefficients:
                     Estimate       Std. Error t value             Pr(>|t|)
(Intercept)  23.4133389947190  2.4662196675545   9.494 < 0.0000000000000002 ***
budget       -0.0000000051640  0.0000000003762 -13.725 < 0.0000000000000002 ***
runtime       0.0103005395210  0.0006982966965  14.751 < 0.0000000000000002 ***
vote_count    0.0004419436921  0.0000177261607  24.932 < 0.0000000000000002 ***
release_year -0.0091979895107  0.0012256192800  -7.505   0.0000000000000835 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.669 on 2655 degrees of freedom
Multiple R-squared:  0.291,     Adjusted R-squared:   0.29
F-statistic: 272.5 on 4 and 2655 DF,  p-value: < 0.00000000000000022

MSE for Vote Average Model:  0.4086104
R2 for Vote Average Model:  0.2910204
```

Рис 3.13 - Результат підгонки множинної лінійної регресії для прогнозування рейтингу

Загальні характеристики моделі:

- Модель має R-квадрат 0.291. Це означає, що приблизно 29% варіативності середньої оцінки голосування пояснюється вибраними предикторами;
- Скоригований R-квадрат становить 0.29, що означає, що після коригування на кількість предикторів, пояснювальна здатність моделі залишається відносно стабільною;
- Коефіцієнт для бюджету є негативним і статистично значимим, що свідчить про те, що збільшення бюджету фільму на одиницю (з урахуванням масштабування) пов’язане з незначним зниженням середньої оцінки. Це може вказувати на тенденцію до критичнішого ставлення глядачів до дорожчих фільмів або на високі очікування, які не завжди виправдовуються великими бюджетами;
- Тривалість фільму має позитивний вплив на оцінку, що може свідчити про те, що довші фільми мають більшу ймовірність отримати вищу оцінку. Це може бути пов'язано з тим, що довші фільми часто є більш глибокими або детальними в розкритті сюжету і персонажів;
- Статистично значущий і позитивний коефіцієнт для “vote_count” підтверджує, що фільми з більшою кількістю голосів мають тенденцію до вищих оцінок, що може відображати схильність популярних фільмів мати вищий рейтинг;
- Негативний коефіцієнт для року випуску фільму вказує на те, що новіші фільми мають тенденцію до нижчих середніх оцінок. Це може відображати зміни у смаках аудиторії або зростання кількості фільмів, що випускаються, з яких менша частка отримує високі оцінки;
- Відносно низька середньоквадратична помилка (MSE) та висока

F-статистика підтверджують адекватність моделі.

На основі цих даних можна зробити висновок, що хоча модель має помірну здатність пояснити різницю в середніх оцінках фільмів, існують певні чіткі тенденції, які можуть бути корисними для розуміння того, що впливає на відгуки глядачів.

Лінійна регресія є класичним інструментом статистичного аналізу, що відомий своєю здатністю виявляти зв'язки між змінними та кількісно оцінювати вплив одних змінних на інші. Її сила полягає в простоті інтерпретації та зручності використання, що робить її популярним вибором для багатьох дослідників. Проте, у контексті нашого дослідження, яке має справу з великою кількістю потенційно взаємопов'язаних предикторів і складною природою даних кінематографії, лінійна регресія може не бути найефективнішим підходом. Альтернативні методи, такі як випадковий ліс, пропонують більшу гнучкість та здатність до моделювання нелінійних зв'язків, що краще відповідає складності і різноманіттю факторів, що впливають на успіх фільму. Таким чином, інші методи можуть забезпечити більш точне і глибоке розуміння динаміки кіноіндустрії.

3.4 Випадковий ліс

Випадковий ліс ефективно обробляє великі набори даних з великою кількістю змінних, що робить його ідеальним для аналізу даних про фільми, де багато потенційних факторів впливають на рейтинг і касові збори. Крім того, цей метод надає інформацію про важливість цих факторів.

Спочатку слід підготувати дані для аналізу касових зборів, а саме видалити непотрібні стовпці, та визначити змінну “revenue” як таку, яку ми будемо прогнозувати.

Після потрібно розділити дані на навчальний та тестовий набори. Це дозволить нам навчити модель на одній частині даних і перевірити її ефективність на іншій, що не використовувалася під час навчання.

Розділимо датасет в співвідношенні 70% на 30%.

Лістинг коду 3.3:

```
# Створення моделі випадкового лісу для касових зборів
rf_revenue <- randomForest(x = train_data_revenue, y = train_revenue, ntree = 100, importance = TRUE)
# Оцінка моделі
revenue_pred <- predict(rf_revenue, test_data_revenue)
revenue_rsquare <- cor(test_revenue, revenue_pred)^2
```

Точність моделі випадкового лісу на тестовому наборі даних становить приблизно 0,76 (рис. 3.14). Це означає, що модель може пояснити близько 76% варіативності касових зборів на основі наших даних

```
"R-squared for revenue model: 0.761066926674026"
```

Рис 3.14 - Результат точності моделі

Далі ми можемо подивитися на важливість змінних, які модель визначила як найбільш значимі для прогнозування касових зборів (рис 3.15).

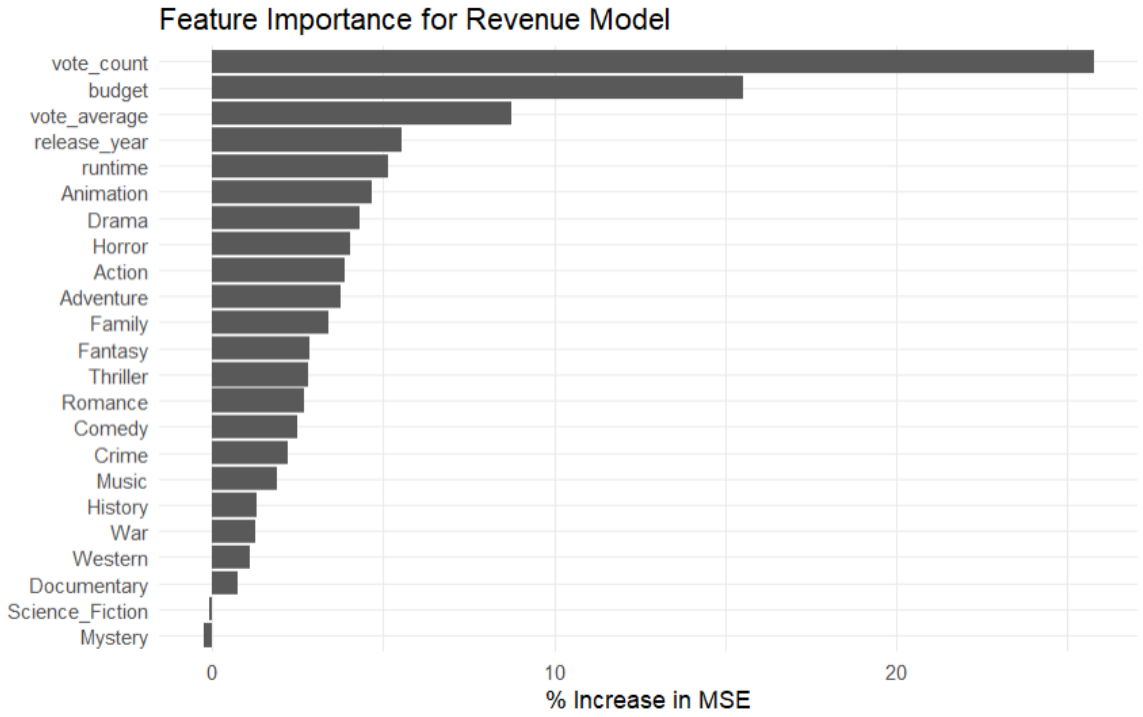

Рис 3.15 - Графік важливості змінних для моделі випадкового лісу

Важливість змінних вимірюється за допомогою відсотка збільшення

в середньоквадратичній помилці (% Increase in MSE). Ця метрика показує, наскільки якість моделі погіршиться, якщо змінну буде видалено з набору даних. Цифри на осі X представляють відсоток збільшення помилки і показують відносну важливість кожної змінної.

З графіка видно, що такі змінні, як “vote_count”, “budget”, “vote_average” є значущими для моделі, оскільки вони мають найбільший відсоток збільшення MSE. Це означає, що ці змінні найбільше впливають на здатність моделі точно прогнозувати касові збори фільмів.

Кількість голосів є важливою змінною з кількох причин:

- Велика кількість голосів часто вказує на високу популярність фільму. Популярні фільми, як правило, мають вищі касові збори, оскільки більше людей дізнаються про фільм і ходять на нього в кінотеатри;
- Фільми з високою кількістю голосів та відгуків можуть сприйматися як більш якісні або привабливі, оскільки багато інших глядачів вже висловили свою думку. Це може спонукати інших людей також подивитися фільм;
- Фільми, які активно обговорюються та отримують багато голосів, також можуть мати вигоду з додаткового маркетингового ефекту. Це може включати “сарафанне радіо”, відгуки та обговорення в соціальних медіа, що може збільшити загальну увагу до фільму і сприяти його касовим зборам;
- Фільми з високою кількістю голосів можуть продовжувати привертати увагу і генерувати дохід протягом тривалого часу, навіть після того, як вони вийшли з прокату в кінотеатрах, завдяки продажам домашніх медіа і стрімінгу.

Далі слід побудувати модель випадкового лісу для прогнозування рейтингу фільму.

Точність моделі випадкового лісу на тестовому наборі даних

становить приблизно 54,5% (рис. 3.16). Це означає, що модель може пояснити близько половини варіативності середнього рейтингу фільму на основі використаних змінних.

```
"R-squared for rating model: 0.54512342336686"
```

Рис 3.16 - Результат точності моделі

Тепер подивимось на важливість змінних для цієї моделі, щоб зрозуміти, які фактори найбільше впливають на рейтинг фільму (рис. 3.17).

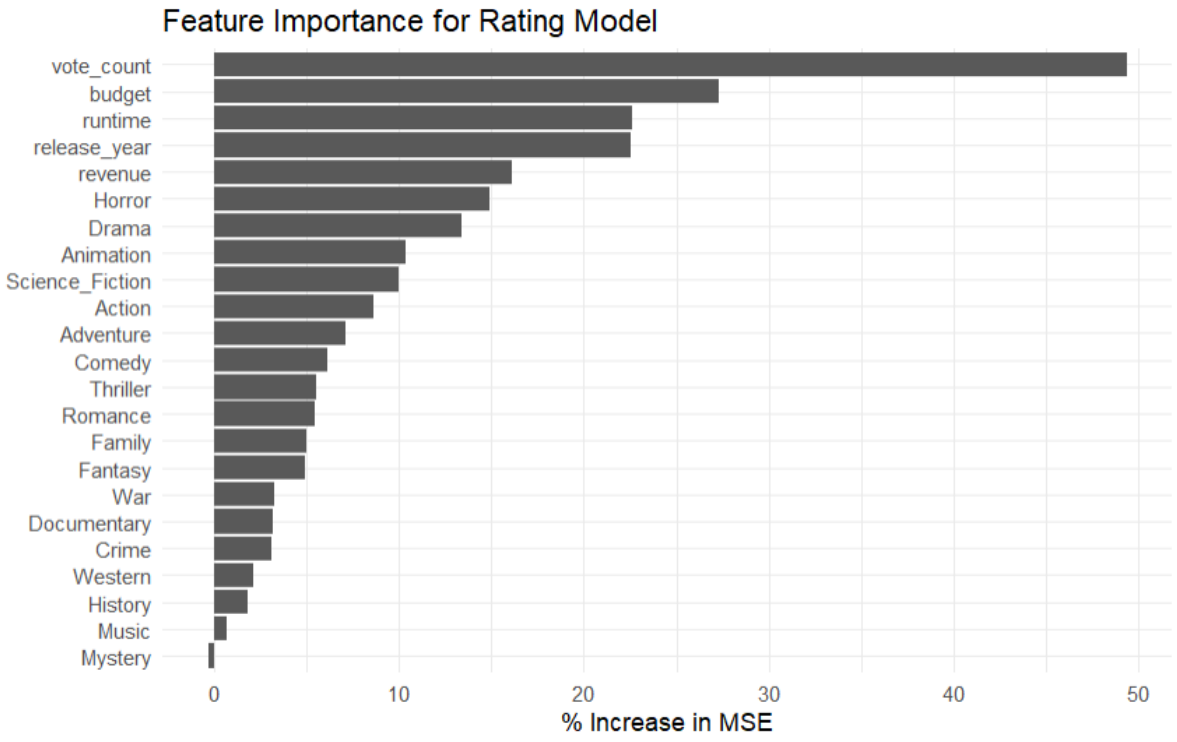

Рис 3.17 - Графік важливості змінних для моделі випадкового лісу

На основі важливості змінних, визначеної моделлю випадкового лісу, найважливішими факторами, які впливають на середній рейтинг фільму, є:

- “vote_count”: як і в моделі для касових зборів, кількість голосів є важливим індикатором, який впливає на рейтинг фільму;
- “budget”: бюджет фільму також має значний вплив, що може бути пов’язано з якістю виробництва;
- “runtime”: тривалість фільму, яка може впливати на сприйняття

глибини та розвитку сюжету та персонажів;

- “release_year”: рік випуску може відображати зміни в смаках аудиторії та стандартах якості з часом.
- “revenue”: касові збори, які можуть бути пов'язані з популярністю фільму та його прийняттям аудиторією.

Інші жанри фільмів, такі як драма і фільми жахів, також вносять свій вклад у визначення рейтингу, що може відображати вподобання аудиторії до певних типів контенту.

Ці висновки можуть бути корисними для розробників фільмів, маркетологів та кінокритиків, які прагнуть розуміти, що впливає на рейтинг та касові збори фільмів. В майбутніх дослідженнях можна додатково дослідити не тільки вплив окремих жанрів, але й інших можливих предикторів, таких як акторський склад, репутація режисера, сезонність випуску та інше.

На основі проведеного аналізу, можна зробити певні рекомендації:

- Продюсерам та інвесторам слід ретельно планувати бюджет, зважаючи на його вплив на потенційні збори та рейтинг;
- Важливе інвестування в маркетинг для збільшення кількості голосів та популярності фільму, оскільки це має великий вплив на досліджувані показники;
- Слід забезпечити оптимальну тривалість фільму, яка забезпечуватиме глибоку розповідь, але не буде втомлювати аудиторію;
- Розробники фільмів могли б зосередитися на жанрах, які історично показують високі рейтинги та касові збори.

Висновок до третього розділу

В даному розділі було проведено всебічний статистичний аналіз даних про кінофільми. Зокрема, було розглянуто описову статистику, в якій було досліджено розподіл та тенденцію даних в датасеті. Проведено

кореляційний аналіз та візуалізацію різних графіків, для дослідження впливу різних змінних на касовий успіх та рейтинг фільмів.

Важливим, було побудова регресійних моделей, а саме багатофакторної лінійної регресії для передбачення рейтингу та касових зборів. З допомогою неї досліджено силу та напрямок впливу різних факторів на успіх фільму та розглянуто їхню взаємодію. Виявлено, які фактори мають вагомий вплив на досліджуванні аспекти.

Також було побудовано моделі випадкового лісу для передбачення досліджуваних показників, які показали хорошу здатність до моделювання нелінійних залежностей.

ВИСНОВОК

У роботі було проведено комплексне дослідження факторів, які визначають успіх фільму у контексті його рейтингу та касових зборів.

В першому розділі було звернено увагу на роль кіноіндустрії у сучасному суспільстві, основи формування рейтингу та касових зборів у минулому та в сучасності. Детальний аналіз публікацій і досліджень у цій сфері дозволив визначити ключові тенденції та зміни в оцінці якості кінопродукції.

У другому розділі основний акцент було зроблено на методологічні аспекти дослідження. Визначено завдання дослідження, представлено обрані інструменти та технології, серед яких мова програмування R, середовище RStudio та пакет Tidyverse. Було підкреслено важливість вибору адекватного набору даних, його підготовки та очищення. Додатково, обґрунтовано вибір методів аналізу даних, спрямованих на досягнення поставлених цілей дослідження.

Третій розділ зосереджується на практичному застосуванні вибраних методів для статистичного аналізу даних про кінофільми. Використання кореляційного аналізу, багатофакторної лінійної регресії та моделей випадкового лісу дозволило виявити ключові фактори, які впливають на рейтинг та касові збори фільмів, виявивши важливі закономірності та тенденції.

Висновки цієї роботи важливі не тільки для розуміння динаміки кіноіндустрії, але й для вибору стратегій розвитку та просування кінопроектів. Застосування методів Data Science виявилося ефективним для аналізу складних і різноманітних даних, що стосуються кіноіндустрії, та дало можливість краще зрозуміти фактори, які визначають успіх фільму на ринку. Ці висновки мають потенціал бути корисними для кінематографістів, маркетологів та дослідників, які цікавляться аналізом і прогнозуванням успіху кінопродукції.

СПИСОК ЛІТЕРАТУРИ

1. JUZER S. TMDb Movies Dataset [Електронний ресурс] / SHAKIR JUZER // kaggle. – 2017. – Режим доступу до ресурсу: https://www.kaggle.com/datasets/juzershakir/tmdb-movies-dataset.
2. Марченко І. Кінематограф в житті сучасної людини [Електронний ресурс] / Ігор Марченко. – 2016. – Режим доступу до ресурсу: https://vbusk.com/cikavo/kinematograf-zhizni-sovremennogo-cheloveka.html.
3. Rilind E. HOW DO MOVIES IMPACT OUR SOCIETIES [Електронний ресурс] / Elezaj Rilind. – 2019. – Режим доступу до ресурсу: https://yourstory.com/mystory/how-movies-impact-societies.
4. Tuttle B. Movie Theaters Make 85% Profit at Concession Stands [Електронний ресурс] / Brad Tuttle. – 2009. – Режим доступу до ресурсу: https://business.time.com/2009/12/07/movie-theaters-make-85-profit-at-concession-stands/.
5. How Do Movies Affect Society? [Електронний ресурс] // PoutyBoy. – 2017. – Режим доступу до ресурсу: https://www.ourmovielife.com/2017/01/15/how-do-movies-affect-society/.
6. DOZIER B. The importance of film industry in the modern world [Електронний ресурс] / BARBRA DOZIER. – 2015. – Режим доступу до ресурсу: https://barbradozier.wordpress.com/2015/02/13/the-importance-of-film-industry-in-the-modern-world/.
7. Drew J. 6 Key Benefits of Using Movies in Education [Електронний ресурс] / Josh Drew. – 2020. – Режим доступу до ресурсу: https://www.liveforfilm.com/2020/11/13/6-key-benefits-of-using-movies-in-education/.
8. MAHER M. Where Do Film Ratings Come From? [Електронний ресурс] / MICHAEL MAHER. – 2016. – Режим доступу до ресурсу: https://www.premiumbeat.com/blog/where-do-film-ratings-come-from/.

9. Шершакова Н. Як працює IMDb і чому йому можна довіряти [Електронний ресурс] / Наталя Шершакова // yummymovie. – 2019. – Режим доступу до ресурсу: https://yummymovie.org/ru/1575/.
10. How IMDb Ratings Work: A Scientific Approach [Електронний ресурс] // Cinema Peedika. – 2022. – Режим доступу до ресурсу: http://cinemapeedika.com/author/dmppdvik9k/.
11. Kumar A. What Is Box Office Collection? How Is Box Office Calculated? – Explained In Detail [Електронний ресурс] / Arun Kumar. – 2023. – Режим доступу до ресурсу: https://topmovierankings.com/box-office/meaning-metrics-history-how-is-box-office-calculated#Historical_Background_of_Box_Office_in_Cinema.
12. He Q. Research on the Influencing Factors of Film Consumption and Box Office Forecast in the Digital Era: Based on the Perspective of Machine Learning and Model Integration / Q. He, B. Hu. // Hindawi. – 2021. – С. 4–5.
13. Litman B. R. Predicting financial success of motion pictures: The '80s experience / B. R. Litman, L. S. Kohl. // Journal of Media Economics. – 1989. – №2. – С. 35–50.
14. De Vany A. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? / A. De Vany, D. Walls. // Journal of Cultural Economics. – 1999. – №23. – С. 285–318.
15. Ravid A. S. Information, Blockbusters, and Stars: A Study of the Film Industry / Abraham Ravid. // The Journal of Business. – 1999. – №72. – С. 463–492.
16. Eliashberg J. Film Critics: Influencers or Predictors? / J. Eliashberg, S. M. Shugan. // Journal of Marketing. – 1997. – №62. – С. 68–78.
17. Reinstein D. The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics / D. Reinstein, S. ChriCristopher Mark. // Journal of Industrial Economics. – 2005. – №53. – С. 27–51.

18. Pangarker N. A. The determinants of box office performance in the film industry revisited / N. A. Pangarker, E. Smit. // South African Journal of Business Management. – 2013. – №44. – С. 47–58.
19. Ramos M. Statistical Patterns in Movie Rating Behavior / M. Ramos, A. Calvão, C. Anteneodo. // PLoS ONE. – 2015. – №10.
20. Carbonnelle P. PYPL PopularitY of Programming Language [Електронний ресурс] / Pierre Carbonnelle // PYPL. – 2023. – Режим доступу до ресурсу: https://pypl.github.io/PYPL.html.
21. R Programming Language – Introduction [Електронний ресурс] // GeeksforGeeks. – 2021. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/r-programming-language-introduction/.
22. What is R? [Електронний ресурс] // The R Foundation – Режим доступу до ресурсу: https://www.r-project.org/about.html.
23. Python Vs R: Know The Difference [Електронний ресурс] // InterviewBit. – 2023. – Режим доступу до ресурсу: https://www.interviewbit.com/blog/python-vs-r/.
24. Ghosh R. What is R Coding Language and Why is it so Important? [Електронний ресурс] / Riku Ghosh // Emeritus. – 2023. – Режим доступу до ресурсу: https://emeritus.org/blog/coding-r-coding-language/.
25. RStudio [Електронний ресурс] // Wikipedia. – 2023. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/RStudio.
26. What Is RStudio? A Beginner's Guide [Електронний ресурс] // CAREERFOUNDRY. – 2023. – Режим доступу до ресурсу: https://careerfoundry.com/en/blog/data-analytics/what-is-rstudio/.
27. R Markdown [Електронний ресурс] // RStudio. – 2020. – Режим доступу до ресурсу: https://rmarkdown.rstudio.com/index.html.
28. Tidyverse [Електронний ресурс] // Wikipedia. – 2023. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Tidyverse.

29. Wickham H. ggplot2: Elegant Graphics for Data Analysis / Hadley Wickham. – New York: Springer International Publishing, 2016. – 260 с.
30. dplyr: A Grammar of Data Manipulation [Електронний ресурс] / [H. Wickham, R. François, L. Henry та ін.] // Posit. – 2023. – Режим доступу до ресурсу: https://dplyr.tidyverse.org/.
31. Wickham H. tidyr: Tidy Messy Data [Електронний ресурс] / H. Wickham, D. Vaughan, M. Girlich // Posit. – 2023. – Режим доступу до ресурсу: https://tidyr.tidyverse.org.
32. Wickham H. readr: Read Rectangular Text Data [Електронний ресурс] / H. Wickham, J. Hester, J. Bryan // Posit. – 2023. – Режим доступу до ресурсу: https://readr.tidyverse.org.
33. Wickham H. purrr: Functional Programming Tools [Електронний ресурс] / H. Wickham, L. Henry // RStudio. – 2023. – Режим доступу до ресурсу: https://purrr.tidyverse.org/.
34. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations [Електронний ресурс] / Hadley Wickham // RStudio. – 2022. – Режим доступу до ресурсу: https://stringr.tidyverse.org.
35. Müller K. tibble: Simple Data Frames [Електронний ресурс] / K. Müller, H. Wickham // RStudio. – 2023. – Режим доступу до ресурсу: https://tibble.tidyverse.org/.
36. Wickham H. forcats: Tools for Working with Categorical Variables (Factors) [Електронний ресурс] / Hadley Wickham // RStudio. – 2023. – Режим доступу до ресурсу: https://forcats.tidyverse.org/.
37. TMDb [Електронний ресурс] // Вікіпедія. – 2023. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/TMDb.
38. Let's talk about TMDB [Електронний ресурс] // TMDB. – 2023. – Режим доступу до ресурсу: https://www.themoviedb.org/about.
39. IMDb [Електронний ресурс] // Wikipedia. – 2023. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/IMDb.

40. Geisler T. M. Data Cleaning Steps & Process to Prep Your Data for Success [Електронний ресурс] / Tobias Mesevage Geisler // MonkeyLearn. – 2021. – Режим доступу до ресурсу: https://monkeylearn.com/blog/data-cleaning-steps/.
41. HILLIER W. What Is Data Cleaning and Why Does It Matter? [Електронний ресурс] / WILL HILLIER // CAREERFOUNDRY. – 2023. – Режим доступу до ресурсу: https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/.
42. Прокляття розмірності [Електронний ресурс] // Wikipedia. – 2023. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%BA%D0%BB%D1%8F%D1%82%D1%82%D1%8F_%D1%80%D0%BE%D0%B7%D0%BC%D1%96%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%96.
43. fix: Fix an Object [Електронний ресурс] // rdocumentation – Режим доступу до ресурсу: https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/fix.
44. Описова статистика [Електронний ресурс] // wikipedia. – 2023. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/%D0%9E%D0%BF%D0%B8%D1%81%D0%BE%D0%B2%D0%B0_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0.
45. Бахрушин В. Є. Методи аналізу даних : навчальний посібник для студентів / В. Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.
46. Регресійний аналіз - що це таке, визначення та поняття [Електронний ресурс] // Economy-Pedia – Режим доступу до ресурсу: https://uk.economy-pedia.com/11032660-regression-analysis.
47. Breiman L. Random Forests / Leo Breiman. // SpringerLink. – 2001. – №45. – С. 5–32.

48. Bondar S. Random Forests in R [Електронний ресурс] / Sasha Bondar // REINTECH. – 2023. – Режим доступу до ресурсу: https://reintech.io/blog/introduction-to-random-forests-in-r.
49. Відрегульований квадрат R (відкоригований коефіцієнт детермінації) [Електронний ресурс] // Economy-Pedia – Режим доступу до ресурсу: https://uk.economy-pedia.com/11038611-adjusted-r-squared-adjusted-coefficient-of-determination.
50. Статистичний F - що це таке, визначення та поняття [Електронний ресурс] // Economy-Pedia – Режим доступу до ресурсу: https://uk.economy-pedia.com/11036174-f-statistic.

Додаток 2 до наказу
від «31» березня 2023 року
№119/05

ДЕКЛАРАЦІЯ
про дотримання академічної доброчесності

Я,_____
_____
Повністю вказується ПІБ та статус (посада для працівників, освітня (освітньо-наукова) програма – для здобувачів вищої освіти)

що нижче підписалась/підписався, розуміючи та підтримуючи загальновизнані засади справедливості, доброчесності та законності,

ЗОБОВ'ЯЗУЮСЬ:

дотримуватися принципів та правил академічної доброчесності, що визначені законодавством України, локальними нормативними актами Донецького національного університету імені Василя Стуса, положеннями, правилами, умовами, визначеними іншими суб'єктами, та не допускати їх порушення.

ПІДТВЕРДЖУЮ:

що мені відомі положення статті 42 Закону України «Про освіту»;

що у даній роботі не представляла/представляв чиїсь роботи повністю або частково як свої власні. Там, де я скористалася/скористався працею інших, я зробила/зробив відповідні посилання на джерела інформації;

що дана робота не передавалась іншим особам і подається вперше, не порушує авторських та суміжних прав закріплених статтями 21-25 Закону України «Про авторське право та суміжні права», а дані та інформація не отримувались в недозволений спосіб.

УСВІДОМЛЮЮ:

що ця робота може бути перевірена університетом на плагіат або інші порушення академічної доброчесності, в тому числі з використанням спеціалізованих сервісів;

що у разі порушення академічної доброчесності, до мене можуть бути застосовані процедури, передбачені законодавством України та Кодексом академічної доброчесності та корпоративної етики Донецького національного університету імені Василя Стуса, іншими локальними нормативними актами університету, та я можу бути притягнута/притягнутий до академічної відповідальності.

_____ _____
(дата) (підпис)